Facets of regularization in high-dimensional learning:
Cross-validation, risk monotonization, and model complexity

Pratik Patil

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

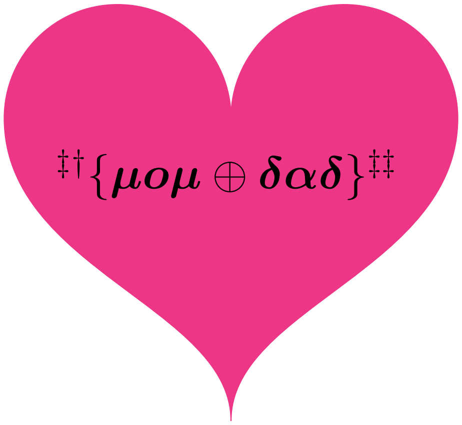
Thesis Committee:
Ryan Tibshirani (Chair)
Alessandro Rinaldo
Arun Kumar Kuchibhotla
Yuting Wei (University of Pennsylvania)
Arian Maleki (Columbia University)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Support: The work in this thesis was supported in part by the Multidisciplinary University Research Initiatives Program by the Office of Naval Research under the grant N00014-20-1-2787.

Disclaimer: Any opinions, findings, conclusions, and recommendations expressed in this work are those of the author and do not necessarily reflect the views of the sponsoring and affiliate institutions.

$\mathscr{T}\,\mathscr{O}$


$^{\ddagger\dagger}\{\boldsymbol{\mu o \mu} \oplus \boldsymbol{\delta\alpha\delta}\}^{\ddagger\ddagger}$

[†] First, before you wonder, even though the order is immaterial as the binary operation is assumed to obey commutativity, it is nonetheless decided by a random draw of Bern(p) for some $p \in (0, 1)$, whose exact value remains unknown to this day ...

[‡] Second, I have tried best to calibrate the color of the heart to match mom's favorite ruby, used greek letters for decorations to sneak in dad's favorite particle, and stylized the dedication letters in my favorite script, so each of us gets one favorite ...

Third, this is a dummy point as LaTeX does not have a nice default footnote symbol that visually indicates three and maintains the symmetry in the lettering that I want. Or you could pretend these go in powers of two and ignore this point ...

[‡‡] Fourth, if you remember, I hand-drew a version isomorphic to this (though not as geeked out) as a kid, then I used Paint in school, then Inkscape in college, and now I am using TikZ in its entirety as a "grown up". So some things have changed: the tools and skills have evolved all round; but it is the same core emotion it was all those years ago. Countless thanks for the countless sacrifices in teaching me the best bits of most all I know. The art of intuition and insights learned from "street-fighting" math then has carried me quite well so far. Apologies that I did not follow footsteps after both of you for a "pure" academic career in physics, but hopefully this thesis convinces you that the 🍎 has not fallen too far from the 🌳 ...

Abstract

This thesis studies aspects of regularization in a high-dimensional regime in which the feature size grows proportionally with the sample size. Several commonly used prediction procedures, such as ridge and lasso, exhibit peculiar risk behavior in this regime: no explicit regularization can be optimal for (random-X) test error, the risk can be non-monotonic in the sample size, and the risk curve can exhibit double or multiple descents in the feature size, treated as a complexity measure. In this thesis, we present results on cross-validation, risk monotonization, and model complexity along these angles.

Cross-validation. We show strong uniform consistency of generalized and leave-one-out cross-validation (GCV and LOOCV) for estimating the squared test error of ridge regression. Consequently, we show that ridge tuning via GCV or LOOCV almost surely delivers the optimal regularization, be it positive, negative, or zero. Furthermore, by suitably extending GCV and LOOCV, we construct consistent estimators of the entire test error distribution and a broad class of its linear and nonlinear functionals. Our results require only minimal moment assumptions on the data distribution and are model-agnostic.

Risk monotonization. We develop a framework that modifies any generic prediction procedure such that its risk is asymptotically monotonic in the sample size. As part of our framework, we propose two data-driven methodologies, namely zero- and one-step, that are akin to bagging and boosting, respectively, and show that under very mild assumptions they achieve monotonic asymptotic risk behavior. Our results are applicable to a wide class of prediction procedures and loss functions, and do not assume a well-specified model. We exemplify our framework with concrete analyses of the ridgeless and lassoless procedures.

Model complexity. We revisit model complexity through the lens of model optimism and degrees of freedom. By re-interpreting degrees of freedom in the fixed-X prediction setting, we extend this concept to the random-X prediction setting. We then define a family of complexity measures, whose two extreme ends we call the emergent and intrinsic degrees of freedom of a prediction model. Through linear and nonlinear example models, we illustrate how the proposed measures may prove useful to align the subtle multiple descents behavior with the typical single descent behavior observed in classical statistical prediction.

Key words: Proportional asymptotics, overparameterized learning, random matrix theory, deterministic equivalents, asymptotic equivalents, ridge regression, lasso, minimum ℓ_2-, ℓ_1-norm interpolators, double descent, optimal regularization, risk monotonicity, cross-validation, bagging, boosting, divide-and-conquer, model optimism, degrees of freedom, model complexity.

Acknowledgments

Graduate school is exciting but also hard, with many peaks and valleys throughout the journey. One has to sift through a large body of prior work and try to find the right "words" for one's intuitions and make them precise. I feel very fortunate to have been in the company of many incredible individuals who enthusiastically shared their learned "vocabulary" and helped to create some new one, and along the way made me a better researcher. Even though my name is on the front page, the work presented in this thesis is collaborative and it would not have been possible without their (direct and indirect) help that I am delighted to acknowledge below!

I am foremost indebted to my advisor, Ryan J. Tibshirani, for all his technical help, guidance, and support throughout this work. I am also grateful to him for patiently teaching me the value of a good process and serving as a role model. During a long endeavor like this, indeed a marathonic mindset has proved (and continues to prove) more useful than a sprintic mindset. The valuable lessons learned during this time will remain with me for the rest of my career. I am grateful to Alessandro Rinaldo for almost serving as a second advisor and his mentorship along every step from the very beginning. His kindness, generosity, and cheerful outlook on life has greatly shaped my own. Many thanks to Arun Kumar Kuchibhotla and Yuting Wei for being awesome mentors and friends, and helping crucially shape the directions in the thesis. I have learned a lot from all our lively and stimulating discussions. I am thankful to Arian Maleki for providing inspirations through his earlier work on cross-validation and for his excellent feedback on this work. I am also thankful to Jin-Hong Du and Daniel LeJeune for fruitful collaborations and many insightful discussions on various asymptotic matrix equivalents used in this thesis.

The interdisciplinary academic environment at Carnegie Mellon has been wonderful for intellectual growth during the course of this work. I am grateful to all the faculty, especially in the Statistics and Machine Learning departments, for excellent courses, and always keeping their door open for any discussions. Special thanks to Larry Wasserman, Sivaraman Balakrishnan, Aaditya Ramdas, Edward Kennedy, Matey Neykov, Valérie Ventura, among others, for many, many enlightening and inspiring conversations. The love around for the field is very contagious. Thanks also to Cosma Shalizi for the endless source of book and paper recommendations that has and will keep me busy for a long time. Many thanks go to the department staff, and in particular, the department mothers, Diane Stidle and Margaret Smykla, for always swiftly taking care of all the administrative things, so we students do not have to worry them. I owe my deepest thanks to all friends here and elsewhere. I apologize for any periods of introversion (exacerbated by the pandemic) and thank them for their patience and understanding. I would rather not list all the names, but truly as long as this thesis is, this section should be longer. If you are reading this, know that you most likely belong here, without me explicitly saying so. Thank you from the bottom of my heart for helping make this journey so fun and enjoyable.

I will conclude with a broader point. Inspirations are often found in the unlikeliest of places. I have gotten many inspirations from various academic things (insightful papers, enlightening talks, elegant arguments, lively discussions, etc.), often very far to the actual thesis content, and various non-academic things (eloquent writing, meticulous artwork, imaginative poetry, perfected recipe, etc.), definitely very very far from the thesis content. The common denominator in all these is that an even outsider can "feel" the passion poured into the activity, no matter how small or big, important or unimportant, but often done solely for the sheer joy and pride of it. Hard inspirations one can properly acknowledge and cite, but such soft inspirations are difficult to credit, but I think were equally or even more important for sustaining energy during this work. For all those inspirations, this is a big thank you! I will be a happy person to have paid forward some of the dues if some small part of this thesis (be it the actual work, writing, figures, or even just formatting) serves the same purpose for someone, even if very softly! Because, even if the shelf life of results in this thesis turn out to be short (in the sense of being subsumed by broader generalizations), I think this process of forward influencing persists!

Contents

Chapter 0

Overview

Modern machine learning models employ a large number of parameters relative to the number of observations. Such overparameterized models typically have the capacity to (nearly) interpolate noisy training data. Despite fitting the models until the training error is nearly zero, they often generalize well on unseen test data in practice (Zhang et al., 2017, 2021). The striking and widespread successes of interpolating models has been a topic of growing interest in the recent mathematical statistics literature (see, e.g., Belkin et al., 2019a, 2018a, 2019b; Bartlett et al., 2020), as it seemingly defies the widely-accepted statistical wisdom that interpolation will generally lead to over-fitting and poor generalization (Hastie et al., 2009, Figure 2.11). A body of recent work has both empirically and theoretically investigated this surprising phenomenon for different models, including linear regression (Hastie et al., 2022; Muthukumar et al., 2020; Belkin et al., 2020; Bartlett et al., 2020), kernel regression (Liang and Rakhlin, 2020), random features regression (Mei and Montanari, 2022), logistic regression (Deng et al., 2022), nearest neighbor methods (Xing et al., 2018, 2022), boosting algorithms (Liang and Sur, 2020a), among others. See the survey papers by Bartlett et al. (2021), Belkin (2021), and Dar et al. (2021) for more related references on overparameterized learning.

In this thesis, we take an "operational" point of view on generalization in overparameterized learning. Specifically, we study three aspects related to regularization in overparameterized models: cross-validation, risk monotonization, and model complexity. Our focus on these aspects is partly motivated by the following three broad questions:

(Q1) **Cross-validation.** Cross-validation is a widely used method for assessing the generalization performance of a learning method. The first question this thesis asks is whether cross-validation still "works" in the overparameterized regime, especially when the optimal regularization and the train error can be zero. Apart from understanding the theoretical properties of cross-validation in overparameterized regime, this is of interest as (near) interpolators can be optimal for (random-X) test error under certain data geometries.

(Q2) **Risk monotonization.** The generalization error of overparameterized models can be non-monotonic in the sample size, suggesting that increasing the sample size might actually yield a worse generalization error. The second question this thesis asks is whether it is possible to modify any given prediction procedure to achieve a monotonic risk behavior. This is of interest because it is highly desirable to rely on prediction procedures that are guaranteed to deliver a risk profile that is monotonically decreasing in the sample size.

(Q3) **Model complexity.** Overparameterized models often exhibit the so-called "double/multiple descent" behavior in the generalization error in the raw number of model parameters. The third question this thesis asks is whether there is a more principled measure of model complexity for overparameterized models. Besides comparing complexity of different interpolating models, this is of interest to attempt to "reconcile" the multiple descents behavior in such models with the common single descent behavior in classical learning.

The thesis provides results related to (Q1)–(Q3). We summarize the main highlights below.

(A1) In Chapter 1, we examine generalized and leave-one-out cross-validation for ridge regression in a proportional asymptotic framework where the dimension of the feature space grows proportionally with the number of observations. Given i.i.d. samples from a linear model with an arbitrary feature covariance and a signal vector that is bounded in ℓ_2 norm, we show that generalized cross-validation for ridge regression converges almost surely to the expected out-of-sample prediction error, uniformly over a range of ridge regularization parameters that includes zero (and even negative values). We prove the analogous result for leave-one-out cross-validation. As a consequence, we show that ridge tuning via minimization of generalized or leave-one-out cross-validation asymptotically almost surely delivers the optimal level of regularization for predictive accuracy, whether it be positive, negative, or zero. In Chapter 2, we study the problem of estimating the distribution of the out-of-sample prediction error associated with ridge regression. We show that both generalized and leave-one-out cross-validation (GCV and LOOCV) for ridge regression can be suitably extended to estimate the full error distribution. This is still possible in a high-dimensional setting where the ridge regularization parameter is zero. In an asymptotic framework in which the feature dimension and sample size grow proportionally, we prove that almost surely, with respect to the training data, our estimators (extensions of GCV and LOOCV) converge weakly to the true out-of-sample error distribution. This result requires mild assumptions on the response and feature distributions. We also establish a more general result that allows us to estimate certain functionals of the error distribution, both linear and nonlinear. This yields various applications, including consistent estimation of the quantiles of the out-of-sample error distribution, which gives rise to prediction intervals with asymptotically exact coverage conditional on the training data.

(A2) In Chapter 3, we develop a general framework for risk monotonization based on cross-validation that takes as input a generic prediction procedure and returns a modified procedure whose out-of-sample prediction risk is, asymptotically, monotonic in the limiting aspect ratio. As part of our framework, we propose two data-driven methodologies, namely zero- and one-step, that are akin to bagging and boosting, respectively, and show that, under very mild assumptions, they provably achieve monotonic asymptotic risk behavior. Our results are applicable to a broad variety of prediction procedures and loss functions, and do not require a well-specified (parametric) model. We exemplify our framework with concrete analyses of the minimum ℓ_2, ℓ_1-norm least squares prediction procedures. As one of the ingredients in our analysis, we also derive novel additive and multiplicative forms of oracle risk inequalities for split cross-validation that are of independent interest. In Chapter 4, we study the prediction risk of variants of bagged predictors in the proportional asymptotics regime, in which the ratio of the number of features to the number of observations converges to a constant. Specifically, we propose a general strategy to analyze prediction risk under squared error loss of bagged predictors using classical results on simple random sampling. Specializing the strategy, we derive the exact asymptotic risk of the bagged ridge and ridgeless predictors with an arbitrary number of bags under a well-specified linear model with arbitrary feature covariance matrices and signal vectors. Furthermore, we prescribe a generic cross-validation procedure to select the optimal subsample size for bagging and discuss its utility to mitigate the non-monotonic behavior of the limiting risk in the sample size (i.e., double or multiple descents). In demonstrating the proposed procedure for bagged ridge and ridgeless predictors, we thoroughly investigate oracle properties of the optimal subsample size, and provide an in-depth comparison between different bagging variants.

(A3) In Chapter 5, we revisit model complexity through the lens of model optimism and degrees of freedom. In particular, we first re-interpret degrees of freedom (a classical notion of complexity in statistics) in the fixed-X prediction setting, which allows us to extend this concept to the random-X prediction setting. We then define a family of complexity measures, whose two extreme ends we call the emergent and intrinsic degrees of freedom of a prediction model. We show the utility of our proposed measures through several example models, both linear and nonlinear, and illustrate how the proposed measures may prove useful to align the subtle multiple descents behavior in modern machine learning with the typical single descent behavior observed in traditional statistical prediction.

A word on organization and notation: Each chapter in the thesis is self-contained and can be read independently. The notation for each chapter is also self-contained. As a result, there may be some repetition in the definitions. The overall dissertation is an essuni[1] of the following (some finished, some ongoing) works of the author on the theme of this thesis, in primary capacity:

(W1) Uniform consistency of cross-validation estimators for high-dimensional ridge regression by Pratik Patil, Yuting Wei, Alessandro Rinaldo, Ryan J. Tibshirani.

(W2) Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression by Pratik Patil, Alessandro Rinaldo, Ryan J. Tibshirani.

(W3) Mitigating multiple descents: A model-agnostic framework for risk monotonization by Pratik Patil, Arun K. Kuchibhotla, Yuting Wei, Alessandro Rinaldo.

(W4) Bagging in overparameterized learning: Risk characterization and risk monotonization by Pratik Patil, Jin-Hong Du, Arun K. Kuchibhotla.

(W5) Revisiting model complexity in the wake of overparameterized learning by Pratik Patil, Ryan J. Tibshirani.

[1] This stands for essential union. And before you start googling, no, this is not a standard term. It is an attempt at some comedic relief through mathematical humor early on before we get serious from this point on, much to the author's dislike.

Chapter 1

Uniform consistency of cross-validation estimators

1.1 Introduction

Fitting high-dimensional statistical models typically requires some form of regularization, both for computational and statistical reasons. For optimization-based models, this can be achieved by adding to the data fitting objective function a tunable regularization term. The optimal level of regularization usually depends on unknown characteristics of the data generating distribution. In practice, one performs regularization tuning based on the observed data. Proper calibration of regularization can significantly affect the performance of the fitted model, and consequently proper data-dependent tuning is one of the core tasks in statistical learning.

Cross-validation (e.g., Allen, 1974; Stone, 1974; Geisser, 1975) is a widely used method for regularization tuning. While it has many variants, the most common variant is arguably k-fold cross-validation (e.g., Hastie et al., 2009; Györfi et al., 2006). Here we split the data into k "folds", leave out the first fold for model fitting so that we can use it to assess the out-of-sample performance of the fitted model, then we leave out the second fold, and so on. By aggregating the errors made across the k folds, we produce a final estimate of the expected out-of-sample error profile as a function of regularization level, and select the tuned regularization level by minimizing the cross-validated error profile.

While a typical choice of k is 5 or 10, such a choice of can suffer from high bias in high-dimensional problems. Setting $k = n$, the number of observations, leads to a variant called leave-one-out cross-validation (LOOCV). This alleviates the bias issues but it is computationally expensive in general, requiring n model fits. Despite recent important advances in the theoretical study of LOOCV and its various approximations in high dimensions (including Kale et al., 2011; Kumar et al., 2013; Meijer and Goeman, 2013; Obuchi and Kabashima, 2016; Miolane and Montanari, 2021; Wang et al., 2018a; Xu et al., 2019; Stephenson and Broderick, 2020a; Wilson et al., 2020; Celentano et al., 2020), the theoretical understanding of these methods, especially statistical properties of the tuned estimators under general distributional assumptions, is still incomplete.

In this work, we focus on ridge regression (Hoerl and Kennard, 1970b), a widely-used estimator in statistics that entails fitting linear regression with ℓ_2 regularization. We consider two commonly used cross-validation procedures, LOOCV and an approximation to LOOCV called generalized cross-validation (GCV) (Golub et al., 1979; Wahba, 1980, 1990). For ridge regression, both procedures can be computed efficiently—in a manner that requires no model refitting whatsoever—and are popular choices in practice. Our main goal is to investigate the theoretical behavior of ridge regression when tuned using one of these cross-validation methods.

For our theoretical analysis, we adopt a proportional asymptotic framework in which the number of features grows linearly with the number of observations (that is, their ratio converges to a constant). We show that both the GCV and LOOCV error curves, as functions of the ridge regularization parameter,

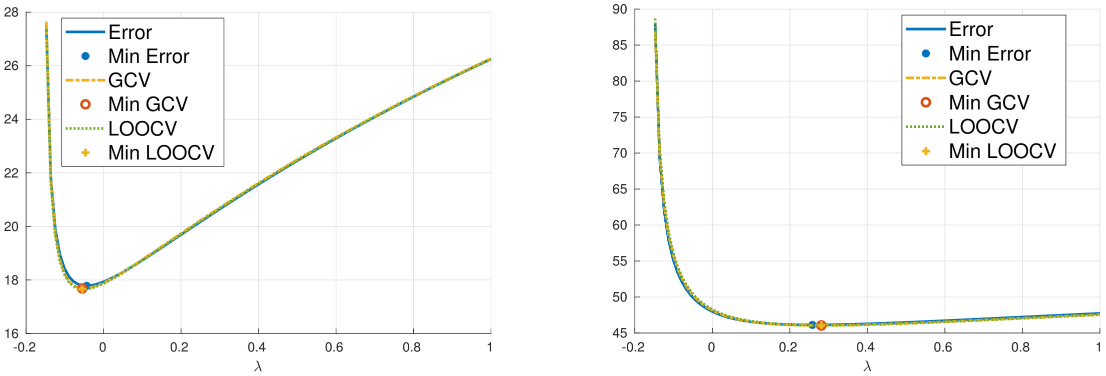


Figure 1.1: Comparison of the GCV and LOOCV estimates of the expected out-of-sample prediction error for ridge regression as a function of the regularization parameter λ. We consider an overparametrized regime where the number of observations is $n = 6000$ and the number of features is $p = 12000$. The features are random with a ρ-autoregressive covariance Σ (such that $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j) with $\rho = 0.25$. The response is generated from a linear model with a nonrandom signal vector β_0. In the left figure, the signal is aligned with the eigenvector corresponding to the largest eigenvalue of Σ, while in the right figure, the signal is aligned with the eigenvector corresponding to the smallest eigenvalue. The effective signal-to-noise ratio is set to $\beta_0^T \Sigma \beta_0 = 60$ to illustrate that, in the overparametrized regime, the optimal regularization could be negative or positive depending on how the signal aligns with the covariance eigenstructure. Note that in both the cases, the GCV and LOOCV curves track the prediction error over the whole range of λ very closely. The optimal regularization is recovered very well by the GCV and LOOCV estimates in both cases.

converge uniformly almost surely to the expected out-of-sample prediction error curve. Our results hold under weaker assumptions on the data generating distribution compared to others in the literature thus far, and provide a rigorous theoretical justification for the use of both GCV and LOOCV for regularization tuning for ridge regression in high dimensions. Below we summarize our main contributions, and illustrate key points with a numerical example in Figure 1.1.

1. **GCV pointwise convergence.** Given n i.i.d. samples from a standard linear model $y = x^T \beta_0 + \varepsilon$, where x is p-dimensional feature such that $x = \Sigma^{1/2} z$ for a covariance matrix Σ, and z contains i.i.d. entries, we establish limiting equivalence of the GCV estimator and the expected out-of-sample prediction error for ridge regression, under proportional asymptotics (p/n converging to a constant). This result holds for an arbitrary sequence of covariance matrices Σ with eigenvalues bounded away from zero and infinity, and an arbitrary sequence of signal vectors β_0 with bounded ℓ_2 norm.

2. **GCV uniform convergence.** Moreover, we show that this GCV convergence holds uniformly over compact intervals of the regularization parameter λ that include zero and negative regularization.

3. **LOOCV convergences.** We establish the analogous properties (pointwise and uniform convergence) for the LOOCV estimator by relating it to GCV.

4. **Optimal tuning.** As a direct consequence of uniform convergence, we demonstrate that the level of regularization chosen based on either of the GCV or LOOCV estimators almost surely delivers a limiting prediction accuracy that an oracle with full knowledge of the out-of-sample prediction error curve would achieve. Thus, in this sense, both methods are asymptotically optimal for tuning the prediction error of ridge regression.

1.2 Related work

Ridge error analysis. The predictive performance of ridge regression has been studied comprehensively in various settings, both asymptotic and non-asymptotic; see, e.g., Hsu et al. (2012); Karoui (2013); Dicker (2016); Dobriban and Wager (2018). More recently, there has been a surge of interest in understanding its prediction error driven by the successes of interpolating models in high dimensions; e.g., Hastie et al. (2022); Mei and Montanari (2022); Wu and Xu (2020); Richards et al. (2020); Tsigler and Bartlett (2020). Interestingly, Wu and Xu (2020); Richards et al. (2020) study the nature of optimal regularization and provide conditions on the feature covariance and signal structure that result in a positive or negative level of optimal regularization.

Ridge cross-validation. In the low-dimensional setting, the consistency of LOOCV and GCV for ridge regression error estimation and regularization tuning has been established in Stone (1974, 1977); Craven and Wahba (1979); Li (1985, 1986, 1987); Dudoit and van der Laan (2005), among others. More recently, statistical and computational aspects of cross-validation for regularized estimators in high dimensions have also been thoroughly studied; see, e.g., Beirami et al. (2017); Rad and Maleki (2020); Wang et al. (2018a); Xu et al. (2019); Rad et al. (2020); Austern and Zhou (2020).

Most similar to our work is probably the result of Hastie et al. (2022) on the asymptotic optimality of LOOCV and GCV tuning for ridge regression in high dimensions. These authors also adopt a proportional asymptotic model, but use stronger assumptions on the data generating distribution: they assume $\Sigma = I$ (independent features) and that the signal β_0 is drawn from a spherical prior (taking a Bayesian view). Under these conditions, the optimal level of regularization is always positive. We significantly generalize the scope of this analysis by allowing for *arbitrary* Σ and *nonrandom* β_0, in which case the optimal regularization level can be positive, negative, or zero.

Our work. We highlight the main contributions of our work below.

- **Analyzing differences.** We do not seek to characterize the limiting risk (we will use the terms risk and prediction error interchangeably), but instead, we analyze the limiting differences between the LOOCV and GCV estimators and the risk, and show that these differences tend to zero. As such, we are able to work in a general regime where it may not even be possible to precisely characterize the limiting risk in the first place.

- **Conditional statements.** Our theory is all conditional on the training data $\{(x_i, y_i)\}_{i=1}^n$ (results hold almost surely with respect to the draws from the training distribution). Most other papers provide cross-validation results that hold in an integrated sense over the training data. Our conditional setup allows for stronger statements about tuning based on the observed data rather than in an average sense.

- **Direct analysis of GCV.** Most previous papers rely on the stability of estimator in question to establish the properties of LOOCV, while we directly tackle the explicit forms of prediction error and GCV, and derive a crucial empirical equivalence lemma to first tie the risk to GCV, and then GCV to LOOCV.

- **Uniform convergence.** To analyze the cross-validation-tuned risks, we establish uniform convergence results, by leveraging the explicit form of the ridge estimator. This aspect has not been focused on in previous cross-validation work to the best of our knowledge, except Hastie et al. (2022).

- **Proof technique.** To reiterate what was mentioned earlier, in comparison to Hastie et al. (2022) (who take $\Sigma = I$ and β_0 drawn from a prior), we allow Σ and β_0 to be essentially arbitrary, only requiring Σ to have bounded eigenvalues and β_0 to have bounded ℓ_2 norm. While the flavor of final results is similar to those in Hastie et al. (2022), the proof techniques are different. We isolate the individual equivalences for the bias- and variance-like components in the GCV and LOOCV estimators, which helps shed light into the structure underlying the overall combined equivalence.

Further, we derive (and rely extensively on) an equivalence that relates certain functionals involving the sample covariance $\widehat{\Sigma}$ and population covariance Σ, in a proportional asymptotic setup. This is in a sense much simpler than the approach taken in Hastie et al. (2022), which relies on equating certain limiting formulae that arise from studying GCV, LOOCV, and ridge risk (equating such formulae involves difficult and unintuitive manipulations with Stieltjes transforms).

- **Result utility.** Recently, it has been observed that models with very small or even zero regularization can generalize well in certain overparametrized settings (e.g., Zhang et al., 2017; Belkin et al., 2019a). This is also the case with ridge regression where the optimal level of regularization can be zero or even negative (Kobak et al., 2020; Richards et al., 2020; Wu and Xu, 2020). Certain nontrivial interactions between the properties of the signal and feature distributions is what leads to these recent surprises. Our framework automatically accommodates these cases and affirms that that GCV and LOOCV can indeed pick risk-optimal interpolators when they need to.

1.3 Problem setup

We consider the standard regression setting in which we observe n i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ is the i^{th} feature vector and $y_i \in \mathbb{R}$ is the corresponding response variable. In matrix notation, we denote by $X \in \mathbb{R}^{n\times p}$ the feature matrix whose i^{th} row is x_i^T and by $y \in \mathbb{R}^n$ the response vector whose i^{th} entry is y_i.

Extended ridge regression. For a regularization parameter $\lambda > 0$, the ridge regression estimate $\widehat{\beta}_\lambda \in \mathbb{R}^p$ based on features X and response y can be formulated as the solution to the convex optimization problem

$$\underset{\beta \in {}^p}{\text{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

The can be explicitly written as

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{-1} X^T y/n,$$

where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix. To allow for an extended range of λ (including $\lambda = 0$), we simply define the extended ridge regression estimate as

$$\widehat{\beta}_\lambda = (X^T X/n + \lambda I_p)^{+} X^T y/n. \tag{1.1}$$

Here A^+ denotes the Moore-Penrose pseudoinverse of a matrix A. Note this definition allows for any $\lambda \in \mathbb{R}$. For $\lambda > 0$, there is no difference between (1.1) and the usual definition of ridge (second to last display). For $\lambda = 0$, we can see that (1.1) reduces to the least squares solution that lies in the row space of X, and hence has minimum ℓ_2 norm among all least squares solutions. Of particular interest is when $\operatorname{rank}(X) = n \le p$: then it reduces to the least squares solution that interpolates the data ($X\widehat{\beta}_\lambda = y$), and has minimum ℓ_2 norm among all such interpolators.

Prediction error. The expected out-of-sample prediction error (or risk) of the ridge model $\widehat{\beta}_\lambda$ is defined as

$$\operatorname{Err}(\widehat{\beta}_\lambda) = \mathbb{E}_{x_0, y_0}\Big[(x_0^T\widehat{\beta}_\lambda - y_0)^2 \mid X, y\Big]. \tag{1.2}$$

Here the expectation is taken with respect to the distribution of a new test pair (x_0, y_0) sampled from the same distribution as the training data $\{(x_i, y_i)\}_{i=1}^n$, and independent of the training data. The prediction error is a random variable (it is conditional on—and thus a function of—X, y) that quantifies how well a given fitted ridge model $\widehat{\beta}_\lambda$ performs in the task of predicting the response.

The prediction error as a function of the regularization parameter λ yields an error curve that we denote by

$$\operatorname{err}(\lambda) = \operatorname{Err}(\widehat{\beta}_\lambda).$$

As far as we are concerned in this work, the optimal regularization parameter is defined as the value that minimizes the risk curve $\operatorname{err}(\lambda)$. This is the value of λ that an oracle with knowledge of the risk curve would pick. We seek to construct a faithful estimate of the risk curve $\mathrm{err}(\lambda)$ based on the available data X and y, uniformly over λ, in order to select the regularization level that leads to prediction error close to that of the oracle prediction error. To do so, we will consider LOOCV and GCV whose definitions we recall next.

LOOCV and GCV. The LOOCV estimate for the risk of a given ridge model $\widehat{\beta}_\lambda$ is defined as

$$\mathrm{loo}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - x_i^T \widehat{\beta}_{-i,\lambda} \Big)^2,$$

where $\widehat{\beta}_{-i,\lambda} = (X_{-i}^T X_{-i}/n + \lambda I_p)^+ X_{-i}^T y_{-i}/n$ denotes the ridge estimate with the i^{th} observation pair (x_i, y_i) excluded from the training set. Computing the LOOCV estimate with this definition requires (re)fitting ridge model n times. Recall that ridge regression is a linear smoother, $X\widehat{\beta}_\lambda = L_\lambda y$, where the smoothing matrix $L_\lambda \in \mathbb{R}^{n \times n}$ is

$$L_\lambda = X(X^T X/n + \lambda I_p)^+ X^T/n. \tag{1.3}$$

Fortunately, there is a so-called shortcut formula for the LOOCV estimate (see, e.g., chapter 7 of Hastie et al., 2009):

$$\mathrm{loo}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right)^2, \tag{1.4}$$

where $[L_\lambda]_{ii}$ denotes the i^{th} diagonal element of L_λ.

The GCV estimate is a further convenient approximation to the LOOCV shortcut formula (1.4) given by

$$\mathrm{gcv}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \right)^2, \tag{1.5}$$

where $\mathrm{tr}[A]$ denotes the trace of a matrix A.

Caution needs to be taken when the smoothing matrix L_λ reduces to the identity matrix I_n, or in other words, ridge regression is an interpolator, with $X\widehat{\beta}_\lambda = y$. This happens when $\lambda = 0$ and X has rank n. In this case, both the numerators and denominators of $\mathrm{loo}(\lambda)$ and $\mathrm{gcv}(\lambda)$ are 0, however, we can define the corresponding LOOCV and GCV estimates as their respective limits as $\lambda \to 0$; see Hastie et al. (2022) for details.

Goal of this work. Our main goal is to analyze the differences between the cross-validation estimators of risk and the risk itself, $\mathrm{loo}(\lambda) - \mathrm{err}(\lambda)$ and $\mathrm{gcv}(\lambda) - \mathrm{err}(\lambda)$. Let $\lambda_I^\star$ denote the optimal oracle ride tuning parameter that minimizes $\mathrm{err}(\lambda)$ over an interval $I \subseteq \mathbb{R}$,

$$\lambda_I^\star = \operatorname*{arg\,min}_{\lambda \in I} \mathrm{err}(\lambda).$$

(If there are multiple minimizers, simply let $\lambda_I^\star$ denote one of them.) Similarly, let $\widehat{\lambda}_I^{\mathrm{gcv}}$ and $\widehat{\lambda}_I^{\mathrm{loo}}$ be the corresponding tuning parameters that minimize GCV and LOOCV over $\lambda \in I$. We seek to compare the prediction errors of the models tuned using GCV and LOOCV, $\mathrm{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\mathrm{gcv}}})$ and $\mathrm{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\mathrm{loo}}})$, against the prediction error under oracle tuning, $\mathrm{Err}(\widehat{\beta}_{\lambda_I^\star})$.

1.4 Main results

In this section, we state and discuss our main results. We first list the required assumptions in Section 1.4.1. In Section 1.4.2, we state the limiting equivalence between the GCV estimator and prediction risk, followed by the limiting equivalence between the LOOCV and GCV estimators in Section 1.4.3.

1.4.1 Assumptions

We begin by stating the assumptions we impose on the structure of response and feature distributions.

Assumption 1.1 (Response distribution)**.** There exists a signal vector $\beta_0 \in \mathbb{R}^p$ such that $y = X\beta_0 + \varepsilon$, where the noise vector $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ is independent of X, and its components are i.i.d. with mean 0, variance σ^2, and finite $4 + \eta$ moment for some $\eta > 0$.

Assumption 1.2 (Feature distribution)**.** The feature vectors (rows of X) can be decomposed as $x = \Sigma^{1/2} z$, where $\Sigma \in \mathbb{R}^{p \times p}$ is a deterministic positive definite matrix, and $z \in \mathbb{R}^p$ is a random vector whose components are i.i.d. with mean zero 0, variance 1, and finite $4 + \eta$ moment for some $\eta > 0$.

We consider a proportional asymptotic framework in which the number of features p grows with the number of observations n in such a way that their ratio p/n approaches a constant $\gamma \in (0, \infty)$. Accordingly, in our asymptotic analysis, we must deal with a sequence of feature covariance matrices Σ and signal vectors β_0. (For ease of readability, we do not make the dependence of these quantities and many others on p explicit in our notation.) We make the following assumptions on the eigenvalues of Σ and the signal energy.

Assumption 1.3 (Extreme eigenvalues of Σ)**.** The maximum and minimum eigenvalues of Σ are upper and lower bounded by constants $r_{\max} < \infty$ and $r_{\min} > 0$, respectively, independent of p.

The lower bound $r_{\min}$ on the minimal eigenvalue of Σ will determine, asymptotically, the smallest possible value of the regularization parameter for which our results hold. We denote it by $\lambda_{\min} = -(\sqrt{\gamma} - 1)^2 r_{\min}$.

Assumption 1.4 (Signal energy)**.** The signal energy $\|\beta_0\|_2^2$ is upper bounded by a constant $\tau < \infty$ independent of p.

We note that it should be possible to relax the assumptions on the maximum and minimum eigenvalues of Σ, to allow a certain fraction of eigenvalues to diverge and others to accumulate near zero. We leave such an extension to future work.

1.4.2 GCV versus prediction error

We are ready to state our first result comparing the GCV estimator to prediction error of ridge regression.

Theorem 1.4.1 (GCV equals prediction error in limit)**.** *Under Assumptions 1.1 to 1.4, for every $\lambda \in (\lambda_{\min}, \infty)$, it holds that*

$$\mathrm{gcv}(\lambda) - \mathrm{err}(\lambda) \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Furthermore, the convergence is uniform in λ over compact subintervals $I \subseteq (\lambda_{\min}, \infty)$; consequently, for any such interval I,

$$\mathrm{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\mathrm{gcv}}}) - \mathrm{Err}(\widehat{\beta}_{\lambda_I^\star}) \xrightarrow{\text{a.s.}} 0,$$

where $\widehat{\lambda}_I^{\mathrm{gcv}}$ and $\lambda_I^\star$ are the corresponding optimal GCV and prediction error tuning parameters, respectively.

We note that in this and in all the other asymptotic statements in this work, the almost sure qualification refers to the randomness in both X and y.

Range of λ. The lower limit $\lambda_{\min}$ in Theorem 1.4.1 is used to ensure that the resulting smoothing matrix L_λ stays positive semidefinite; this is simply a function of the behavior of the minimum non-zero eigenvalue of the sample covariance matrix $\widehat{\Sigma}$ (see Bai and Silverstein, 1998).

Note that this range of λ allows for potentially negative regularization (when $\gamma \neq 1$), including zero; the latter case, in particular, results in the least squares interpolator when $p > n$. The fact that GCV works in this case is interesting because both the numerator and denominator in the expression (1.5) for $\mathrm{gcv}(\lambda)$ are 0—implying the particular form of the ridge estimator somehow preserves the information about the predictive performance in the GCV limit even when the training error is 0.

The statement in Theorem 1.4.1 does not cover the behavior of GCV at the endpoints $\lambda = \lambda_{\min}$ and $\lambda \to \infty$. In fact, it is easy to check that the limiting behavior of GCV and prediction error matches at these endpoints as well. In particular, under the same assumptions as the theorem, if $r_{\min}$ is the limit inferior of minimum eigenvalues of the Σ sequence, then indeed both

$$\operatorname{gcv}(\lambda_{\min}) \to \infty \quad \text{and} \quad \operatorname{err}(\lambda_{\min}) \to \infty$$

as $n, p \to \infty$ with $p/n \to \gamma$. Similarly, both

$$\operatorname{gcv}(\lambda) \to c^2 \quad \text{and} \quad \operatorname{err}(\lambda) \to c^2$$

as $\lambda \to \infty$ and $n, p \to \infty$ with $p/n \to \gamma$, where $c^2 = \mathbb{E}[y_0^2]$ is the prediction error of the null estimator. In this regard, the pointwise equivalence between GCV and prediction error extends to the entire range of λ.

1.4.3 LOOCV versus GCV

As a byproduct of our analysis, we establish a limiting equivalence between the LOOCV and GCV estimators. This implies a limiting equivalence between LOOCV and prediction error.

Theorem 1.4.2 (LOOCV equals GCV in limit)**.** *If the components of the response vector $y \in \mathbb{R}^n$ have mean zero and finite second moment, and Assumptions 1.2 to 1.3 hold, then for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\operatorname{loo}(\lambda) - \operatorname{gcv}(\lambda) \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Furthermore, the convergence is uniform in λ over compact subintervals $I \subseteq (\lambda_{\min}, \infty)$.

It is worth pointing out that, compared to Theorem 1.4.1, the last guarantee only requires that the response variables have a finite second moment. In particular, it does not postulate a linear model. So the equivalence between the GCV and LOOCV estimators holds even when the model is misspecified.

In general, the analysis of LOOCV is challenging because of complex dependencies between its summands. Fortunately, for ridge regression, the equivalent shortcut expression given in (1.4) for the LOOCV estimate simplifies such dependence. Unlike GCV in (1.5), which weights training errors by $1 - \operatorname{tr}[L_\lambda]/n$, the shortcut expression for LOOCV weights the i^{th} training error by $1 - [L_\lambda]_{ii}$. Theorem 1.4.1 effectively shows that this different reweighting does not affect the limiting behavior, providing a way to directly tie GCV to LOOCV.

An important consequence of the last theorem is the following.

Corollary 1.4.3 (LOOCV equals prediction error in limit)**.** *Under the assumptions as Theorem 1.4.1, the same results hold but for LOOCV in place of GCV.*

(The same remarks about the range of λ that were made following the GCV theorem also apply here.)

In light of this corollary, we conclude that both the GCV and the LOOCV estimators are uniformly close to the true risk in the limit. Thus regularization tuning using either method will be asymptotically optimal for ridge regression.

1.5 Proof outlines

In this section, we outline the main ideas behind the proofs of Theorem 1.4.1 and Theorem 1.4.2. The complete proofs are provided in the supplement.

1.5.1 GCV versus prediction error

The proof of Theorem 1.4.2 involves two steps. In the first step, we decompose both the prediction error and the GCV estimator into asymptotic bias- and variance-like components as summarized in Lemma 1.5.1 and Lemma 1.5.2. In the second step, we establish limiting equivalences for both the bias and variance components as summarized in Lemma 1.5.3 and Lemma 1.5.4. The key reason why the limiting bias-variance equivalences hold is a certain property obeyed by the denominator of GCV as elucidated in Lemma 1.5.5.

Prediction error decomposition. We begin with a familiar asymptotic bias-variance decomposition for the prediction risk. For convenience, let $\widehat{\Sigma} = X^T X/n$ denote the sample covariance matrix. Also, define bias- and variance-like components as follows:

$$\mathrm{err}_b(\lambda) = \beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0,$$

$$\mathrm{err}_v(\lambda) = \frac{\varepsilon^T}{\sqrt{n}}\left(\frac{X(\widehat{\Sigma} + \lambda I_p)^+ \Sigma (\widehat{\Sigma} + \lambda I_p)^+ X^T}{n}\right)\frac{\varepsilon}{\sqrt{n}} + \sigma^2.$$

The decomposition of the prediction error can now be summarized as follows.

Lemma 1.5.1 (Error bias-variance decomposition)**.** *Under Assumptions 1.1 to 1.4, for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}(\lambda) - \mathrm{err}_b(\lambda) - \mathrm{err}_v(\lambda) \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ and $n/p \to \gamma \in (0, \infty)$.

GCV decomposition. We decompose GCV into terms that mimic the bias- and variance-like terms in the decomposition for the risk. For $\lambda \neq 0$, define GCV bias- and variance-like components as follows:

$$\mathrm{gcv}_b(\lambda) = \beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0,$$

$$\mathrm{gcv}_v(\lambda) = \frac{\varepsilon^T}{\sqrt{n}}\left(I_n - \frac{X(\widehat{\Sigma} + \lambda I_p)^+ X^T}{n}\right)^2 \frac{\varepsilon}{\sqrt{n}}.$$

Additionally, write the GCV denominator as:

$$\mathrm{gcv}_d(\lambda) = \big(1 - \mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+]/n\big)^2.$$

When $\lambda = 0$, the corresponding quantities after taking the limit $\lambda \to 0$ take the form:

$$\mathrm{gcv}_b(0) = \beta_0^T \widehat{\Sigma}^+ \beta_0,$$

$$\mathrm{gcv}_v(0) = \frac{\varepsilon^T}{\sqrt{n}}\big(\widehat{\Sigma}^+\big)^2 \frac{\varepsilon}{\sqrt{n}},$$

$$\mathrm{gcv}_d(0) = \big(\mathrm{tr}[\widehat{\Sigma}^+]/n\big)^2.$$

(We remark that the limiting expressions for the bias- and variance-like components and the denominator for the $\lambda = 0$ case can alternately be written in terms of the gram matrix XX^T/n. The representation in terms of the sample covariance matrix $\widehat{\Sigma}$ is for consistency with the $\lambda \neq 0$ case.)

Next we establish the decomposition of GCV into bias- and variance-like quantities.

Lemma 1.5.2 (GCV bias-variance decomposition)**.** *Under Assumptions 1.1 to 1.4, for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{gcv}(\lambda) - \frac{\mathrm{gcv}_b(\lambda) + \mathrm{gcv}_v(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Bias-variance equivalences. The two bias terms $\mathrm{err}_b(\lambda)$ and $\mathrm{gcv}_b(\lambda)$ differ in the sense that the latter has the unknown Σ replaced by its natural plug-in estimator $\widehat{\Sigma}$ and a rescaling by the denominator $\mathrm{gcv}_d(\lambda)$. The difference between the variance terms is analogous, albeit slightly more involved. For both, the denominator adjustment, which can be thought of as a correction for optimism in the training error by the number of effective degrees of freedom, turns out to be critical. Indeed, it is only through this normalization that $\mathrm{gcv}_b(\lambda)$ and $\mathrm{gcv}_v(\lambda)$ become consistent estimators of their population counterparts, as summarized next and illustrated in Figure 1.2.

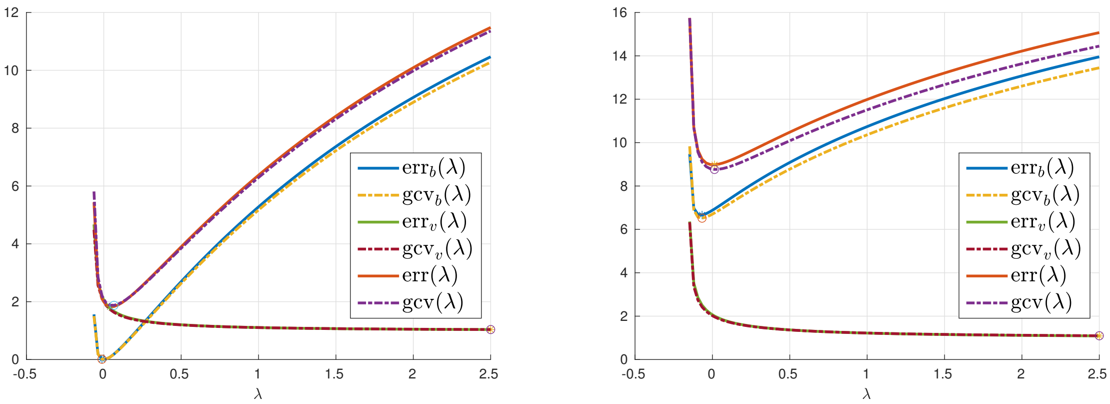


Figure 1.2: Comparison of the bias and variance decompositions of the GCV estimate and the prediction error. Similar to Figure 1.1, the features are random from a ρ-autoregressive covariance matrix Σ with $\rho = 0.25$. The response is generated from a linear model where the signal is nonrandom and aligned with the principal eigenvector of Σ. The effective signal-to-noise ratio is $\beta_0^T \Sigma \beta_0 = 25$. The left figure illustrates an underparametrized regime (with $n = 6000$ and $p = 3000$ such that $\gamma = 0.5$) while the right illustrates an overparametrized regime (with $n = 6000$ and $p = 12000$ such that $\gamma = 2$). In both cases, the bias-variance-like components of the GCV risk estimate track the bias-variance components in the prediction risk over the entire range of λ very well. In the underparametrized regime, the bias of the prediction risk is 0 at $\lambda = 0$ and increases on either sides when $\lambda \neq 0$, while the variance always decreases as λ increases (from the most negative allowed λ), resulting in a positive optimal regularization. On the other hand, in the overparametrized regime, the bias is no longer minimized at $\lambda = 0$, but at a negative λ, while the variance is again a decreasing function of λ. Since the bias dominates the total prediction risk, it results in negative optimal regularization.

Lemma 1.5.3 (Bias equivalence)**.** *Under Assumptions 1.2 to 1.4, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}_b(\lambda) - \frac{\mathrm{gcv}_b(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Lemma 1.5.4 (Variance equivalence)**.** *Under Assumptions 1.1 to 1.3, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}_v(\lambda) - \frac{\mathrm{gcv}_v(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Basic GCV equivalence. At the heart of why the rescaling in the GCV bias and variance-like terms yields consistency is a certain asymptotic equivalence of random matrices as summarized below.

Lemma 1.5.5 (Basic GCV equivalence)**.** *Under Assumption 1.2 and Assumption 1.3, for any sequence of matrices $B_p \in \mathbb{R}^{p \times p}$ that are bounded in trace norm (independent of p), and for $\lambda \in (\lambda_{min}, \infty) \setminus \{0\}$, it holds that*

$$\mathrm{tr}\left[B_p\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big)\Sigma\right] - \frac{\mathrm{tr}\left[B_p\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big)\widehat{\Sigma}\right]}{1 - \mathrm{tr}[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. When $\lambda = 0$,

$$\mathrm{tr}\left[B_p(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma\right] - \frac{\mathrm{tr}\left[B_p\widehat{\Sigma}^{+}\widehat{\Sigma}\right]}{\mathrm{tr}[\widehat{\Sigma}^{+}]/n} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Finally, to prove uniform convergence in λ, we show that both the prediction risk $\mathrm{err}(\lambda)$ and the GCV estimator $\mathrm{gcv}(\lambda)$, and their derivatives, as functions of λ, are uniformly bounded over compact subintervals of $(\lambda_{\min}, \infty)$. This yields equicontinuity of the family of functions $\lambda \to \mathrm{err}(\lambda)$ and $\lambda \to \mathrm{gcv}(\lambda)$ almost surely and the result then follows from an application of the Arzela-Ascoli theorem. The uniform convergence subsequently leads to the convergence of the tuned errors.

1.5.2 LOOCV versus GCV

There are two steps involved in establishing the limiting equivalence between LOOCV and GCV. The first is to show that the LOOCV estimator in the limit is equal to a scalar corrected factor of the training error. The second is that the correction happens to match with the factor that appears in the GCV estimator in the limit. The following lemma provides the LOOCV limit.

Lemma 1.5.6 (LOOCV limit as rescaled train error)**.** *If the components of the response $y \in \mathbb{R}^n$ have mean zero and finite second moment, and Assumptions 1.2 to 1.3 hold, then for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{loo}(\lambda) - \Big(1 + \mathrm{tr}[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma]/n\Big)^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\widehat{\beta}_\lambda)^2 \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

The limiting equivalence then follows by tying the scale factor in the GCV estimator to the scale factor in the limiting LOOCV using an instantiation of Lemma 1.5.5.

1.6 Discussion

In this work, we established uniform consistency of the GCV and LOOCV estimators for ridge regression prediction error under a proportional asymptotic framework. At a high level, the key reason why the limiting equivalences hold is a certain asymptotic equivalence of random matrices, where on one side we have a quantity that involves both the feature covariance Σ and the sample covariance $\widehat{\Sigma}$, while on the other side, we have a quantity that only involves $\widehat{\Sigma}$, appropriately normalized. That is,

$$(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma \asymp \frac{(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}}{1 - \mathrm{tr}[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}]/n}$$

where for two sequences of matrices A_p and B_p, $A_p \asymp B_p$ is used to mean that $\lim_{n\to\infty} \mathrm{tr}[C_p A_p] - \mathrm{tr}[C_p B_p] = 0$ almost surely for any sequence of deterministic matrices C_p of bounded trace norm.

A similar notion of equivalence has appeared in the random matrix theory literature (e.g., Serdobolskii, 1983; Silverstein and Choi, 1995; Hachem et al., 2007; Ledoit and Péché, 2011; Rubio and Mestre, 2011; Couillet and Debbah, 2011), and recently, has been utilized and developed further in Dobriban and Sheng (2021, 2020). Our work takes a slightly differently approach in that, instead of expressing the resolvents in terms of limits of unknown population quantities (which has been called a *deterministic equivalence*), we relate two sets of resolvents, neither of which needs to have a computable asymptotic limit in the first place.

For statistical applications, we believe this could have broad utility because it allows to tie potentially interesting out-of-sample quantities to purely data-dependent quantities. For example, it should be possible to asymptotically equate more general functionals involving Σ and $\widehat{\Sigma}$ in terms of $\widehat{\Sigma}$ alone. Exploring such connections for both a wider class of statistical problems and for metrics other than the expected out-of-sample error is a future direction.

Beyond asymptotics, it is also of interest to carry out a finite sample analysis that explicitly reveals how the interaction between the signal vector and the feature covariance affects rates of convergence. This may, for example, facilitate constructions of confidence intervals for the tuned parameters. It may also reveal that GCV and LOOCV—though consistent across a very broad set of problem settings, as demonstrated in this work—can struggle in terms of their speed of convergence for certain problems, like (say) when the optimal regularization parameter is around zero. Finally, the assumptions on the feature and response distribution should be able to be relaxed; pursuing minimal assumptions that allow for equivalences is of general interest.

Chapter 2

Estimating functionals of out-of-sample error distribution

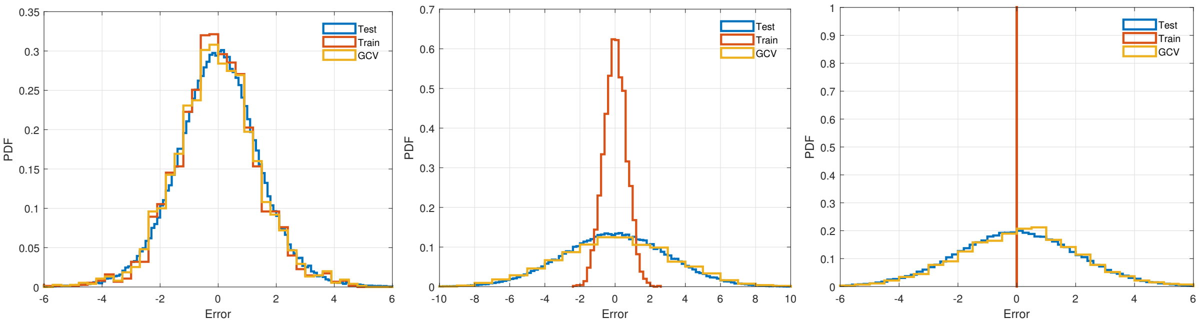
2.1 Introduction

The out-of-sample error associated with a predictive model is the difference between the true (unobserved) response and the predicted response at a new draw from the feature distribution. Being able to accurately estimate functionals of the out-of-sample error distribution is of critical importance in practice, both for model assessment and model selection purposes. By far the most common functional considered is the uncentered second moment of this error distribution—the mean squared error of the predictive model. Estimating this quantity has been the focus of many decades of research in the statistics and machine learning communities, which has yielded numerous advances in both theory and methodology. A central method in practice for estimating the mean squared prediction error is cross-validation (CV), which comes in many variants, including *generalized* and *leave-one-out* cross-validation (GCV and LOOCV, respectively). Classic references on CV include Allen (1974); Stone (1974, 1977); Geisser (1975); Golub et al. (1979); Wahba (1980, 1990); Li (1985, 1986, 1987). See Arlot and Celisse (2010) for a general review of CV.

In this work, we study the problem of estimating the entire out-of-sample error distribution. Part of reason why so much past work in risk estimation has focused on mean squared out-of-sample error is undoubtedly the special analytical structure that it affords and the associated bias-variance decomposition. A main goal of this work is to understand what other functionals of the out-of-sample error distribution can be reliably estimated using cross-validation. Such an understanding is useful for not only theoretical purposes (necessitating novel proof techniques to analyze generic functionals), but practical ones as well, since cross-validation estimators that work under such general settings then open up the possibility of employing a wider range of metrics for model evaluation and selection, which may be informative for the data analyst in any given problem setting at hand.

Throughout, we will focus on *ridge regression* (Hoerl and Kennard, 1970b,a) for the predictive model, a special form of Tikhonov regularization (Tikhonov, 1943, 1963), which is very widely used in statistics and machine learning. We choose to focus on ridge regression because GCV and LOOCV admit special forms for this estimator, and also because ridge has recently attracted much attention—especially in the limiting case of zero regularization, often called the "ridgeless" limit—due to its somewhat exotic behavior in the overparametrized regime (see, e.g., Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2022; Muthukumar et al., 2020, and references therein). Importantly, it has been recently shown that the ridgeless (minimum ℓ_2 norm) interpolator can be optimal for mean squared out-of-sample error, among all ridge models, for well-specified linear models with certain data geometries and high signal-to-noise ratios (Wu and Xu, 2020; Richards et al., 2020). This has been corroborated empirically using real data sets for ridge regression (Kobak et al., 2020) and kernel ridge regression (Liang and Rakhlin, 2020). Thus, providing theory that covers that ridgeless case is both of foundational and practical importance.

Before summarizing our main contributions, we give some empirical examples in Figure 2.1 to motivate our study.



(a) Low dimension ($p/n = 0.04$) (b) Moderate dimension ($p/n = 0.8$) (c) High dimension ($p/n = 2$)

Figure 2.1: A simulation with $n = 2500$ samples and $p \in \{100, 2000, 5000\}$ features (a different p per panel above). In each setting, we generated the feature vectors x_i to have independent components from a t-distribution with 5 degrees of freedom, and generated the responses y_i by adding t-distributed noise with 5 degrees of freedom to a nonlinear (quadratic) function of x_i. We then fit the minimum ℓ_2 norm least squares solution, as in (1.1) with $\lambda = 0$. The blue curve in each panel is a histogram of the true prediction error distribution, computed from 10^5 independent test samples. The red curve is a histogram of the training errors; when $p > n$, this is just a point mass at zero. The yellow curve is a histogram of GCV-reweighted training errors, as in (2.11) (for $p < n$, in the first two panels) and (2.13) (for $p > n$, in the last panel). This tracks the blue curve very well in all settings. Empirical results for LOOCV are given in the supplement.

2.1.1 Summary of contributions

An overview of our main contributions is as follows.

- We define natural extensions of GCV and LOOCV in order to estimate the out-of-sample prediction error distribution associated with ridge regression. These are empirical distributions over reweighted training errors (where the reweighting is tied to GCV or LOOCV).
- Under an asymptotic framework where the feature dimension p and sample size n grow proportionally, $p/n \to \gamma \in (0, \infty)$, we prove that, almost surely with respect to the training data, these extensions of GCV and LOOCV converge weakly to the true out-of-sample error distribution of ridge regression. This result requires mild assumptions; we do not need the true regression model to be linear.
- The GCV and LOOCV extensions and the theory we prove about them all accommodate the choice of zero (or even negative) ridge regularization in high dimensions, where $p > n$.
- For certain linear functionals of the error distribution P, which take the form $\int t \, dP$ for a function t, we prove that suitable plug-in estimators (based on the GCV and LOOCV estimators of the entire error distribution) are asymptotically consistent, almost surely. This result requires t to satisfy certain continuity and growth conditions, but it can be unbounded.
- Finally, we use a uniform convergence argument to handle certain nonlinear functionals of the error distribution (that can be written in a variational form involving linear functionals). This allows us to consistently estimate, as an application, quantiles of the ridge error distribution.

2.1.2 Related work

Among the different CV variants to assess prediction accuracy, k-fold CV is widely used in practice (Györfi et al., 2006; Hastie et al., 2009). However, in a high-dimensional regime where the feature dimension p is comparable to the sample size n, small values of k (such as $k = 5$ or 10) lead to bias in error estimation (see, e.g., Rad and Maleki, 2020). LOOCV (where $k = n$) mitigates these bias issues, and consequently LOOCV and various approximations to it (that circumvent its computational burden) have been of interest in recent work, including Meijer and Goeman (2013); Liu et al. (2014); Obuchi and Kabashima (2016); Beirami et al. (2017); Wang et al. (2018b); Stephenson and Broderick (2020a); Giordano et al. (2019); Wilson et al. (2020); Rad et al. (2020); Xu et al. (2021). For recent results on ridge regression in particular, where LOOCV can be done efficiently via a "shortcut" formula, see Patil et al. (2021).

On the inferential side, Bayle et al. (2020) prove central limit theorems for CV error and a derive a consistent estimator of its asymptotic variance under certain stability assumptions, similar to Kale et al. (2011); Kumar et al. (2013); Celisse and Guedj (2016). Their results yield asymptotic confidence intervals for the prediction error and apply to k-fold CV (for a fixed k) as well as LOOCV. See also Austern and Zhou (2020) for similar guarantees. A prominent and distinctive aspect of our work compared to these works and others is the focus on properties of the entire empirical distribution of the CV errors, rather than specific functionals such as the mean squared CV error.

In a contribution that is quite relevant to this work, Steinberger and Leeb (2016, 2018) construct prediction intervals from quantiles of the empirical distribution of the LOOCV errors and provide conditional coverage guarantees, which hold in expectation. Their key assumptions are algorithmic stability, as in Bousquet and Elisseeff (2002), along with a bound in probability on the prediction error at a new test point. Under a more restrictive asymptotic regime in which $p/n \to \gamma < 1$, they show that the Kolmogorov-Smirnov distance between the empirical distribution of LOOCV errors and the conditional prediction error distribution vanishes in expectation. This general result is then applied to yield corollaries for various predictive models, including ridge regression, by leveraging model-specific stability and error results from the literature.

In comparison, our work focuses on ridge regression alone, but we deliver stronger and broader guarantees. To be specific, our results (1) accommodate the high-dimensional regime, $p/n \to \gamma \geq 1$; (2) assume quite weak conditions on the data (e.g., we do not require a well-specified linear model); (3) hold uniformly over the choice of regularization parameter (which includes no regularization—the ridgeless limit); (4) yield not only consistent estimation of the prediction error distribution itself, but of a broad class of functionals of this distribution (which includes unbounded and nonlinear ones); and (5) produces guarantees that hold almost surely—rather than in expectation or in probability—with respect to the training data.

2.2 Preliminaries

We adopt a standard regression setting, with i.i.d. samples (x_i, y_i), for $i = 1, \dots, n$, where each $x_i \in \mathbb{R}^p$ is a feature vector and $y_i \in \mathbb{R}$ is its corresponding response value. We will denote by $X \in \mathbb{R}^{n \times p}$ the feature matrix whose i^{th} row is $x_i^\top$, and by $y \in \mathbb{R}^n$ the response vector whose i^{th} entry is y_i.

2.2.1 Ridge regression

The *ridge regression* estimator $\widehat{\beta}_\lambda \in \mathbb{R}^p$, based on X, y, is defined as the solution to the following problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Here λ is a regularization parameter. When $\lambda > 0$, the above optimization problem is strictly convex and has a unique solution:

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^{-1} X^\top y/n.$$

When $\lambda = 0$, and $X^\top X$ is rank deficient (which will always be the case when $p > n$), there will be infinitely many solutions, and we focus on the solution with the minimum ℓ_2 norm, which we refer to as the *min-norm solution* for short. By defining the ridge estimator as

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^\dagger X^\top y/n, \tag{2.1}$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix A, we simultaneously accommodate the case of $\lambda > 0$, in which case (2.1) reduces to the second to last display, and the case of $\lambda = 0$, in which case (2.1) becomes the min-norm solution (it lies in the column space of $(X^\top X)^\dagger$, i.e., the row space of X, so it has the minimum ℓ_2 norm among all least squares solutions). In fact, the above display even accommodates the case of $\lambda < 0$, in which case (2.1) remains well-defined.

The case of zero regularization is of particular interest when $\mathrm{rank}(X) = n$, because then any least squares solution interpolates the training data, and the min-norm solution $\widehat{\beta}_0$ (by construction) has the minimum ℓ_2 norm among all such interpolators.

2.2.2 Out-of-sample error

Let (x_0, y_0) denote a test point drawn independently from the same distribution as the training data (x_i, y_i), $i = 1, \dots, n$, and denote the out-of-sample prediction error of ridge regression at tuning parameter λ by

$$e_\lambda = y_0 - x_0^\top \widehat{\beta}_\lambda. \tag{2.2}$$

This is a scalar random variable, and we denote by P_λ its distribution conditional the training data:[1]

$$P_\lambda = \mathcal{L}\big(e_\lambda \mid X, y\big). \tag{2.3}$$

We are interested in estimating P_λ using the training data. A naive estimator would be to use the empirical distribution over the training errors expressed as

$$\widehat{P}_\lambda = \frac{1}{n} \sum_{i=1}^n \delta\big(y_i - x_i^\top \widehat{\beta}_\lambda\big). \tag{2.4}$$

Here we use $\delta(z)$ for a point mass at z. Of course, this can be very inaccurate in high dimensions (as we saw in Figure 2.1); at the extreme case of $\mathrm{rank}(X) = n$ and $\lambda = 0$, the naive estimator $\widehat{P}_\lambda$ trivially places all mass at zero. In the next subsection, we will introduce more sensible estimators based on cross-validation.

Aside from estimating P_λ itself, we may be interested in estimating a particular *functional* of P_λ, denoted by $\psi(P_\lambda)$. Recall, a functional ψ acting on distributions is such that $P \mapsto \psi(P) \in \mathbb{R}$ for all distributions P.

In the context of the out-of-sample error distribution P_λ, the most common functional of interest is its uncentered second moment,

$$\psi(P_\lambda) = \int z^2 \, dP_\lambda(z) = \mathbb{E}\big[e_\lambda^2 \mid X, y\big],$$

which is simply the mean squared prediction error. We will consider general linear functionals of the form

$$\psi(P_\lambda) = \int t(z) \, dP_\lambda(z) = \mathbb{E}\big[t(e_\lambda) \mid X, y\big], \tag{2.5}$$

for functions t (possibly nonlinear and unbounded, but subject to certain continuity and growth conditions). We will also consider certain nonlinear functionals such as the level-τ quantile, for $\tau \in (0, 1)$:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\}, \tag{2.6}$$

where F_λ denotes the cumulative distribution function (CDF) of P_λ.

2.2.3 Cross-validation

GCV and LOOCV are two popular versions of cross-validation that are used to estimate the mean squared prediction error. GCV is traditionally defined for linear smoothers only, but LOOCV is fully general: it

[1] To be clear, P_λ is itself a random quantity, because it depends on the training data X, y. However, we suppress this dependence notationally, for simplicity.

applies to any predictive model. In order to describe the details for ridge regression, we introduce the notation:

$$L_\lambda = X(X^\top X/n + \lambda I_p)^\dagger X^\top/n, \tag{2.7}$$

for the ridge smoother matrix at regularization level λ. Thus, by definition, we can express the fitted values (predicted values at the training points x_i, $i = 1, \dots, n$) from ridge regression as $X\widehat{\beta}_\lambda = L_\lambda y$.

The LOOCV estimate for the mean squared prediction error of a given ridge model $\widehat{\beta}_\lambda$ can now be written as

$$\frac{1}{n}\sum_{i=1}^n \left(y_i - x_i^T \widehat{\beta}_{-i,\lambda}\right)^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)^2, \tag{2.8}$$

where $\widehat{\beta}_{-i,\lambda}$ denotes the ridge estimate when the i^{th} pair (x_i, y_i) is excluded from the training data set, and $[L_\lambda]_{ii}$ denotes the i^{th} diagonal element of L_λ. The left-hand side in (2.8) is the usual definition of LOOCV for any predictive model; the right-hand side is a so-called "shortcut" formula that holds for ridge (and a handful of other special linear smoothers; see, e.g., Chapter 7 of Hastie et al., 2009).

The GCV estimate for the mean squared error of $\widehat{\beta}_\lambda$ is given by

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right)^2, \tag{2.9}$$

where $\mathrm{tr}[A]$ denotes the trace of a matrix A.

Caution needs to be taken in (2.8) and (2.9) when $\lambda = 0$ and $\mathrm{rank}(X) = n$, in which case $L_\lambda = I_n$, and both of the numerators and denominators in every summand of (2.8), (2.9) are zero. To avoid this problem we redefine them by their respective limits as $\lambda \to 0$, which gives (see the supplement for details):

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)^2 \text{ and } \frac{1}{n}\sum_{i=1}^n \left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right)^2, \tag{2.10}$$

for LOOCV and GCV, respectively.

2.2.4 Proposed estimators

We propose estimators for the out-of-sample prediction error distribution P_λ in (2.3), building off the empirical distributions of reweighted training errors, inspired by GCV in (2.9) and LOOCV in (2.8). Precisely, we define

$$\widehat{P}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right), \tag{2.11}$$

which we refer to as the GCV estimate of the out-of-sample error distribution, and

$$\widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right), \tag{2.12}$$

which we refer to as the LOOCV estimate of the out-of-sample error distribution.

When $\lambda = 0$ and $\mathrm{rank}(X) = n$, the above expressions are ill-defined, and we redefine them based on the forms of GCV and LOOCV in (2.10):

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right), \tag{2.13}$$

$$\widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^n \delta\left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right). \tag{2.14}$$

To estimate a generic functional of $\psi(P_\lambda)$ of the error distribution, we simply use

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}). \tag{2.15}$$

For $\psi(P_\lambda) = \int z^2 \, dP_\lambda(z)$, the plug-in estimates above reduce to the standard GCV and LOOCV estimates of the mean squared prediction error.

2.3 Distribution estimation

We first cover distributional convergence results. We impose the following mild structural and moment assumptions on the feature and response distributions.

Assumption 2.1 (Feature distribution)**.** Each feature vector can be decomposed as $x_i = \Sigma^{1/2} z_i$, for a deterministic symmetric matrix $\Sigma \in \mathbb{R}^{p \times p}$ whose maximum eigenvalue is bounded above by $r_{\max} < \infty$, and minimum eigenvalue is bounded below by $r_{\min} > 0$, where $r_{\max}$ and $r_{\min}$ are constants, and for a random vector $z_i \in \mathbb{R}^p$ whose entries are i.i.d. with mean zero, unit variance, and $\mathbb{E}[|z_{ij}|^{4+\mu}] \le M_z < \infty$, where $\mu > 0$ and M_z are constants.

The maximum eigenvalue bound for the feature covariance matrix Σ is used to control the magnitude of ridge predictions; the minimum eigenvalue bound is used in the analysis of the min-norm interpolator. Both of these can be relaxed further for some of our results, but we do not pursue such refinements here.

Assumption 2.2 (Response distribution)**.** Each y_i has mean zero and satisfies $\mathbb{E}[|y_i|^{4+\nu}] \le M_y < \infty$, where $\nu > 0$ and M_y are constants.

The condition that each y_i is centered is only used for simplicity. When y_i does not have mean zero, we would simply include an intercept in the model defined in (2.1), and all of our results would translate accordingly.

We work in an asymptotic regime where the number the samples n and the number of features p both diverge to ∞, and yet their ratio p/n converges to $\gamma \in (0, \infty)$. Such asymptotic regime has received considerable attention recently in high-dimensional statistics and machine learning theory, which is commonly referred to as proportional asymptotics. The range of regularization parameter values λ over which our results will hold is a function of γ and $r_{\min}$. In preparation for the coming theorem statements, we define $\lambda_{\min} = -(1 - \sqrt{\gamma})^2 r_{\min}$.

We are now ready to state the result concerning weak convergence of the empirical distributions (2.11)–(2.14) to the true out-of-sample error distribution (2.3).

Theorem 2.3.1 (Distribution estimation)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for $\lambda > \lambda_{\min}$,*

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda, \tag{2.16}$$

almost surely (which means, here and henceforth, almost surely with respect to the distribution of X, y), as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

In (2.16), note the left- and right-hand sides both depend on n, p. To explain what we mean by convergence in distribution here: if $\widehat{P}_n$ and P_n are univariate distributions depending on n (where we make the notational dependence explicit for concreteness), and their CDFs are $\widehat{F}_n$ and F_n respectively, then we write $\widehat{P}_n \xrightarrow{\mathrm{d}} P_n$ as $n \to \infty$ to mean that $|\widehat{F}_n(z) - F_n(z)| \to 0$ for every z that is a continuity point of F_n for all n large enough.

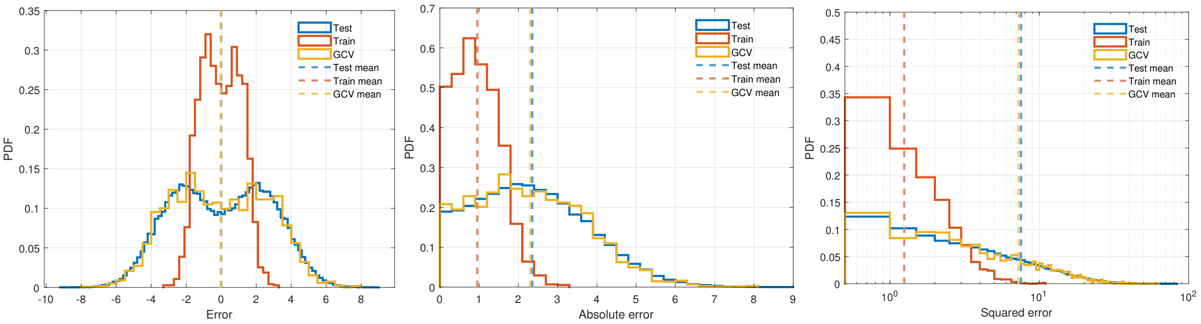
We remark that if we make the stronger assumption that P_λ converges weakly to a continuous distribution, then Theorem 2.3.1 can be strengthened from pointwise to uniform convergence in the following sense: in place of (2.16), we have $\sup_{z \in \mathbb{R}} |\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F_\lambda(z)| \to 0$, where F_λ and $\widehat{F}_\lambda^{\mathrm{gcv}}$ are the distribution functions associated with P_λ and $\widehat{P}_\lambda^{\mathrm{gcv}}$, respectively. The analogous result holds for LOOCV as well. This follows from standard arguments (e.g., Chapter 3 of Durrett, 2019), and we omit the details.

An extension (resembling the continuous mapping theorem) of Theorem 2.3.1 is given next.

Corollary 2.3.2. *Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function, and H_λ denote the distribution of the transformed error $h(e_\lambda)$ conditional on the training data. Let $\widehat{H}_\lambda^{\mathrm{gcv}}$ and $\widehat{H}_\lambda^{\mathrm{loo}}$ denote the empirical distributions as in (2.11)–(2.14), but where the point mass in each summand is evaluated at h of its argument. Then, under Assumptions 2.1 and 2.2, for $\lambda > \lambda_{\min}$,*

$$\widehat{H}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} H_\lambda \quad \text{and} \quad \widehat{H}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} H_\lambda, \tag{2.17}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.



(a) $h(e) = e$ (b) $h(e) = |e|$ (c) $h(e) = e^2$

Figure 2.2: An example with $n = 2500$, $p = 5000$. We generated each x_i according to a Bernoulli distribution, and y_i by adding Bernoulli noise to a nonlinear (quadratic) function of x_i. The ridge tuning parameter was fixed at $\lambda = 1$. Each panel above examines weak convergence per (2.17) for a different function h of the error variable (identity, absolute value, and square, from left to right). In each case, the GCV estimate (yellow) tracks the true distribution (blue) closely. Empirical results for LOOCV are given in the supplement.

Some remarks on the above results are in order. The assumptions required on the distributions of response and features are very weak. Notably, we do not require that the response comes from a well-specified model. Further, the distributions of the response and feature components could be arbitrary so long as they satisfy the moment bounds. As an illustration, we consider examples with binary features and noise in Figure 2.2. Finally, since $\lambda_{\min} < 0$, the results cover the case of the min-norm interpolator (except when $\gamma = 1$).

We next provide some intuition as to why the above results are true. Consider the special case of an underlying linear model $y_0 = x_0^\top \beta_0 + \varepsilon_0$, where $\beta_0 \in \mathbb{R}^p$ is deterministic unknown parameter vector and ε_0 is independent of x_0. In this case, the out-of-sample prediction error simplifies to $e_\lambda = x_0^\top (\beta_0 - \widehat{\beta}_\lambda) + \varepsilon_0$, and

$$P_\lambda = \mathcal{L}\big(x_0^\top (\beta_0 - \widehat{\beta}_\lambda)\big) \star \mathcal{L}(\varepsilon_0),$$

where $\star$ denotes convolution. Further assuming that the features x_0 are Gaussian, as is the noise ε_0, with mean zero and variance σ^2, this law will be Gaussian with mean zero and variance $\|\beta_0 - \widehat{\beta}_\lambda\|_\Sigma^2 + \sigma^2$, where $\|a\|_\Sigma^2 = a^\top \Sigma a$. The variance here is the same as the mean squared prediction error of $\widehat{\beta}_\lambda$. As LOOCV and GCV (in their usual forms (2.8) and (2.9)) track this variance term, Theorem 2.3.1 can be viewed as establishing asymptotic normality of the empirical distributions of LOOCV and GCV errors, in this special case.

However, Theorem 2.3.1 is considerably more general and applies even when $\mathcal{L}(x_0^\top (\beta_0 - \widehat{\beta}_\lambda))$ does not have an analytically known asymptotic limit (and to reiterate, applies even when $\mathbb{E}[y_0 \mid x_0]$ is not linear in x_0). In fact, Theorem 2.3.1 is itself a consequence of a more general result on the convergence of certain functionals of the error distribution, which is covered next.

2.4 Functional estimation

Now we derive convergence theory on the estimation of linear functionals (2.5) of the out-of-sample prediction error distribution. In addition to serving as the main ingredient for proving Theorem 2.3.1, it forms a building block for establishing convergence results that apply to certain nonlinear functionals of the error distribution, discussed in the next section.

2.4.1 Pointwise convergence

We impose the following assumption on the error function t in (2.5).

Assumption 2.3 (Growth rate for the error function)**.** There are constants $a, b, c > 0$ such that $|t(z)| \leq az^2 + b|z| + c$ for any $z \in \mathbb{R}$.

The quadratic growth condition on the error function t in Assumption 2.3 is tied to the moment conditions in Assumptions 2.1 and 2.2. In particular, both assumptions together let us bound $\mathbb{E}[|t(e_\lambda)|^{2+\xi}]$, where $\xi > 0$. One can thus relax the requirement on the growth rate by assuming higher moments in Assumptions 2.1 and 2.2.

Henceforth, let T_λ denote the linear functional in (2.5) corresponding to an error function t, and let $\widehat{T}_\lambda^{\mathrm{gcv}}, \widehat{T}_\lambda^{\mathrm{loo}}$ denote the associated plug-in estimators in (2.15). Next we give the first functional convergence result.

Theorem 2.4.1 (Linear functional estimation)**.** *Suppose Assumptions 2.1 and 2.2 hold, and the function t is continuous and satisfies Assumption 2.3. Then, for $\lambda > \lambda_{\min}$,*

$$\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \to 0 \quad \textit{and} \quad \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \to 0, \tag{2.18}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

Several remarks on the above result follow. As before, the allowed range of tuning parameter values includes the min-norm estimator, since $\lambda_{\min} < 0$ (except when $\gamma = 1$). Moreover, the convergence result in (2.18) holds almost surely (with respect to the training data X, y). This is stronger than many previous results for CV that hold either in probability or expectation over the training data. Lastly, the error function t can be any arbitrary continuous, subquadratic function. In particular, it does *not* need to be bounded (which, by the Portmanteau theorem, would be equivalent to the weak convergence result in Theorem 2.3.1).

A special case of the last result was recently given in Patil et al. (2021) for squared error, $t(e) = e^2$, who assume a much more restricted setting of a well-specified linear model. The current result greatly extends this last one, by allowing for general error functions as well as nonlinear models. The proofs in Patil et al. (2021) exploit the bias-variance decomposition that accompanies squared error, analyze the asymptotic behavior of GCV first, and then tie this to LOOCV. Our approach in this work is completely different (as it must be, due to the general lack of bias-variance decompositions for non-squared error functions). Below we highlight key steps involved in the proof of Theorem 2.4.1.

Proof overview. Our strategy is to study LOOCV first, and then connect it to GCV. It helps to introduce an intermediate quantity:

$$\widetilde{T}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big], \tag{2.19}$$

where we use X_{-i} and y_{-i} for the feature matrix and response vector with the i^{th} row and element removed, respectively, and $\widehat{\beta}_{-i,\lambda}$ for the ridge estimator trained on X_{-i} and y_{-i}. One can interpret (2.19) as the average of the functionals of the leave-one-out estimators $\widehat{\beta}_{-i,\lambda}$, $i = 1, \dots, n$. The result then follows from establishing that: (i) $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$, (ii) $\widetilde{T}_\lambda - \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$, and (iii) $\widehat{T}_\lambda^{\mathrm{loo}} - \widehat{T}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$. In step (i), we use the modulus of continuity of a suitably truncated error function and the stability of the ridge regression estimator. Step (ii) is based on identifying a martingale difference sequence and applying the Burkholder concentration inequality. In step (iii), we use a key lemma from Patil et al. (2021) on the asymptotic equivalence of certain functionals of sample covariance matrices. The full proof is deferred to the supplement (as with all others in this work).

2.4.2 Uniform convergence

The result in Theorem 2.4.1, which is pointwise in λ, can be made uniform in λ under a stronger assumption on the error function t.

Assumption 2.4 (Growth rate for the derivative of the error function)**.** There are constants $g, h > 0$ such that $|t'(z)| \leq g|z| + h$ for any $z \in \mathbb{R}$.

Theorem 2.4.2 (Linear functional estimation, uniform in λ)**.** *Assume the conditions of Theorem 2.4.1, and that t is differentiable and satisfies Assumption 2.4. Then, for any compact $\Lambda \subseteq (\lambda_{\min}, \infty)$,*

$$\sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \right| \to 0 \quad \textit{and} \quad \sup_{\lambda \in \Lambda} \left| \widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \right| \to 0, \tag{2.20}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

We remark that it is not essential that the error function t be differentiable. We can prove a similar result assuming that the error function t is Lipschitz continuous. We assume a global Lipschitz error function t to simplify the proof, but it should be possible to further relax this to a locally Lipschitz assumption, where we have control over the average Lipschitz constant. We do not pursue this in the current work.

Theorem 2.4.3 (Linear functional estimation, uniform in λ, nonsmooth t)**.** *Assume the conditions of Theorem 2.4.1, and that t is Lipschitz continuous. Then, for any compact $\Lambda \subseteq (\lambda_{\min}, \infty)$, the same result as in (2.20) holds, almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

Such uniform convergence will come in handy in the applications discussed next.

2.5 Other applications

The main application of Theorem 2.4.1 discussed thus far is the weak convergence in Theorem 2.3.1. Several other applications are possible, as detailed in this section.

2.5.1 Variational functional estimation

We consider estimation of certain nonlinear functionals that can be represented in variational form as minimizers of parametrized linear functionals over a sufficiently "nice" family of error functions. The main idea behind such an approach is to exploit uniform convergence of the plug-in estimators over the family.

Let $\mathcal{T}_{\mathcal{V}} = \{t(\cdot, v) : \mathbb{R} \to \mathbb{R} \,:\, v \in \mathcal{V}\}$ denote a family of functions indexed by a set $\mathcal{V} \subseteq \mathbb{R}$. Corresponding to each error function $t(\cdot, v)$ in $\mathcal{T}_{\mathcal{V}}$, let $T_\lambda(v)$ denote the linear functional (2.5) associated with $\widehat{\beta}_\lambda$. A variational error functional, denoted by V_λ, is defined as

$$V_\lambda = \operatorname*{arg\,min}_{v \in \mathcal{V}} T_\lambda(v). \tag{2.21}$$

This is assumed to be unique.[2] Meanwhile, denoting by $\widehat{T}_\lambda^{\mathrm{gcv}}(v)$ and $\widehat{T}_\lambda^{\mathrm{loo}}(v)$ the plug-in estimators (2.15) associated with the error function $t(\cdot, v)$, for $v \in \mathcal{V}$, we can then define:

$$\widehat{V}_\lambda^{\mathrm{gcv}} \in \operatorname*{arg\,min}_{v \in \mathcal{V}} \widehat{T}_\lambda^{\mathrm{gcv}}(v), \tag{2.22}$$

$$\widehat{V}_\lambda^{\mathrm{loo}} \in \operatorname*{arg\,min}_{v \in \mathcal{V}} \widehat{T}_\lambda^{\mathrm{loo}}(v). \tag{2.23}$$

Note that we do not assume that these are unique (as is reflected by the element notation above). Our main result in the variational setting is as follows.

Theorem 2.5.1 (Variational functional estimation)**.** *Suppose Assumptions 2.1 and 2.2 hold. Let $\mathcal{T}_{\mathcal{V}}$ be a pointwise equicontinuous family of functions, where $\mathcal{V}$ is compact, and each $t(\cdot, v)$ satisfies Assumption 2.3. For $\lambda > \lambda_{\min}$,*

$$\widehat{V}_\lambda^{\mathrm{gcv}} - V_\lambda \to 0 \quad \textit{and} \quad \widehat{V}_\lambda^{\mathrm{loo}} - V_\lambda \to 0, \tag{2.24}$$

almost surely as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

The proof of Theorem 2.5.1 builds on the previous results. We apply Theorem 2.4.1 on $t(\cdot, v)$ to establish the convergence of $\widehat{T}_\lambda^{\mathrm{gcv}}(v)$ to $T_\lambda(v)$ for each $v \in \mathcal{V}$. The pointwise equicontinuity of functions in $\mathcal{T}_{\mathcal{V}}$ leads to stochastic equicontinuity of $\widehat{T}_\lambda^{\mathrm{gcv}}(v) - T_\lambda(v)$, which then provides GCV part of (2.24). Similar arguments hold for LOOCV.

[2] This is done for simplicity, so we do not have to appeal to set-theoretic notation for convergence of minimizers in the statements that follow. More general formulations that do not assume uniqueness, via variational analysis, should be possible.

2.5.2 Quantile estimation

To illustrate the use of Theorem 2.5.1, we consider estimating quantiles of the out-of-sample prediction error distribution. For $\tau \in (0,1)$, let $Q_\lambda(\tau)$ denote the level-τ conditional quantile (2.6), assumed unique for simplicity. While this is a nonlinear functional of P_λ, we will exploit the fact that (2.6) can expressed in an equivalent variational form (Koenker and Bassett Jr., 1978):

$$Q_\lambda(\tau) = \operatorname*{arg\,min}_{u \in \mathcal{U}} \mathbb{E}\big[t_\tau\big(y_0 - x_0^\top \widehat{\beta}_\lambda - u\big) \mid X, y\big], \tag{2.25}$$

where $t_\tau(u) = u(\tau - \mathbb{I}(u < 0))$, sometimes called the pinball or tilted ℓ_1 loss. If $\mathcal{U}$ is any set containing the true quantile, we can recognize $Q_\lambda(\tau)$ as being in the form (2.21), for the family $\mathcal{T}_\mathcal{U} = \{t_\tau(\cdot, u) : u \in \mathcal{U}\}$. We can then define plug-in estimators $\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau)$ and $\widehat{Q}_\lambda^{\mathrm{loo}}(\tau)$ as in (2.22) and (2.23), or to be fully explicit:

$$\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau) \in \operatorname*{arg\,min}_{u \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^n t_\tau\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \frac{\operatorname{tr}[L_\lambda]}{n}} - u\right), \tag{2.26}$$

$$\widehat{Q}_\lambda^{\mathrm{loo}}(\tau) \in \operatorname*{arg\,min}_{u \in \mathcal{U}} \frac{1}{n} \sum_{i=1}^n t_\tau\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} - u\right), \tag{2.27}$$

with suitable adaptations based on (2.13), (2.14) if $\lambda = 0$. These are essentially just the sample quantiles of GCV and LOOCV residuals, up to discretization issues (the sample quantiles not being unique for integral τn).

Corollary 2.5.2 (Quantile estimation)**.** *Suppose Assumptions 2.1 and 2.2 hold. Given $\tau \in (0,1)$, assume the level-τ quantile $Q_\lambda(\tau)$ of P_λ is unique, and assume $\mathcal{U}$ in (2.26), (2.27) is any compact set that contains the true quantile. For any $\lambda > \lambda_{\min}$,*

$$\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau) - Q_\lambda(\tau) \to 0 \quad \textit{and} \quad \widehat{Q}_\lambda^{\mathrm{loo}}(\tau) - Q_\lambda(\tau) \to 0, \tag{2.28}$$

almost surely as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Thanks to the general result in Theorem 2.5.1, the proof of (2.28) reduces to verifying the pointwise equicontinuity of the family of pinball loss functions.

Estimating quantiles gives us a way to construct prediction intervals for the out-of-sample response y_0, of the form:

$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big], \tag{2.29}$$

$$\mathcal{I}_\lambda^{\mathrm{loo}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{loo}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{loo}}(\tau_u)\big], \tag{2.30}$$

where $\tau_l < \tau_u$ are appropriate lower and upper quantile levels chosen to provide the desired coverage. These intervals have asymptotically exact coverage conditional on the training set, as a consequence of Corollary 2.5.2. See Figure 2.3 for empirical results.

2.5.3 Regularization tuning

One important application of convergence results that are uniform in λ, for given functionals, is that we can tune the amount of regularization according to those functionals, and uniformity will imply that any minimizer of the plug-in estimator converges to a minimizer of the population functional. A typical strategy is to tune by minimizing the mean squared GCV or LOOCV error; but we can also tune via more robust measures such as absolute error, Huber error, or the length of the prediction intervals.

The next corollary certifies that the the level of regularization tuned by using the plug-in GCV and LOOCV estimators is almost surely optimal for a wide range of error functions.

Corollary 2.5.3 (Convergence of tuned errors)**.** *Suppose Assumptions 2.1 and 2.2 hold. Suppose the error function t satisfies Assumption 2.3, and furthermore, it is either differentiable and satisfies Assumption 2.4,*

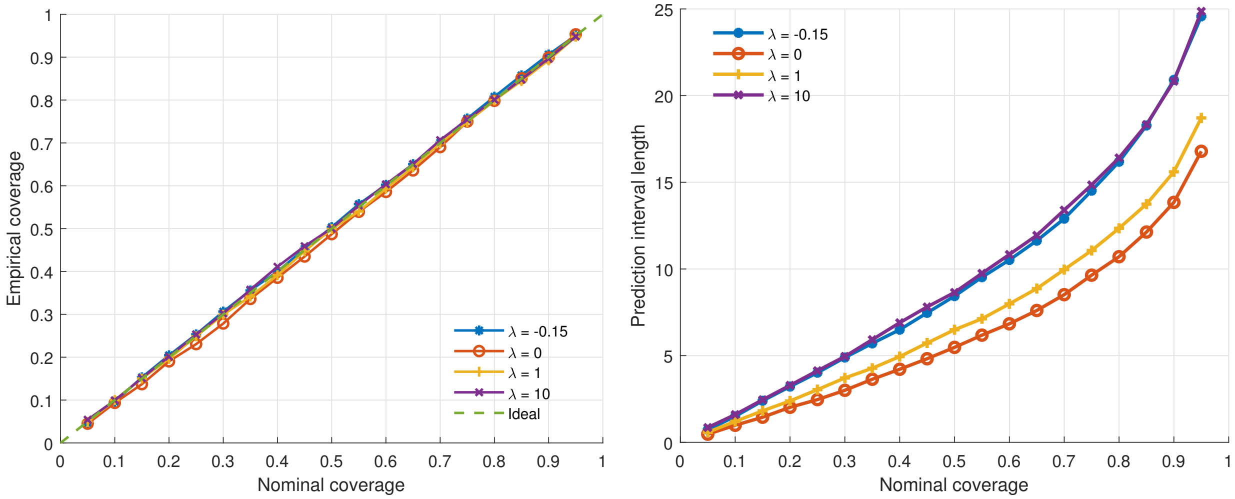


Figure 2.3: Illustration of empirical coverage and length of GCV prediction intervals (2.29) against nominal coverage, where $n = 2500$, $p = 5000$. The data model has a latent structure with autoregressive feature covariance and true signal aligned with the principal eigenvector, similar to that in Kobak et al. (2020) (the supplement gives details), who investigated the empirical optimality of the min-norm interpolator. Here we see that intervals for any λ have excellent finite-sample coverage (left), and the case of $\lambda = 0$ provides the smallest interval lengths (right).

or else it is Lipschitz. Let $\Lambda \subseteq (\lambda_{\min}, \infty)$ be compact, and let $\lambda^\star$ be a minimizer of T_λ over Λ. Similarly, let $\widehat{\lambda}^{\mathrm{gcv}}$ and $\widehat{\lambda}^{\mathrm{loo}}$ denote minimizers of $\widehat{T}^{\mathrm{gcv}}_\lambda$ and $\widehat{T}^{\mathrm{loo}}_\lambda$ over Λ, respectively. Then,

$$T_{\widehat{\lambda}^{\mathrm{gcv}}} - T_{\lambda^\star} \to 0 \quad \textit{and} \quad T_{\widehat{\lambda}^{\mathrm{loo}}} - T_{\lambda^\star} \to 0, \tag{2.31}$$

almost surely as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

2.6 Discussion

In this work, we investigate the distribution of errors arising from both generalized and leave-one-out cross-validation in the context of ridge regression. We show that these distributions converge to the out-of-sample prediction error distribution, under generic conditions. A core result in our work is on consistent estimation of linear functionals of the error distribution, yielding wide implications, including an extension to estimating certain nonlinear functionals which has applications in conditional predictive inference.

Amazingly (and surprisingly, even to us), these results continue to hold in an high-dimensional setting when $p > n$. LOOCV for ridge regression takes on a special form, based on the beautiful "shortcut" relation:

$$y_i - x_i^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}.$$

When $p > n$ and $\lambda = 0$, the numerator and denominator in both fractions here are zero. However, as $\lambda \to 0$ the numerator and denominator (in each fraction) tend to zero at exactly the same rate, allowing us to "cancel" the dependence on λ infinitesimally, leading to:

$$y_i - x_i^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}.$$

This fact was first derived in Hastie et al. (2022), and it is key for our results.

The most immediate next direction is to study kernel ridge regression, which yields a similar "shortcut" formula (Hastie, 2020) where $XX^\top$ gets replaced by the kernel gram matrix. For other predictive models that do not yield exact leave-one-out formulae (in terms of training errors), examining to what degree similar results hold true is an interesting direction for future study. This is especially interesting for "benign" interpolators, now an active area of research, which decompose into a "simple" component useful for prediction and a "spiky" component that interpolates the training data (Bartlett et al., 2021). As interpolators gain a central role in modern machine learning, adapting CV methods to work seamlessly with them is becoming of foundational importance. This current work serves as a step in that direction.

Chapter 3

Mitigating multiple descents

3.1 Introduction

A striking feature of overparameterized models is the so-called "double/multiple descent" behavior in the generalization error curve when plotted against the number of parameters or as a function of the aspect ratio of the number of parameters to the sample size (Belkin et al., 2019a). In a typical double descent scenario, the generalization or test error initially increases as a function of the aspect ratio. It peaks and in some cases explodes as this ratio crosses the *interpolation threshold*, where the learning algorithm achieves a degree of complexity that allows for perfect interpolation of the data. Past the interpolation threshold, the test error tapers down as the complexity of the algorithm increases relative to the sample size. Furthermore, for some algorithms and settings, e.g., the lasso and the minimum ℓ_1-norm least square (e.g., Li and Wei, 2021) or various structures of the design matrix (Adlam and Pennington, 2020; Chen et al., 2020), multiple descents may occur. Double and multiple descent phenomena have been first demonstrated empirically, e.g., for decision trees, random features and two-layer and deep neural networks, and some of these findings have now been corroborated by rigorous theories in a growing body of work: see, e.g., Neyshabur et al. (2014); Nakkiran et al. (2019); Belkin et al. (2018b, 2019a); Mei and Montanari (2022); Adlam and Pennington (2020); Chen et al. (2020); Li and Wei (2021), among others. However, in general, the shape and number of local minima associated with a non-monotonic risk profile due to double descent depend non-trivially on the learning problem, the algorithm deployed, and to an extent, the properties of the data generating distribution in ways that are only partially understood.

The non-monotonic behavior of the generalization error as a function of the aspect ratio in the over-parameterized settings suggests the jarring conclusion that, in high dimensions, increasing the sample size might actually yield a worse generalization error. In contrast, it is highly desirable to rely on prediction procedures that are guaranteed to deliver, at least asymptotically, a risk profile that is monotonically increasing in the aspect ratio, over a large class of data generating distributions. (Note that increasing in aspect ratio is same as decreasing in sample size for a given number of features.) To that effect, some authors have considered ridge-regularized estimators; see Nakkiran et al. (2021); Hastie et al. (2022). In those cases, under fairly restrictive settings and distributional assumptions, a monotonic risk profile can be assured. However, in general settings and for any given procedure, it is unclear how to determine whether the associated risk profile is at least approximately non-monotonic and, if so, how to mitigate it. The ubiquity of the double and multiple descent phenomenon in over-parameterized settings begs the question:

Is it possible to modify any given prediction procedure in order to achieve a monotonic risk behavior?

In this work, we answer this question in the affirmative. More specifically, we develop a simple, general-purpose framework that takes as input an arbitrary learning algorithm and returns a modified version whose out-of-sample risk will be asymptotically no larger than the smallest risk achievable beyond the aspect ratio for the problem at hand. In particular, the asymptotic risk of the returned procedure, as a function of the aspect ratio, will stay below the "monotonized" asymptotic risk profile of the original procedure corresponding to its largest non-decreasing minorant (see Figure 3.1 for an illustration). As a result, when the risk function of the original procedure exhibits double or multiple descents, our modification will guarantee, asymptotically, a far smaller out-of-sample risk near the peaks of the risk function. Our approach is applicable to a large class of data generating distributions and learning problems, with mild to no assumptions on the learning algorithm of choice.

To illustrate the type of guarantees obtained in this work, we provide a preview of one of our main results from Section 3.3.3.1 and comment on its implication. Adopting a standard regression framework, we assume that the data $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are comprised of n i.i.d. pairs of a p-dimensional covariate and a response variable from an unknown distribution. Using $\mathcal{D}_n$, suppose one fits a predictor $\widehat{f}$ — a random function that maps $x \in \mathbb{R}^p \mapsto \widehat{f}(x) \in \mathbb{R}$. Given a loss function $\ell \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, we evaluate the performance of $\widehat{f}$ by its conditional predictive risk given the data, defined by $R(\widehat{f}; \mathcal{D}_n) = \mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]$, where (X_0, Y_0) is an unseen data point, drawn independently from the data generating distribution. Note the risk is a random variable, as it depends on the data $\mathcal{D}_n$. We are interested in the limiting behavior of the risk under the proportional asymptotic regime in which $n, p \to \infty$ with the aspect ratio p/n converging to a constant $\gamma \in (0, \infty)$. As noted above, in such regime the asymptotic risk profile of $\widehat{f}$ has been recently shown to be non-monotonic for a wide variety of problems and procedures. In order to mitigate such behavior, we devise a modification of the original procedure $\widehat{f}$ that results into a new procedure $\widehat{f}^{\mathrm{zs}}$, called zero-step procedure (described in Algorithm 2), whose asymptotic risk profile is provably monotonic in γ. The following informal result can be derived as a consequence of results in Section 3.3.3.1.

Theorem 3.1.1 (Informal monotonization result)**.** *Suppose there exists a deterministic function $R^{\mathrm{det}}(\cdot; \widehat{f}) : (0, \infty] \to [0, \infty]$ such that for any $\phi \in (0, \infty]$ for any dataset $\mathcal{D}$ consisting of m i.i.d. observations with p_m features, $R(\widehat{f}; \mathcal{D}) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi; \widehat{f})$, whenever $m, p_m \to \infty$ and $p_m/m \to \phi$. Then, under mild assumptions on R^{det}, the loss function ℓ, and the data generating distribution, the zero-step procedure $\widehat{f}^{\mathrm{zs}}$ satisfies*

$$\left| R(\widehat{f}^{\mathrm{zs}}; \mathcal{D}_n) - \min_{\zeta \geq \gamma} R^{\mathrm{det}}(\zeta; \widehat{f}) \right| \xrightarrow{\mathrm{p}} 0$$

as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

Figure 3.1 illustrates the above result for the minimum ℓ_2-norm least squares estimator (Hastie et al., 2022) and the minimum ℓ_1-norm least squares estimator (Li and Wei, 2021). The light-blue lines show the asymptotic risk profiles of the two procedures, which are non-monotonic as they diverge to infinity around the interpolation threshold of 1, at which the sample size and the number of features are equal. The red lines depict the risk profiles of the zero-step procedure $\widehat{f}^{\mathrm{zs}}$, which corresponds to the map

$$\gamma \in (0, \infty) \;\mapsto\; \min_{\zeta \geq \gamma} R^{\mathrm{det}}(\zeta; \widehat{f}). \tag{3.1}$$

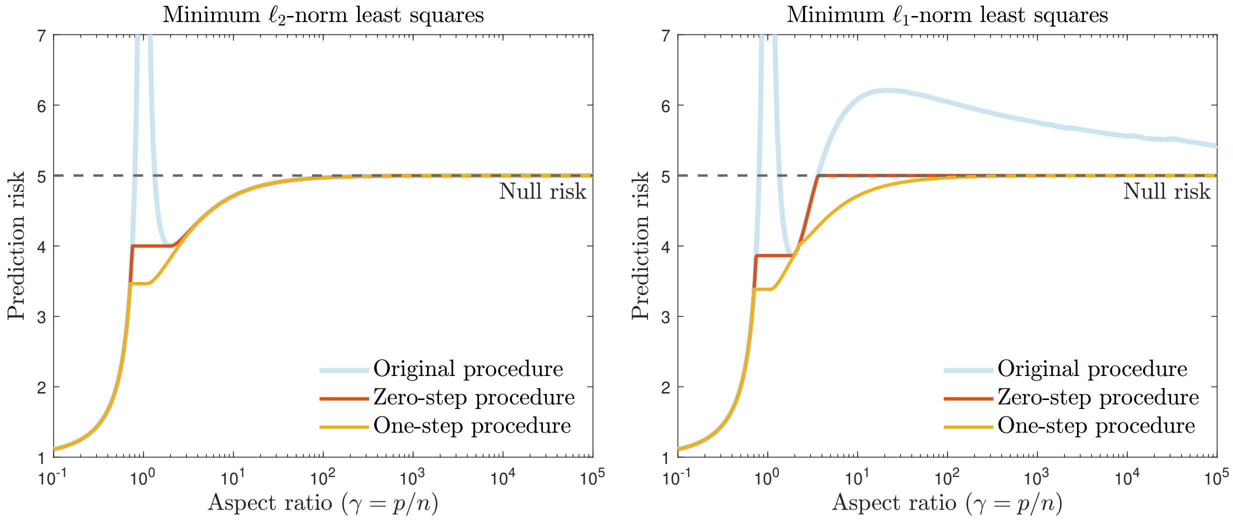
The function (3.1) is a monotonically non-decreasing function of γ, regardless of whether $\gamma \mapsto R^{\mathrm{det}}(\gamma; \widehat{f})$ is non-monotonic. Furthermore, since

$$\min_{\zeta \geq \gamma} R^{\mathrm{det}}(\zeta; \widehat{f}) \leq R^{\mathrm{det}}(\gamma; \widehat{f}), \text{ for all } \gamma > 0,$$

the asymptotic risk of $\widehat{f}^{\mathrm{zs}}$ is no worse than that of $\widehat{f}$. We refer to the function described in (3.1) as the *monotonized risk of the base procedure $\widehat{f}$.*

The assumptions required in Theorem 3.1.1 are very mild, and apply to a broad range of procedures and settings. Indeed, as remarked above, the risk profile $R^{\mathrm{det}}(\cdot; \widehat{f})$ of several estimators have been recently identified under proportional asymptotics regime; see Remark 3.3.16. The requirements on the loss functions are also mild and can be verified for common loss functions. In fact, our results do not require proportional asymptotics and hold more generally.

We also develop a more sophisticated methodology whose asymptotic risk profile is not only monotonic in the aspect ratio but can be strictly smaller than the monotonized risk profile (3.1), a fact that we again verify for the minimum ℓ_2, ℓ_1-norm least squares procedures. See Section 3.4.


Minimum ℓ_2-norm least squares

Minimum ℓ_1-norm least squares

Figure 3.1: Monotonized asymptotic conditional prediction risk of the zero-step procedure (described in Algorithm 2) and one-step procedure (described in Algorithm 3) for the minimum ℓ_2-norm and ℓ_1-norm least squares procedures. The figure in the left panel follows the setup of Figure 2 of Hastie et al. (2022), and the figure in the right panel follows the setup of Figure 3 of Li and Wei (2021) (at sparsity level = 0.1). Both settings assume isotropic features and a linear model with noise variance $\sigma^2 = 1$ and linear coefficients of squared Euclidean norm $\rho^2 = 4$. Note that the risk is lower bounded by $\sigma^2 = 1$ and the risk of the null predictor (null risk) is $\rho^2 + \sigma^2 = 5$.

Core idea: the zero-step procedure. Our methodology is conceptually straightforward, as it relies on a combination of sample splitting, sub-sampling, and cross-validation. The core principle is as follows. Starting off with an aspect ratio of p/n, if the risk were to be lower at, say, twice this aspect ratio $2p/n$, then we could just use half the data to evaluate the predictor, enjoying a smaller risk than the one obtained when training with the entire data. To decide whether the out-of-sample error is lower at any larger aspect ratio, we use cross-validation to "glean at" the values of the risk function at all aspect ratios larger than the one for the full data. To elaborate, we next give an informal description of one of our main methods, the zero-step procedure that we study in Section 3.3.

We initially split the data into a training and a validation set in such a way that the size of the validation set is a vanishing proportion of that of the training set. In the first step, we compute a collection of predictors, each resulting from applying the same base prediction procedure on a sub-sample of size k_n varying over a grid of values in $\mathcal{K}_n$. Depending on the size of the sub-sample, we are able to mimic the behavior of the risk at larger aspect ratios $(p/k_n,\ k_n \in \mathcal{K}_n)$. In the second step, we estimate the out-of-sample risk of each of these predictors using the validation set. With $\{p/k_n : k_n \in \mathcal{K}_n\}$ approximating the set $[p/n, \infty]$, these estimated out-of-sample risks act as proxies for the true generalization error at larger aspect ratios. In the final step, we perform model selection by minimizing the estimated test error across the candidate aspect ratios. In order to make full use of the data, one can use more than one sub-sample for each $k_n \in \mathcal{K}_n$, a practice that closely resembles bagging. To prove the "correctness" of the split-sample cross-validation, we develop novel oracle inequalities in additive and multiplicative forms that are of independent interest.

Because the core components of our approach are sub-sampling and cross-validation, our methodology is applicable to virtually any algorithm – even the black-box type – and its validity holds under minimal assumptions on the data generating distribution.

3.1.1 Summary of results

Below we summarize the main contributions of this work.

- **Novel guarantees for split-sample cross-validation.** At its core, our methodology performs model selection of arbitrary learning procedures built over sub-samples of different sizes, with the size of the sub-samples treated as a tuning parameter to optimize. Towards that goal, we rely on split-sample cross-validation, which we analyze in Section 3.2. In Proposition 3.2.1, we provide deterministic inequalities for the risk of split cross-validated predictors in both additive and multiplicative form. We remark that multiplicative oracle inequalities allow for the possibility of unbounded oracle risk values, and are therefore well suited to incorporate prediction procedures exhibiting the double descent phenomena around the interpolating threshold. Leveraging concentration inequalities for both the mean estimator of the prediction risk and the median-of-means estimator, in Section 3.2.3, we show how these bounds imply finite-sample oracle inequalities for split-sample cross-validation that are applicable to a broad range of loss functions and under minimal assumptions on the learning procedure. In particular, our results do not require well-specified (parametric) models. We exemplify our bounds on various loss functions for both regression and classification, and in Theorem 3.2.22, we give a general multiplicative oracle inequality for arbitrary linear predictors under mild distributional assumptions.

- **Zero-step procedure.** Using oracle inequalities for split-sample cross-validation, we put forth a general methodology that takes as input an arbitrary prediction procedure and minimizes the prediction risk of its bagged version over a grid of sub-sample sizes. We call this the "zero-step" prediction procedure. We analyze the asymptotic risk behavior of the zero-step procedure under proportional asymptotics, in which the number of features grows proportionally with the number of observations. In Theorem 3.3.11, we prove that the risk of predictor returned by the zero-step procedure is upper bounded by the monotonized risk given in (3.1). Unlike most contributions in the literature on over-parameterized learning, our results do not depend on well-specified (parametric) models and only require the existence of a sufficiently well-behaved asymptotic risk profile.

- **One-step procedure.** In Section 3.4, we further generalize the zero-step procedure by considering an adjustment of the original predictor that is inspired by the one-step estimation method used in parametric statistics to improve efficiency (Van der Vaart, 2000, Section 5.7). This modification, which can be thought of as a single-iterate boosting of the baseline procedure, is shown, both in theory and in simulations, to produce an asymptotic monotonized risk that is smaller than the monotonized risk of the zero-step procedure; see Theorem 3.4.4. We derive explicit expressions of the asymptotic risk profile of the one-step procedure for the minimum ℓ_2, ℓ_1-norm least squares prediction procedures. The main insight we draw from the minimum ℓ_2-norm least squares example is that the one-step procedure in addition to changing the aspect ratio of the predictor also reduces the signal energy leading to a smaller asymptotic risk; see Remark 3.4.12.

- **Risk profiles**. In our study of the performance of the zero-step and one-step procedures, we derive several auxiliary results that might of independent interest. Specifically, we provide a systematic way to certify the continuity or lower semicontinuity of the asymptotic risk profile of any prediction procedure, assuming only point-wise convergence of the conditional prediction risk under proportional asymptotics; see Proposition 3.3.10. This is often hard to prove directly from the asymptotic risk profiles as they are usually defined implicitly via one or more fixed-point equations. Also of independent interest is a representation that we prove, for the conditional prediction risk of an arbitrary linear predictor with a one-iterate boosting with minimum ℓ_2-norm least squares, using the recent tools from random matrix theory. This, in particular, involves deriving deterministic equivalents for the generalized bias and variance of the ridgeless predictor which may be of independent interest; see Lemmas 3.4.8 and C.5.3.

We corroborate our theoretical results with several illustrative simulations. An intriguing finding emerging from our numerical studies is the fact that bagging, i.e., aggregation over sub-sample, appears to have a significant positive impact on the asymptotic risk profile of both the zero- and one-step procedure: averaging over an increasing number of sub-samples results in a downward shift of the risk asymptotic profile, especially around the interpolation threshold: see, e.g., Figures 3.3 and 3.4. Though we do not provide a theoretical justification for this interesting phenomenon, we offer some conjectures in the discussion section; see Section 3.5.

3.1.2 Other related work

In this section, we review some related work on risk non-monotonicity, cross-validation, as well as exact asymptotic risk characterization. Explicit references to these works, when appropriate, are also made in the main sections.

Non-monotonicity of generalization performance. The study of non-monotone risk behavior is largely motivated by empirical evidence in standard statistical learning tasks such as classification and prediction, where instances of non-monotonic risk profiles were originally discovered and reported. See Trunk (1979); Duin (1995); Opper and Kinzel (1996) and Loog et al. (2020) for some earlier findings on the double descent risk behavior. Recently, it has garnered growing interest due to the remarkable successes of neural networks where similar non-monotonic behavior has also been observed; see LeCun et al. (1990); Geiger et al. (2019); Zhang et al. (2017, 2021) and references therein. The non-monotonic behavior of the test error as a function of the model size in general context was brought up by Belkin et al. (2019a) and has since been theoretically established for many other classical estimators such as linear/kernel regression, ridge regression, logistic regression, and under stylized models such as linear model or random features model. Besides the work discussed in our main sections, see also Kini and Thrampoulidis (2020); Mei and Montanari (2022); Mitra (2019); Derezinski et al. (2020); Frei et al. (2022) and the survey work Bartlett et al. (2021). When it comes to the sample-wise non-monotonic performances, a recent line of work asks and provides partial answers to the question: given additional observation points, when and to what extend will the generalization performance improve (Viering et al., 2019; Nakkiran, 2019; Nakkiran et al., 2021; Mhammedi, 2021). In particular, Nakkiran et al. (2021) investigates the role of optimal tuning in the context of ridge regression, and for a class of linear models, demonstrated that the optimally-tuned ℓ_2 regularization achieves monotonic generalization performance.

Data-splitting and cross-validation. The framework developed in the current work crucially depends on split-sample cross-validation, which compares different predictors trained on one part of the sample using out-of-sample risk estimates from the remaining part. The split-sample cross-validation is a well-known methodology studied in several works (e.g., Stone (1974); Györfi et al. (2006); Yang (2007); Arlot and Celisse (2010)). Split-sample cross-validation is theoretically easier to analyze compared to the k-fold cross-validation and is shown to yield optimal rates in the context of non-parametric regression (Yang, 2007; Van der Laan et al., 2007; Van der Vaart et al., 2006). These works have derived oracle inequalities that show that split-sample cross-validation based predictor has asymptotically the smallest risk among the collection of predictors up to an additive error (that converges to zero). The oracle inequalities are either called exact or inexact depending on whether the constant multiplying the smallest risk is 1 or $1 + \delta$ (for an arbitrarily δ); see, e.g., Lecué and Mendelson (2012). All these works have used split-sample cross-validation for the purpose of choosing predictors with good prediction risk, and the existing oracle inequalities are all additive in nature.

Application of cross-validation for over-parameterized learning is more recent and here special care is required in choosing the split sizes because splitting in half would change the aspect ratios in the proportional asymptotics regime. In contrast to the low dimensional or non-parametric setting, it is well-known that the classical k-fold cross-validation framework suffers from severe bias and thus requires careful modification or a diverging choice of k (see, e.g., Mücke et al. (2022); Rad and Maleki (2020)). In particular, when k is taken to be n, the resulting procedure is also known as leave-one-out cross-validation (LOOCV), which mitigates these bias issue and has proven to be effective in a variety of settings; see Beirami et al. (2017); Wang et al. (2018b); Giordano et al. (2019); Stephenson and Broderick (2020b); Wilson et al. (2020); Austern and Zhou (2020); Xu et al. (2021); Patil et al. (2021, 2022b) and references therein.

Our use of cross-validation is slightly different: the goal is to choose the "optimal" sub-sample size for a single prediction procedure. Furthermore, supplementing the existing oracle inequalities for cross-validation, we also provide a multiplicative oracle inequality which shows that the split-sample cross-validated predictor attains the smallest risk in the collection up to a factor converging to 1 with the sample size. This multiplicative version is crucial for our study, allowing us to consider ingredient predictors whose risk might diverge with sample size.

Risk characterization. In developing our zero-step and one-step procedures, we assume existence of a deterministic risk profile function for every aspect ratio. As discussed, the exact formulas for the risk profile functions have been obtained for various estimators in both classification and regression settings. In the past decade, several distinct techniques and tools have been developed to explicitly describe and analyze these risk functions. Prominent examples include the leave-one-out type perturbation analysis (e.g., Karoui (2013, 2018)), the approximate message passing machinery (e.g., Donoho et al. (2009); Donoho and Montanari (2016); Bayati and Montanari (2011)), and the convex Gaussian min-max theorem (e.g., Stojnic (2013); Thrampoulidis et al. (2015, 2018)). These techniques rely critically upon a well-specified model, as well as the assumption that the entries of the design matrix are drawn i.i.d. from standard normal distribution, while some restricted universality results are developed in Bayati et al. (2015); Montanari and Nguyen (2017); Chen and Lam (2021); Hu and Lu (2020). In this work, however, we take a more direct approach and develop some non-asymptotic oracle risk inequalities. Leveraging upon these oracle inequalities, our results do not require well-specified models, and only assume the existence of a relatively well-behaved risk profile, which presumably allows for weaker distributional assumptions.

3.1.3 Organization and notation

Organization.

- In Section 3.2, we describe the general cross-validation and model selection algorithm, derive associated oracle risk inequalities, and provide probabilistic bounds on the error terms. We then obtain concrete results for a variety of classification and regression loss functions.
- In Section 3.3, we describe the zero-step prediction procedure, and provide its risk monotonization guarantee. We then explicitly verify the related assumptions for the ridgeless and lassoless prediction procedures, and show corresponding numerical illustrations.
- In Section 3.4, we describe the one-step prediction procedure, and provide its risk monotonization guarantee. We then explicitly verify assumptions for arbitrary linear predictors, the special cases of ridgeless and lassoless prediction procedures, and show corresponding numerical illustrations.
- In Section 3.5, we conclude and provide three concrete directions for future work.

Nearly all the proofs are deferred to the supplement.

Notation. We use $\mathbb{N}$ to denote the set of natural numbers, $\mathbb{R}$ to denote the set of real numbers, $\mathbb{R}_{\geq 0}$ to denote the set of non-negative real numbers, $\mathbb{R}_{>0}$ to denote the set of positive real numbers, and $\overline{\mathbb{R}}$ to denote the extended real number system, i.e., $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. For a real number a, $(a)_+$ denotes its positive part, $\lfloor a \rfloor$ denotes its floor, $\lceil a \rceil$ denotes its ceiling. For a set $\mathcal{A}$, we use $\mathbb{I}_{\mathcal{A}}$ to denote its indicator function. We denote convergence in probability by $\xrightarrow{\text{p}}$, almost sure convergence by $\xrightarrow{\text{a.s.}}$, and weak convergence by $\xrightarrow{\text{d}}$. We use generic letters $C, C_1, C_2, \dots$ to denote constants whose values may change from line to line.

For a comprehensive list of notation used in this work, see Appendix C.9.

3.2 General cross-validation and model selection

The primary focus of this work is to develop a framework to improve upon prediction procedures in the overparameterized regime in which the number of features p is comparable to and often exceeds the number of observations n, and where the predictive risk may be non-monotonic in the aspect ratio p/n. As discussed in Section 3.1, a fundamental component of our methodology is the selection of an optimal size of the sub-samples through cross-validation. To that effect, we begin by deriving some general, non-asymptotic oracle risk inequalities for split-sample cross-validation, as described in Algorithm 1, that hold under minimal assumptions. While our bounds apply to a wide range of learning problems and may be of independent interest, they are crucial in demonstrating the risk monotonization properties of the procedures presented in Sections 3.3 and 3.4.

Though cross-validation is a well-known and well-studied procedure (see, e.g., Van der Laan et al., 2007; Györfi et al., 2006; Yang, 2007), our work extends the previous results on cross-validation in a couple of ways: (1) We derive two forms of oracle risk inequalities: the additive form that is better suited for bounded loss functions (especially classification losses), and the multiplicative form that is better suited unbounded loss functions (especially regression losses); (2) In addition to common sample mean based estimation of the prediction risk, we also analyze the median-of-means based estimation of the prediction risk that proves to be useful in relaxing strong moment assumption on the predictors.

3.2.1 Oracle risk inequalities

Setting the stage, suppose we are given n samples of labeled data $\mathcal{D}_n = \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$, where $X_i \in \mathbb{R}^p$ is a p-dimensional feature vector and $Y_i \in \mathbb{R}$ is a scalar response variable for $i = 1, \dots, n$. Let $\widehat{f}$ be a prediction procedure that maps $\mathcal{D}_n$ to a predictor $\widehat{f}(\cdot; \mathcal{D}_n) : \mathbb{R}^p \to \mathbb{R}$ (a measurable function of the data $\mathcal{D}_n$). For any predictor $\widehat{f}(\cdot; \mathcal{D}_n)$, trained on the data set $\mathcal{D}_n$, that takes in a feature vector $x \in \mathbb{R}^p$ and outputs a real-valued prediction $\widehat{f}(x; \mathcal{D}_n)$, we measure its predictive accuracy via a non-negative loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$. Given a new feature vector $X_0 \in \mathbb{R}^p$ with associated response variable $Y_0 \in \mathbb{R}$ so that (X_0, Y_0) is independent of $\mathcal{D}_n$,[1] the prediction error or out-of-sample error incurred by $\widehat{f}(\cdot; \mathcal{D}_n)$ is $\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n))$. Note that the prediction error $\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n))$ is a random variable that is a function of both $\mathcal{D}_n$ and (X_0, Y_0).

We will quantify the performance of $\widehat{f}(\cdot; \mathcal{D}_n)$ using the conditional expected prediction loss. The conditional expected prediction loss given the data $\mathcal{D}_n$, or the conditional prediction risk for short, of $\widehat{f}(\cdot; \mathcal{D}_n)$ is defined as

$$R(\widehat{f}(\cdot; \mathcal{D}_n)) \ := \ \mathbb{E}_{X_0, Y_0}[\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n)) \mid \mathcal{D}_n] \ = \ \int \ell(y, \widehat{f}(x; \mathcal{D}_n)) \, \mathrm{d}P(x, y), \tag{3.5}$$

where P denotes the joint probability distribution of (X_0, Y_0). Note that $R(\widehat{f}(\cdot; \mathcal{D}_n))$ is a random variable that depends on $\mathcal{D}_n$. An empirical estimator of $R(\widehat{f}(\cdot; \mathcal{D}_n))$ is denoted by $\widehat{R}(\widehat{f}(\cdot; \mathcal{D}_n))$. In this work, we mainly consider two such estimators: the average estimator and the median-of-means estimator as defined in (3.2) and (3.3), respectively.

Consider any prescribed index set Ξ, where each $\xi \in \Xi$ corresponds to a specific model that will be clear from the context. Based on the training data, a predictor $\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})$ is fitted for each model ξ and estimated risks of $\widehat{f}^{\xi}$, $\xi \in \Xi$ are compared on a validation data set as described in Algorithm 1. Let $\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n)$ be the final predictor returned by Algorithm 1. We shall consider two types of oracle inequalities: one in an additive form and the other in a multiplicative form. More specifically, for any prescribed model set Ξ, define the additive error term and multiplicative error term respectively as follows:

$$\Delta_n^{\mathrm{add}} := \max_{\xi \in \Xi} \left| \widehat{R}(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})) - R(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})) \right|, \tag{3.6a}$$

$$\Delta_n^{\mathrm{mul}} := \max_{\xi \in \Xi} \left| \frac{\widehat{R}(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}}))}{R(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}}))} - 1 \right|. \tag{3.6b}$$

The following proposition relates the performance of $\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n)$ to the "oracle" prediction risk in terms of these errors terms.

Proposition 3.2.1 (Deterministic oracle risk inequalities)**.** *The prediction risk of $\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n)$ satisfies the following deterministic oracle inequalities:*

[1] We will reserve the notation (X_0, Y_0) to denote a random variable that is drawn independent of $\mathcal{D}_n$.

Algorithm 1 General cross-validation and model selection procedure

Inputs:

- a dataset $\mathcal{D}_n = \{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le n\}$;
- a positive integer $n_{\mathrm{te}} < n$;
- an index set Ξ;
- a set of prediction procedures $\{\widehat{f}^{\xi}\colon\ \xi \in \Xi\}$;
- a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$;
- a centering procedure $\mathtt{CEN} \in \{\mathtt{AVG}, \mathtt{MOM}\}$;
- a real number $\eta > 0$ if $\mathtt{CEN}$ is $\mathtt{MOM}$.

Output:

- a predictor $\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n) : \mathbb{R}^p \to \mathbb{R}$.

Procedure:

1. Randomly split the index set $\mathcal{I}_n = \{1, \dots, n\}$ into two disjoint sets $\mathcal{I}_{\mathrm{tr}}$ and $\mathcal{I}_{\mathrm{te}}$ such that $|\mathcal{I}_{\mathrm{tr}}| = n - n_{\mathrm{te}}$ (which we denote by n_{tr}), $|\mathcal{I}_{\mathrm{te}}| = n_{\mathrm{te}}$. Denote the corresponding splitting of the dataset $\mathcal{D}_n$ by $\mathcal{D}_{\mathrm{tr}} = \{(X_i, Y_i) : i \in \mathcal{I}_{\mathrm{tr}}\}$ (for training) and $\mathcal{D}_{\mathrm{te}} = \{(X_j, Y_j) : j \in \mathcal{I}_{\mathrm{te}}\}$ (for testing).

2. For each $\xi \in \Xi$, fit the prediction procedure $\widehat{f}^{\xi}$ on $\mathcal{D}_{\mathrm{tr}}$ to obtain the predictor $\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}}) : \mathbb{R}^p \to \mathbb{R}$.

3. For each $\xi \in \Xi$,

   - if $\mathtt{CEN} = \mathtt{AVG}$, estimate the conditional prediction risk of $\widehat{f}^{\xi}$ using
   $$\widehat{R}(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})) = \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{j \in \mathcal{I}_{\mathrm{te}}} \ell(Y_j, \widehat{f}^{\xi}(X_j; \mathcal{D}_{\mathrm{tr}})). \tag{3.2}$$

   - if $\mathtt{CEN} = \mathtt{MOM}$, estimate the conditional prediction risk of $\widehat{f}^{\xi}$ using
   $$\widehat{R}(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})) = \mathtt{MOM}\big(\big\{\ell(Y_j, \widehat{f}^{\xi}(X_j; \mathcal{D}_{\mathrm{tr}})),\ j \in \mathcal{I}_{\mathrm{te}}\big\},\ \eta\big). \tag{3.3}$$

     See discussion after Lemma C.8.2 for the definition of $\mathtt{MOM}(\cdot, \cdot)$.

4. Set $\widehat{\xi} \in \Xi$ to be the index that minimizes the estimated prediction risk using
$$\widehat{\xi} \in \operatorname*{arg\,min}_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}(\cdot; \mathcal{D}_{\mathrm{tr}})). \tag{3.4}$$

   Note that $\widehat{\xi}$ need not be unique (hence the set notation) and any choice that leads to the minimum estimated risk enjoys the subsequent theoretical guarantees.

5. Return the predictor $\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n) = \widehat{f}^{\widehat{\xi}}(\cdot; \mathcal{D}_{\mathrm{tr}})$.

1. *additive form:*

$$\begin{aligned} R(\widehat{f}^{\mathrm{cv}}(\cdot;\mathcal{D}_n)) &\le \min_{\xi\in\Xi} R(\widehat{f}^{\xi}(\cdot;\mathcal{D}_{\mathrm{tr}}) + 2\Delta_n^{\mathrm{add}}, \\ \mathbb{E}[R(\widehat{f}^{\mathrm{cv}}(\cdot;\mathcal{D}_n))] &\le \min_{\xi\in\Xi} \mathbb{E}[R(\widehat{f}^{\xi}(\cdot;\mathcal{D}_{\mathrm{tr}})] + 2\mathbb{E}[\Delta_n^{\mathrm{add}}]. \end{aligned} \tag{3.7}$$

2. *multiplicative form:*

$$R(\widehat{f}^{\mathrm{cv}}(\cdot;\mathcal{D}_n)) \le \frac{1+\Delta_n^{\mathrm{mul}}}{(1-\Delta_n^{\mathrm{mul}})_+}\cdot \min_{\xi\in\Xi} R(\widehat{f}^{\xi}(\cdot;\mathcal{D}_{\mathrm{tr}}). \tag{3.8}$$

Proposition 3.2.1 provides oracle bounds on the prediction risk of $\widehat{f}^{\mathrm{cv}}(\cdot;\mathcal{D}_n)$ in terms of the error terms Δ_n^{add} and Δ_n^{mul}. Note that Proposition 3.2.1 does not make any assumptions about the underlying model of the data or the dependence structure between the observations. Under some general conditions on the data, one can show that Δ_n^{add} and/or Δ_n^{mul} converge to zero in probability as $n\to\infty$. The exact rate of convergence depends on the number of observations n_{te} in the test data and also on the tail behavior of $\ell(Y_0,\widehat{f}^{\xi}(X_0;\mathcal{D}_{\mathrm{tr}}))$ conditional on $\widehat{f}^{\xi}(\cdot;\mathcal{D}_{\mathrm{tr}})$. For notational convenience, from now, we will write $\widehat{f}^{\mathrm{cv}}$ and $\widehat{f}^{\xi}$ to denote $\widehat{f}^{\mathrm{cv}}(\cdot;\mathcal{D}_n)$ and $\widehat{f}^{\xi}(\cdot;\mathcal{D}_{\mathrm{tr}})$, respectively.

Remark 3.2.2 (Lower bound on $R(\widehat{f}^{\mathrm{cv}})$)**.** Proposition 3.2.1 provides upper bounds on the (conditional) prediction risk of $\widehat{f}^{\mathrm{cv}}$ in terms of the minimum risk of $\widehat{f}^{\xi}$. It can be readily seen that the risk of $\widehat{f}^{\mathrm{cv}}$ is always lower bounded by the minimum risk. More formally, note that $\widehat{f}^{\mathrm{cv}} = \sum_{\xi\in\Xi}\widehat{f}^{\xi}\mathbb{I}_{\widehat{\xi}=\xi}$, and, therefore,

$$R(\widehat{f}^{\mathrm{cv}}) = \sum_{\xi\in\Xi} R(\widehat{f}^{\xi})\mathbb{I}_{\widehat{\xi}=\xi} \ge \min_{\xi\in\Xi} R(\widehat{f}^{\xi})\sum_{\xi\in\Xi}\mathbb{I}_{\widehat{\xi}=\xi} = \min_{\xi\in\Xi} R(\widehat{f}^{\xi}).$$

Combined with Proposition 3.2.1, we conclude that

$$\min_{\xi\in\Xi} R(\widehat{f}^{\xi}) \le R(\widehat{f}^{\mathrm{cv}}) \le \begin{cases} \min_{\xi\in\Xi} R(\widehat{f}^{\xi}) + \Delta_n^{\mathrm{add}} \\ \min_{\xi\in\Xi} R(\widehat{f}^{\xi})\cdot(1+\Delta_n^{\mathrm{mul}})/(1-\Delta_n^{\mathrm{mul}})_+. \end{cases}$$

Thus, convergence (in probability) of either Δ_n^{add} or Δ_n^{mul} to 0 implies that the risk of $\widehat{f}^{\mathrm{cv}}$ is asymptotically the same as the minimum risk of $\widehat{f}^{\xi}$, $\xi\in\Xi$ in either additive or multiplicative sense, respectively.

The additive and multiplicative form of oracle inequalities have their own advantages. Traditionally, the additive form is more common. The additive oracle inequality for the prediction risk readily implies the additive oracle inequality on the excess risk. In other words,

$$R(\widehat{f}^{\mathrm{cv}}) - R(f^\star) \le \min_{\xi\in\Xi} R(\widehat{f}^{\xi}) - R(f^\star) + \Delta_n^{\mathrm{add}},$$

for any predictor $f^\star$. In particular, this will hold for the best (oracle) predictor for the prediction risk. This is not true of the multiplicative oracle inequality, which instead only implies the bound

$$R(\widehat{f}^{\mathrm{cv}}) - R(f^\star) \le c_n\big\{\min_{\xi\in\Xi} R(\widehat{f}^{\xi}) - R(f^\star)\big\} + (c_n-1)R(f^\star),$$

where $f^\star$ is any predictor (in particular, the one with the best prediction risk) and

$$c_n = \frac{1+\Delta_n^{\mathrm{mul}}}{(1-\Delta_n^{\mathrm{mul}})_+}, \quad c_n - 1 = \frac{2\Delta_n^{\mathrm{mul}}}{(1-\Delta_n^{\mathrm{mul}})_+}.$$

In terms of claiming that $\widehat{f}^{\mathrm{cv}}$ has prediction risk close to the best in the collection of predictors $\{\widehat{f}^{\xi},\xi\in\Xi\}$, the multiplicative form has certain advantages compared to the additive form. In the case that $\min_{\xi\in\Xi} R(\widehat{f}^{\xi})$ converges to 0, the additive oracle inequality (3.7) implies that the risk of the selected predictor $\widehat{f}^{\rm cv}$ asymptotically matches the risk of the *best* predictor among the collection $\{\widehat{f}^{\xi}, \xi \in \Xi\}$ only if $\Delta_n^{\rm add}$ converges to zero faster than $\min_{\xi\in\Xi} R(\widehat{f}^{\xi})$. If, however, $\Delta_n^{\rm add}$ converges to zero slower than the minimum risk in the collection, then the additive oracle inequality does not imply a favorable result. In this case, a multiplicative oracle inequality helps. As long as $\Delta_n^{\rm mul}$ converges to 0, the multiplicative oracle inequality implies that $\widehat{f}^{\rm cv}$ matches in risk with the best predictor in the collection, irrespective of whether the minimum risk converges to zero or not. Note that $\Delta_n^{\rm add}$ only controls the additive error of the risk estimator $\widehat{R}(\widehat{f}^{\xi})$, which is easier to control than the multiplicative error; think of controlling the error of sample mean of Bernoulli(p) random variables with $p = p_n \to 0$; See Remark 3.2.12 for a more mathematical discussion. Even when $\min_{\xi\in\Xi} R(\widehat{f}^{\xi})$ does not converge to zero, the multiplicative form might be advantageous compared to the additive form. Indeed, suppose that $\widehat{f}^{\xi_0}$ is in the collection and its risk diverges as $n \to \infty$. Then, it may not be true that

$$\big|\widehat{R}(\widehat{f}^{\xi_0}) - R(\widehat{f}^{\xi_0})\big| \xrightarrow{p} 0,$$

because both $\widehat{R}(\widehat{f}^{\xi_0})$ and $R(\widehat{f}^{\xi_0})$ are diverging. This implies that $\Delta_n^{\rm add}$ does not converge to 0 and in fact, might diverge. However, the minimum risk in the collection could still be finite, and the additive oracle inequality fails to capture this. On the other hand, $\widehat{R}(\widehat{f}^{\xi_0})/R(\widehat{f}^{\xi_0})$ can still converge to 1 as $n \to \infty$ even if $R(\widehat{f}^{\xi_0})$ diverges to ∞. In our applications in overparameterized learning, we will encounter this situation where the number of features (p) is close to the number of observations (n), i.e., $p/n \approx 1$. See Remark 3.2.23 for more details.

Remark 3.2.3 (From multiplicative to additive oracle inequality)**.** Note that if $\Delta_n^{\rm mul} = o_p(1)$, then $(1+\Delta_n^{\rm mul})/(1-\Delta_n^{\rm mul})_+ = 1 + O_p(1)\Delta_n^{\rm mul} = 1 + o_p(1)$, then the multiplicative oracle inequality (3.8) yields

$$R(\widehat{f}^{\rm cv}) \;\le\; (1 + O_p(1)\Delta_n^{\rm mul}) \min_{\xi\in\Xi} R(\widehat{f}^{\xi}) \;=\; (1 + o_p(1)) \min_{\xi\in\Xi} R(\widehat{f}^{\xi}).$$

Observe that this multiplicative form can be converted into an additive form as

$$R(\widehat{f}^{\rm cv}) \;\le\; \min_{\xi\in\Xi} R(\widehat{f}^{\xi}) \;+\; O_p(1)\Delta_n^{\rm mul} \min_{\xi\in\Xi} R(\widehat{f}^{\xi}),$$

where the second term on the right hand side is always smaller order compared to the first term as long as $\Delta_n^{\rm mul}$ converges in probability to zero.

From this discussion, it follows that one can choose a predictor with the *best* prediction risk in a collection if either $\Delta_n^{\rm add}$ or $\Delta_n^{\rm mul}$ converges in probability to zero. The application of Algorithm 1 for risk monotonizing procedures will be discussed in the next three sections. In the next two subsections, we provide some general sufficient conditions to verify $\Delta_n^{\rm add} = o_p(1)$ and $\Delta_n^{\rm mul} = o_p(1)$ for independent data. We also provide examples of common loss functions and show that under some mild moment assumptions, they satisfy $\Delta_n^{\rm add} = o_p(1)$ and $\Delta_n^{\rm mul} = o_p(1)$.

3.2.2 Control of $\Delta_n^{\rm add}$ and $\Delta_n^{\rm mul}$

In order to characterize $R(\widehat{f}^{\rm cv})$, by Proposition 3.2.1 it is sufficient to control $\Delta_n^{\rm add}$ and $\Delta_n^{\rm mul}$. In this section, we demonstrate that under certain assumptions on the loss function ℓ, the error terms are small both in probability and in expectation, which in turn yields optimality of $\widehat{f}^{\rm cv}$ among the predictors in $\{\widehat{f}^{\xi}, \xi \in \Xi\}$.

To facilitate our discussion, for each $\xi \in \Xi$, define the conditional ψ_1-Orlicz norm of $\ell(Y_0, \widehat{f}^{\xi}(X_0))$ given $\mathcal{D}_n$ as

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}_n} := \inf\big\{C > 0 : \mathbb{E}\big[\exp\big(|\ell(Y_0, \widehat{f}^{\xi}(X_0))|/C\big) \mid \mathcal{D}_n\big] \le 2\big\}. \tag{3.9}$$

Similarly, for $r \ge 1$, define the conditional L_r-norm as

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_r|\mathcal{D}_n} := \big(\mathbb{E}\big[\big|\ell(Y_0, \widehat{f}^{\xi}(X_0))\big|^r \mid \mathcal{D}_n\big]\big)^{1/r}. \tag{3.10}$$

It is well-known (Vershynin, 2018, Proposition 2.7.1) that

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}_n} \asymp \sup_{r\geq 1} r^{-1}\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_r|\mathcal{D}_n},$$

i.e., there are absolute constants C_l and C_u such that

$$0 < C_l \leq \frac{\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}}}{\sup_{r\geq 1} r^{-1}\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_r|\mathcal{D}_n}} \leq C_u < \infty.$$

3.2.2.1 Control of Δ_n^{add}

Let $\widehat{f}^{\xi}$, n_{te}, and `CEN` be as defined in Algorithm 1, and Δ_n^{add} be as defined in (3.6a).

Lemma 3.2.4 (Control of Δ_n^{add} and its expectation for losses with bounded conditional ψ_1 norm)**.** *Suppose $(X_i, Y_i), i \in \mathcal{I}_{\mathrm{te}}$ are sampled i.i.d. from P. Suppose the loss function ℓ is such that*

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}_n} \leq \widehat{\sigma}_{\xi}$$

for $(X_0, Y_0) \sim P$ and set $\widehat{\sigma}_{\Xi} := \max_{\xi\in\Xi} \widehat{\sigma}_{\xi}$. Fix any $0 < A < \infty$. Then, for `CEN` $=$ `AVG`, or `CEN` $=$ `MOM` with $\eta = n^{-A}/|\Xi|$, [2] *there exists an absolute constant $C_1 > 0$ such that*

$$\mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1\widehat{\sigma}_{\Xi} \max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\}\right) \leq n^{-A}.$$

Additionally, if for some $A > 0$, there exists a $C_2 > 0$ such that $\mathbb{P}(\widehat{\sigma}_{\Xi} \geq C_2) \leq n^{-A}$, then there exists an absolute constant $C_3 > 0$ such that

$$\mathbb{E}[\Delta_n^{\mathrm{add}}] \leq C_1C_2 \max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\} + C_3n^{-A/r}|\Xi|^{1/t} \max\left\{\sqrt{\frac{t}{n_{\mathrm{te}}}}, \frac{t}{n_{\mathrm{te}}}\right\} \max_{\xi\in\Xi} \|\widehat{\sigma}_{\xi}\|_{L_t} \tag{3.11}$$

for every $r, t \geq 2$ and $1/r + 1/t = 1$.

Lemma 3.2.5 (Control of Δ_n^{add} and its expectation for losses with bounded conditional L_2 norm)**.** *Suppose $(X_i, Y_i), i \in \mathcal{I}_{\mathrm{te}}$ are sampled i.i.d. from P. Suppose the loss function ℓ is such that*

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_2|\mathcal{D}_n} \leq \widehat{\sigma}_{\xi}$$

for $(X_0, Y_0) \sim P$ and set $\widehat{\sigma}_{\Xi} := \max_{\xi\in\Xi} \widehat{\sigma}_{\xi}$. Fix any $0 < A < \infty$. Then, for `CEN` $=$ `MOM` with $\eta = n^{-A}/|\Xi|$, there exists an absolute constant $C_1 > 0$ such that

$$\mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1\widehat{\sigma}_{\Xi}\sqrt{\frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}}\right) \leq n^{-A}. \tag{3.12}$$

Additionally, if for some $A > 0$ there exists a $C_2 > 0$ such that $\mathbb{P}(\widehat{\sigma}_{\Xi} \geq C_2) \leq n^{-A}$, then for `CEN` $=$ `MOM`,

$$\mathbb{E}\left[\Delta_n^{\mathrm{add}}\right] \leq C_1C_2\sqrt{\frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}} + C_3n^{-A/2}|\Xi|^{1/2}\sqrt{\frac{\log^2(|\Xi|n^A)}{n_{\mathrm{te}}}} \max_{\xi\in\Xi} \|\widehat{\sigma}_{\xi}\|_{L_2} \tag{3.13}$$

for some absolute constant $C_3 > 0$.

[2] See Remark 3.2.7.

Remark 3.2.6 (Comparison of assumptions for $\texttt{CEN} = \texttt{AVG}$ and $\texttt{CEN} = \texttt{MOM}$.)**.** Comparing Lemmas 3.2.4 and 3.2.5, we note that the median-of-means method of risk estimation only requires control of the L_2 moments of the loss function compared to the ψ_1 (exponential) moments of the loss function. This is not surprising given that the median-of-means was developed as a sub-Gaussian estimator of the mean, only assuming finite variance (Lemma C.8.2). The L_2 moment assumption in Lemma 3.2.5 can be further relaxed to an $L_{1+\alpha}$ moment assumption for $\alpha \in (0, 1]$ (Lugosi and Mendelson, 2019, Theorem 3) at the cost of weaker rate of convergence of Δ_n^{add}. One can, of course, replace the median-of-means estimator with any other sub-Gaussian or sub-exponential mean estimator (Catoni, 2012; Minsker, 2015; Fan et al., 2017) and obtain a similar weakening of the moment assumptions. Same remark continues to hold for Δ_n^{mul} discussed in Section 3.2.2.2.

Remark 3.2.7 (Restriction on A for $\texttt{CEN} = \texttt{MOM}$)**.** In Lemmas 3.2.4 and 3.2.5, we allow for a free parameter A. However, in order for the choice of η to be feasible in the MOM construction (see, e.g., Lemma C.8.2 in Appendix C.8), we need $B = \lceil 8\log(1/\eta)\rceil \le n_{\mathrm{te}}$, which puts the following constraint on A:

$$8\log(n^A|\Xi|) \le n_{\mathrm{te}} \quad\iff\quad A\log n \le \frac{n_{\mathrm{te}}}{8} - \log(|\Xi|) \quad\iff\quad A \le \frac{n_{\mathrm{te}}}{8\log n} - \frac{\log(|\Xi|)}{\log n}.$$

For a large enough n, this allows for a large range of A. In addition, the right hand side is large enough to imply exponentially small probability bound for the event that Δ_n^{add} is large. The same remark holds for Lemmas 3.2.9 and 3.2.10 below.

The key quantities that drive the tail probability and expectation bound on Δ_n^{add} in both Lemmas 3.2.4 and 3.2.5 are $\widehat{\sigma}_\Xi$ and $|\Xi|$. The following remark specifies the permissible growth rates on $\widehat{\sigma}_\Xi$ and $|\Xi|$ to ensure that Δ_n^{add} is asymptotically small in probability.

Remark 3.2.8 (Tolerable growth rates on $\widehat{\sigma}_\Xi$ for $\Delta_n^{\mathrm{add}} = o_p(1)$)**.** Suppose $|\Xi| \le n^S$ for some constant $S > 0$ independent of n, p. If

$$\widehat{\sigma}_\Xi = o_p\left(\sqrt{\frac{n_{\mathrm{te}}}{\log n}}\right),$$

then under the setting of Lemmas 3.2.4 and 3.2.5, $\Delta_n^{\mathrm{add}} = o_p(1)$ as $n \to \infty$. The remark follows simply by noting that the dominating term in the probabilistic bound on Δ_n^{add} in (3.12) is of order

$$\widehat{\sigma}_\Xi\sqrt{\frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}} \le \widehat{\sigma}_\Xi\sqrt{\frac{(S+A)\log n}{n_{\mathrm{te}}}} = O\left(\widehat{\sigma}_\Xi\sqrt{\frac{\log n}{n_{\mathrm{te}}}}\right).$$

See Appendix C.6.9 for feasible rates for $\widehat{\sigma}_\Xi$ to ensure that $\mathbb{E}[\Delta_n^{\mathrm{add}}] = o(1)$.

3.2.2.2 Control of Δ_n^{mul}

Moving on to Δ_n^{mul}, analogously to Lemmas 3.2.4 and 3.2.5, the following results provide high probability bounds on Δ_n^{mul} in terms of a coefficient of variation parameter κ which is the relative standard deviation of $\ell(Y_0, \widehat{f}^\xi(X_0))$ conditional on $\mathcal{D}_n$. Let $\widehat{f}^\xi$, n_{te}, $\texttt{CEN}$ be as defined Algorithm 1, and Δ_n^{mul} be as in (3.6b).

Lemma 3.2.9 (Control of Δ_n^{mul} for losses with bounded conditional ψ_1 norm)**.** *Suppose (X_j, Y_j), $j \in \mathcal{I}_{\mathrm{te}}$ are sampled i.i.d. from P. Suppose the loss function ℓ is such that*

$$\|\ell(Y_0, \widehat{f}^\xi(X_0))\|_{\psi_1|\mathcal{D}_n} \le \widehat{\sigma}_\xi \quad \textit{for } (X_0, Y_0) \sim P.$$

Define $\widehat{\kappa}_\xi = \widehat{\sigma}_\xi / R(\widehat{f}^\xi)$ and $\widehat{\kappa}_\Xi = \max_{\xi\in\Xi}\widehat{\kappa}_\xi$. Fix any $0 < A < \infty$. Then, for $\texttt{CEN} = \texttt{AVG}$, or $\texttt{CEN} = \texttt{MOM}$ with $\eta = n^{-A}/|\Xi|$,

$$\mathbb{P}\left(\Delta_n^{\mathrm{mul}} \ge C\widehat{\kappa}_\Xi \max\left\{\sqrt{\frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}}, \frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}\right\}\right) \le n^{-A}$$

for a positive constant C.

Lemma 3.2.10 (Control of Δ_n^{mul} for losses with bounded conditional L_2 norm). *Suppose (X_j, Y_j), $j \in \mathcal{I}_{\mathrm{te}}$ are sampled i.i.d. from P. Suppose the loss function ℓ is such that*

$$\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_2|\mathcal{D}_n} \le \widehat{\sigma}_{\xi} \quad \textit{for } (X_0, Y_0) \sim P.$$

Define $\widehat{\kappa}_{\xi} := \widehat{\sigma}_{\xi}/R(\widehat{f}^{\xi})$ and $\widehat{\kappa}_{\Xi} := \max_{\xi \in \Xi} \widehat{\kappa}_{\xi}$. Fix any $0 < A < \infty$. Then, for $\texttt{CEN} = \texttt{MOM}$ with $\eta = n^{-A}/|\Xi|$,

$$\mathbb{P}\left(\Delta_n^{\mathrm{mul}} \ge C\widehat{\kappa}_{\Xi}\sqrt{\frac{\log(|\Xi|n^A)}{n_{\mathrm{te}}}}\right) \le n^{-A}$$

for a positive constant C.

Remark 3.2.11 (Tolerable growth rate on $\widehat{\kappa}_{\Xi}$ for probabilistic bound). Suppose $|\Xi| \le n^S$ for some $S < \infty$. If

$$\widehat{\kappa}_{\Xi} = o_p\left(\sqrt{\frac{n_{\mathrm{te}}}{\log n}}\right),$$

then under the setting of Lemmas 3.2.9 and 3.2.10, $\Delta_n^{\mathrm{mul}} = o_p(1)$ as $n \to \infty$.

Remark 3.2.12 (Comparing the control of Δ_n^{add} versus Δ_n^{mul}). Note that from Lemmas 3.2.4 and 3.2.9, controlling Δ_n^{add} requires controlling $\widehat{\sigma}_{\Xi}$, while controlling Δ_n^{mul} requires controlling $\widehat{\kappa}_{\Xi}$. The former is on the scale of the standard deviation of the loss, while the latter is normalized standard deviation (where the normalization is with respect to the expectation of the loss). The advantage of the latter is that, even if the standard deviation diverges, the normalized standard deviation can be finite. This, in fact, happens for the case of minimum ℓ_2-norm least squares predictor when $\gamma \approx 1$, in which case the control of Δ_n^{mul} is feasible. See also the discussion in Remark 3.2.23.

Remark 3.2.13 (Choice of n_{te}). The above results hold true as long as $n_{\mathrm{te}} \to \infty$. Of course, the choice n_{te} restricts the allowable growth rate of $\widehat{\sigma}_{\Xi}$ and $\widehat{\kappa}_{\Xi}$ as discussed in Remarks 3.2.8 and 3.2.11. In our later applications in overparameterized learning, we adopt the proportional asymptotics framework in which the number of covariates to the number of observations converges to a non-zero constant. For this reason, we restrict ourselves to the choices of n_{te} such that $n_{\mathrm{te}}/n \to 0$ as $n \to \infty$; for example, one can take $n_{\mathrm{te}} = n^{\nu}$ for some $\nu < 1$. This allows us to have training models with the same limiting aspect ratio (dimension/sample size) as that of the original data without splitting. However, the larger the n_{te}, the more accurate our estimator of the prediction risk. For this reason, we suggest $n_{\mathrm{te}} = O(n/\log n)$ rather than $n_{\mathrm{te}} = n^{\nu}$.

3.2.3 Applications to loss functions

Below we consider several examples of common predictors and loss functions, and bound the corresponding conditional $\widehat{\sigma}$ parameters used in Lemmas 3.2.4 and 3.2.5, and conditional $\widehat{\kappa}$ parameters used in Lemmas 3.2.9 and 3.2.10. Recall the conditional ψ_1 and L_r norms from (3.9) and (3.10), respectively. In addition, let ψ_2 denote the ψ_2-Orlicz norm.

Recall that the quantity $\widehat{\sigma}_{\Xi}$ is the maximum of either of the two conditional norms $\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}_n}$ or $\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_2|\mathcal{D}_n}$ over $\xi \in \Xi$. Also recall that the quantity $\widehat{\kappa}_{\Xi}$ is the maximum of either of the two ratios of conditional norms $\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{\psi_1|\mathcal{D}_n}/\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_1|\mathcal{D}_n}$ or $\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_2|\mathcal{D}_n}/\|\ell(Y_0, \widehat{f}^{\xi}(X_0))\|_{L_1|\mathcal{D}_n}$ over $\xi \in \Xi$. In the following, we control each of these quantities for one of the predictors $\widehat{f}^{\xi}$, $\xi \in \Xi$, which we denote simply by $\widehat{f}$ for brevity.

3.2.3.1 Bounded classification loss functions

Proposition 3.2.14 (Generic classifier and 0-1 loss and hinge loss). *Let $\widehat{f}$ be any predictor.*

1. *Suppose $\ell(Y_0, \widehat{f}(X_0)) = \max\{0, 1 - Y_0\widehat{f}(X_0)\}$ is the hinge loss. Assume $|Y_0| \le 1$ and $|\widehat{f}(X_0)| \le 1$. Then,*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} \le 2, \quad \textit{and} \quad \|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n} \le 2.$$

2. *Suppose $\ell(Y_0, \widehat{f}(X_0)) = \mathbb{1}\{Y_0 \neq \widehat{f}(X_0)\}$ is the 0-1 loss. Then,*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} \le 1, \quad \text{and} \quad \|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n} \le 1. \tag{3.14}$$

More generally, any loss function that is bounded by 1 *satisfies* (3.14).

Proposition 3.2.14 implies that the parameter $\widehat{\sigma}_\Xi$ is bounded by 1 (with probability 1) for any collection of bounded classifiers $\{\widehat{f}^\xi, \xi \in \Xi\}$. Hence, Lemmas 3.2.4 and 3.2.5 imply that $\Delta_n^{\mathrm{add}} = O_p(\sqrt{\log(|\Xi|)/n_{\mathrm{te}}})$. Therefore, the additive form of oracle inequality from Proposition 3.2.1 can be used to conclude the following result.

Theorem 3.2.15 (Oracle inequality for arbitrary classifiers)**.** *For any collection of classifiers $\{\widehat{f}^\xi, \xi \in \Xi\}$ with $\log(|\Xi|) = o(n_{\mathrm{te}})$ and the loss being the mis-classification or hinge loss with bounded response and predictor,*

$$\left|R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi\in\Xi} R(\widehat{f}^\xi)\right| \;=\; O_p\left(\sqrt{\frac{\log(|\Xi|)}{n_{\mathrm{te}}}}\right).$$

Theorem 3.2.15 can be used to argue that tuning of hyperparameters in an arbitrary classifier using Algorithm 1 leads to an "optimal" classifier under the $0-1$ or hinge loss. Moreover, Proposition 3.2.14 extends to arbitrary bounded loss functions.

For logistic or the cross-entropy loss, being unbounded, is not covered by Proposition 3.2.14. However, we can use the multiplicative form of the oracle risk inequality (3.8) as done in the next section in Proposition 3.2.18.

3.2.3.2 Unbounded regression loss functions

Proposition 3.2.16 (Linear predictor and square loss)**.** *Let $\widehat{f}$ be a linear predictor, i.e., for any $x_0 \in \mathbb{R}^p$, $\widehat{f}(x_0) = x_0^\top \widehat{\beta}$ for some estimator $\widehat{\beta} \in \mathbb{R}^p$ fitted on $\mathcal{D}_n$. Suppose $\ell(Y_0, \widehat{f}(X_0)) = (Y_0 - \widehat{f}(X_0))^2$ is the square loss. Let $(X_0, Y_0) \sim P$. Assume $\mathbb{E}[X_0] = 0_p$ and let $\Sigma := \mathbb{E}[X_0 X_0^\top]$. Then, the following statements hold:*

1. *If $(X_0, Y_0) \in \mathbb{R}^p \times \mathbb{R}$ satisfies $\psi_2 - L_2$ equivalence, i.e., $\|aY_0 + b^\top X_0\|_{\psi_2} \le \tau \|aY_0 + b^\top X_0\|_{L_2}$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$, then*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} \le \tau^2 \inf_{\beta\in\mathbb{R}^p} (\|Y_0 - X_0^\top \beta\|_{\psi_2} + \|\widehat{\beta} - \beta\|_\Sigma)^2, \quad \text{and} \quad \frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \tau^2. \tag{3.15}$$

2. *If (X_0, Y_0) satisfies the $L_4 - L_2$ equivalence, i.e., $\|aY_0 + b^\top X_0\|_{L_4} \le \tau \|aY_0 + b^\top X_0\|_{L_2}$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$, then*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n} \le \tau^2 \inf_{\beta\in\mathbb{R}^p} (\|Y_0 - X_0^\top \beta\|_{L_2} + \|\widehat{\beta} - \beta\|_\Sigma)^2, \quad \text{and} \quad \frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \tau^2. \tag{3.16}$$

Proposition 3.2.17 (Linear predictor and absolute loss)**.** *Let $\widehat{f}$ be a linear predictor corresponding to estimator $\widehat{\beta}$ fitted on $\mathcal{D}_n$. Suppose $\ell(Y_0, \widehat{f}(X_0)) = |Y_0 - X_0^\top \widehat{\beta}|$ is the absolute loss. Let $(X_0, Y_0) \sim P$. Assume $\mathbb{E}[X_0] = 0_p$ and let $\Sigma := \mathbb{E}[X_0 X_0^\top]$. Then, the following statements hold:*

1. *If $(X_0, Y_0) \in \mathbb{R}^p \times \mathbb{R}$ satisfies $\psi_1 - L_1$ equivalence, i.e., $\|aY_0 + b^\top X_0\|_{\psi_1} \le \tau \|aY_0 + b^\top X_0\|_{L_1}$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$, then*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} \le \tau \inf_{\beta\in\mathbb{R}^p} (\|Y_0 - X_0^\top \beta\|_{L_1} + \|X_0^\top(\widehat{\beta} - \beta)\|_{L_1|\mathcal{D}_n}), \quad \frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \tau. \tag{3.17}$$

2. *If (X_0, Y_0) satisfies $L_2 - L_1$ equivalence, i.e., $\|aY_0 + b^\top X_0\|_{L_2} \le \tau\|aY_0 + b^\top X_0\|_{L_1}$, for all $a \in \mathbb{R}^p$ and $b \in \mathbb{R}p$, then*

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n} \le \tau \inf_{\beta\in\mathbb{R}^p}(\|Y_0 - X_0^\top\beta\|_{L_1} + \|X_0^\top(\widehat{\beta} - \beta)\|_{L_1|\mathcal{D}_n}), \quad \frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \tau. \tag{3.18}$$

Proposition 3.2.18 (Linear predictor and logistic loss)**.** *Let $Y_0 \in [0, 1]$ almost surely. Let $\widehat{f}$ be a linear predictor corresponding to an estimator $\widehat{\beta}$ fitted on $\mathcal{D}_n$. Suppose $\ell(Y_0, \widehat{f}(X_0))$ is the logistic or cross-entropy loss:*

$$\ell(Y_0, \widehat{f}(X_0)) = -Y_0 \log\left(\frac{1}{1 + e^{-X_0^\top\widehat{\beta}}}\right) - (1 - Y_0)\log\left(1 - \frac{1}{1 + e^{-X_0^\top\widehat{\beta}}}\right).$$

Assume there exists $p_{\min} \in (0, 1)$ such that $p_{\min} \le \mathbb{E}[Y_0 \mid X_0 = x] \le 1 - p_{\min}$ for all x. Then, the following statements hold:

1. *If $X_0 \in \mathbb{R}^p$ satisfies $\psi_1 - L_1$ equivalence, i.e., $\|b^\top X_0\|_{\psi_1} \le \tau\|b^\top X_0\|_{L_1}$ for all $b \in \mathbb{R}^p$, then*

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le 2\tau p_{\min}^{-1}.$$

2. *If $X_0 \in \mathbb{R}^p$ satisfies $L_2 - L_1$ equivalence, i.e., $\|b^\top X_0\|_{L_2} \le \tau\|b^\top X_0\|_{L_1}$ for all $b \in \mathbb{R}^p$, then*

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le 2\tau p_{\min}^{-1}.$$

In the remarks that follow we offer a discussion of the different types of norm equivalences assumed in Propositions 3.2.16 to 3.2.18.

Remark 3.2.19 (Discussion of $\psi_2 - L_2$ and $L_4 - L_2$ equivalences)**.** A centered random vector $Z \in \mathbb{R}^p$ is said to be τ-sub-Gaussian if

$$\sup_{a\in\mathbb{R}^p} \frac{\|a^\top Z\|_{\psi_2}}{\|a\|_{\Sigma_Z}} \le \tau < \infty \quad \text{where} \quad \Sigma_Z := \mathrm{Cov}(Z). \tag{3.19}$$

See for instance Definition 1.2 and Remark 1.3 of Mendelson and Zhivotovskiy (2020) for more details. The $L_4 - L_2$ equivalence assumption is popular in robust estimation of covariance matrices. See, for example, Minsker and Wei (2020); Minsker (2018); Mendelson and Zhivotovskiy (2020). This is weaker than the sub-Gaussianity assumption in (3.19) in the sense that $\psi_2 - L_2$ equivalence implies $L_4 - L_2$ equivalence. This follows from the well-known fact that

$$C_l \le \frac{\|W\|_{\psi_2}}{\sup_{r\ge1} r^{-1/2}\|W\|_{L_r}} \le C_u$$

for some universal constants C_l and C_u; see Vershynin (2018, Proposition 2.5.2). The $L_4 - L_2$ equivalence assumption is also weaker than a commonly used assumption in the random matrix theory (RMT) literature. In RMT, one typically assumes features of the form $\Sigma^{1/2}Z$, where Z have i.i.d. entries and Σ is feature covariance matrix. If the components of Z are independent and have bounded kurtosis, then this typical RMT assumption implies $L_4 - L_2$ equivalence.

Remark 3.2.20 (Discussion of $\psi_1 - L_1$ and $L_2 - L_1$ equivalences)**.** In Remark 3.2.19, we have given examples of distributions that satisfy $\psi_2 - L_2$ and/or $L_4 - L_2$ equivalence. From the fact that, for any random variable W, the function $r \mapsto \log\mathbb{E}[|W|^r]$ $(r \ge 1)$ is convex (Loeve, 2017, Section 9, inequality (b)), we can conclude that $\psi_2 - L_2$ equivalence implies $\psi_1 - L_1$ equivalence, and $L_4 - L_2$ equivalence implies $L_2 - L_1$ equivalence; see Proposition C.6.21. We further note that distributions satisfying $\psi_1 - L_2$

equivalence also satisfy $\psi_1 - L_1$ and $L_2 - L_1$ equivalence. See Figure C.7 for a visual summary of these equivalences and their proofs in Appendix C.6.10.

We will now discuss other distributions that satisfy $\psi_1 - L_2$ equivalence (which implies $\psi_1 - L_1$ equivalence). A random vector $Z \in \mathbb{R}^q$ is log-concave if for any two measurable subsets A and B of $\mathbb{R}^q$, and for any $\theta \in [0, 1]$,

$$\log \mathbb{P}(Z \in \theta A + (1-\theta)B) \;\geq\; \theta \cdot \mathbb{P}(Z \in A) + (1-\theta) \cdot \mathbb{P}(Z \in B),$$

whenever the set $\theta A + (1-\theta)B = \{\theta x_1 + (1-\theta)x_2 : x_1 \in A, x_2 \in B\}$ is measurable; see Definition 2.2 of Adamczak et al. (2010). There exist a universal constant C such that all log-concave random vectors $Z \in \mathbb{R}^q$ with mean 0 satisfy

$$\|a^\top Z\|_{\psi_1} \leq C\|a^\top Z\|_{L_1}$$

for all $a \in \mathbb{R}^q$. This follows from the results of Adamczak et al. (2010) and Latała (1999); see also Nayar and Oleszkiewicz (2012, Corollary 3), Proposition 2.1.1 of Warsaw (2003), and Proposition 2.14 of Ledoux (2001). In particular, Lemma 2.3 of Adamczak et al. (2010) implies that there exists a universal constant C such that for all $a \in \mathbb{R}^q$

$$\|a^\top Z\|_{\psi_1} \leq C\|a^\top Z\|_{L_2}.$$

Finally, note that since $L_4 - L_2$ equivalence implies $L_2 - L_1$ equivalence, and the RMT features as described in Remark 3.2.19 satisfy $L_4 - L_2$ equivalence, they in turn satisfy $L_2 - L_1$ equivalence.

Remark 3.2.21 (Model-free nature of assumptions)**.** It is worth emphasizing that we do not require a well-specified linear model for Propositions 3.2.16 and 3.2.17. Hence, our results are model agnostic.

Propositions 3.2.16 to 3.2.18 imply that, under the stated assumptions, for any collection of predictors $\{\widehat{f}^\xi : \widehat{f}^\xi(x) = x^\top \widehat{\beta}^\xi, \xi \in \Xi\}$, $\widehat{\kappa}_\Xi$ is bounded if (X_0, Y_0) satisfies a requisite moment equivalence assumption. On the other hand, the control of $\widehat{\sigma}_\Xi$ depends crucially on behavior of $\max_{\xi\in\Xi} \|\widehat{\beta}^\xi - \beta_0\|_\Sigma$. Because $\widehat{\kappa}_\Xi$ is bounded with probability 1, Lemmas 3.2.9 and 3.2.10 can be used to conclude $\Delta_n^{\mathrm{mul}} = O_p(K_{X,Y}\sqrt{\log(|\Xi|)/n_{\mathrm{te}}})$, where $K_{X,Y}$ is the constant in the moment equivalence. Hence, the multiplicative form of the oracle inequality from Proposition 3.2.1 can used to conclude the following general result for an arbitrary collection of linear predictors.

Theorem 3.2.22 (Oracle inequality for arbitrary linear predictors)**.** *Fix any collection of predictors $\{\widehat{f}^\xi : \widehat{f}^\xi(x) = x^\top \widehat{\beta}^\xi, \xi \in \Xi\}$. Let $\widehat{f}^{\mathrm{cv}}$ be the output of Algorithm 1 with $\widehat{f}^\xi, \xi \in \Xi$ as the ingredient predictors. Suppose one of the following conditions hold:*

1. *The loss is squared error, (X_0, Y_0) satisfies $\psi_2 - L_2$ equivalence when $\texttt{CEN} = \texttt{AVE}$ and $L_4 - L_2$ equivalence when $\texttt{CEN} = \texttt{MOM}$.*

2. *The loss is absolute error, (X_0, Y_0) satisfies $\psi_1 - L_2$ equivalence when $\texttt{CEN} = \texttt{AVE}$ and $L_2 - L_1$ equivalence when $\texttt{CEN} = \texttt{MOM}$.*

3. *The loss is logistic error and $p_{\min} \leq \mathbb{E}[Y_0 \mid X = x] \leq 1 - p_{\min}$ for some $p_{\min} \in (0, 1)$, X_0 satisfies $\psi_1 - L_1$ equivalence when $\texttt{CEN} = \texttt{AVE}$ and $L_2 - L_1$ equivalence when $\texttt{CEN} = \texttt{MOM}$.*

Then, there exists a constant C depending only on the moment equivalence condition such that for any $A > 0$ and for $\widehat{f}^{\mathrm{cv}}$ returned by Algorithm 1, we have with probability at least $1 - n^{-A}$,

$$\left|\frac{R(\widehat{f}^{\mathrm{cv}})}{\min_{\xi\in\Xi} R(\widehat{f}^\xi)} - 1\right| \;\leq\; C\sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}}.$$

Here, for $\texttt{CEN} = \texttt{AVE}$, there are no restrictions on A. For $\texttt{CEN} = \texttt{MOM}$, we need η to be $n^{-A}/|\Xi|$ in Algorithm 1.

Theorem 3.2.22 implies that a multiplicative form of oracle inequality holds true for any collection of linear predictors with three commonly used loss functions – square, absolute, or logistic loss – under certain moment equivalence conditions on the underlying data. It is worth stressing that Theorem 3.2.22 does not require any parametric model assumption on the data. The moment equivalence conditions required are quite mild as indicated in Remarks 3.2.19 and 3.2.20. Theorem 3.2.22 can be used to argue that tuning of hyperparameters for an arbitrary linear predictor using Algorithm 1 leads to an "optimal" linear predictor. In particular, this includes variable selection in linear regression, and penalty selection in ridge regression or lasso.

Remark 3.2.23 (Divergence of Δ_n^{add})**.** As mentioned above, control of $\widehat{\sigma}_\Xi$ for a collection of linear predictors depends crucially on $\max_{\xi\in\Xi}\|\widehat{\beta}^\xi - \beta_0\|_\Sigma$. Controlling this maximum is not difficult in the "low-dimensional" regime, where the number of features is asymptotically negligible compared to the number of observations. If, however, the collection of linear predictors involves the least squares estimator with the number of features approximately same as the number of observations, then Corollaries 1 and 3 of Hastie et al. (2022) implies that $\max_{\xi\in\Xi}\|\widehat{\beta}^\xi - \beta_0\|_\Sigma \to \infty$ almost surely under some regularity assumptions. The case of number of features approximately the same as the number of observations can be seen in the problem of tuning the number of basis functions in series regression (see also Mei and Montanari (2022); Bartlett et al. (2021) for similar results on random features regression and kernel regression). In this case, Δ_n^{add} diverges while Δ_n^{mul} is bounded hinting the advantages of the multiplicative form of the oracle inequality over the additive form.

3.2.4 Illustrative prediction procedures

In the following two sections, we provide concrete applications of the results from this section in the context of overparameterized learning. The main motivation of our applications is to synthesize a predictor whose prediction risk is approximately monotonically non-increasing in the sample size. Although this represents the basic idea of "more data does not hurt," many commonly studied predictors such as minimum ℓ_2-norm least squares, minimum ℓ_1-norm least squares in the overparameterized regime do not satisfy this property. In the following sections, we will provide two different ways to synthesize a predictor with this property starting from any given base prediction procedure.

Definition 3.2.24 (Prediction procedure)**.** A prediction procedure, denoted by $\widetilde{f}$ is a real-valued map, with two arguments: (1) a feature vector; and (2) a dataset. If $\mathcal{D}_m = \{(X_i, Y_i) : 1 \le i \le m\}$ represents a dataset of size m, then $\widetilde{f}(x; \mathcal{D}_m)$ represents prediction at x of the prediction procedure $\widetilde{f}$ trained on the dataset $\mathcal{D}_m$.

Example 3.2.25 (Minimum ℓ_2-norm least squares prediction procedure)**.** Suppose $\mathcal{D}_m = \{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le m\}$. The minimum ℓ_2-norm least squares (MN2LS) estimator trained on $\mathcal{D}_m$ is defined as

$$\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_m) := \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p}\left\{\|\beta\|_2 : \beta \text{ is a minimizer of the function } \theta \mapsto \sum_{i=1}^m (Y_i - X_i^\top\theta)^2\right\}.$$

The estimator can be written explicitly in terms of (X_i, Y_i), $i = 1, \dots, m$ as

$$\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_m) = \left(\frac{1}{m}\sum_{i=1}^m X_i X_i^\top\right)^\dagger \left(\frac{1}{m}\sum_{i=1}^m X_i Y_i\right), \tag{3.20}$$

where $A^\dagger$ denotes the Moore-Penrose inverse of A. It is also the "ridgeless" least squares estimator because of the fact that $\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_m) = \lim_{\lambda\to 0^+}\widetilde{\beta}_{\mathrm{ridge},\lambda}(\mathcal{D}_m)$, where $\widetilde{\beta}_{\mathrm{ridge},\lambda}(\mathcal{D}_m)$ is the ridge estimator at a regularization parameter $\lambda > 0$ trained on $\mathcal{D}_m$:

$$\widetilde{\beta}_{\mathrm{ridge},\lambda}(\mathcal{D}_m) := \operatorname*{arg\,min}_{\theta\in\mathbb{R}^p}\left\{\frac{1}{m}\sum_{i=1}^m (Y_i - X_i^\top\theta)^2 + \lambda\|\theta\|_2^2\right\}. \tag{3.21}$$

The MN2LS estimator has been attracted attention in the last few years and its risk behavior has been studied by Bartlett et al. (2020); Belkin et al. (2020); Hastie et al. (2022); Muthukumar et al. (2020), among others. The MN2LS predictor is now defined as

$$\widetilde{f}_{\mathrm{mn2}}(x; \mathcal{D}) := x^\top\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}), \tag{3.22}$$

for any vector $x \in \mathbb{R}^p$ and dataset $\mathcal{D}$ containing random vectors from $\mathbb{R}^p \times \mathbb{R}$.

Example 3.2.26 (Minimum ℓ_1-norm least squares prediction procedure)**.** Suppose $\mathcal{D}_m = \{(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le m\}$. The minimum ℓ_1-norm least squares (MN1LS) estimator trained on $\mathcal{D}_m$ is defined as

$$\widetilde{\beta}_{\mathrm{mn1}}(\mathcal{D}_m) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \|\beta\|_1 : \beta \text{ is a minimizer of the function } \theta \mapsto \sum_{i=1}^{m} (Y_i - X_i^\top \theta)^2 \right\}. \tag{3.23}$$

It is also the "lassoless" least squares estimator because of the fact that $\widetilde{\beta}_{\mathrm{mn1}}(\mathcal{D}_m) = \lim_{\lambda \to 0^+} \widetilde{\beta}_{\mathrm{lasso},\lambda}$, where $\widetilde{\beta}_{\mathrm{lasso},\lambda}(\mathcal{D}_m)$ is the lasso estimator at a regularization parameter $\lambda > 0$ trained on $\mathcal{D}_m$:

$$\widetilde{\beta}_{\mathrm{lasso},\lambda}(\mathcal{D}_m) := \operatorname*{arg\,min}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2m} \sum_{i=1}^{m} (Y_i - X_i^\top \theta)^2 + \lambda \|\theta\|_1 \right\}. \tag{3.24}$$

The MN1LS estimator connects naturally to the basis pursuit estimator in compressed sensing literature (e.g. Candes and Tao (2006); Donoho (2006)) and its risk in the proportional regime has been recently analyzed in Mitra (2019); Li and Wei (2021). The MN1LS predictor is now defined as

$$\widetilde{f}_{\mathrm{mn1}}(x; \mathcal{D}) := x^\top \widetilde{\beta}_{\mathrm{mn1}}(\mathcal{D}), \tag{3.25}$$

for any vector $x \in \mathbb{R}^p$ and dataset $\mathcal{D}$ containing random vectors from $\mathbb{R}^p \times \mathbb{R}$.

Note that the MN2LS and MN1LS estimators coincide when there is a unique minimizer of the function $\theta \mapsto \sum_{i=1}^{m} (Y_i - X_i^\top \theta)^2$, in which case both the estimators become the least squares estimator.

We focus mostly on the case of linear predictors and squared error loss, although all our results are easily extendable to general predictors and loss functions. (See Remark 3.3.16 for more details.)

3.3 Application 1: Zero-step prediction procedure

3.3.1 Motivation

Suppose R_n represents the prediction risk of a given prediction procedure $\widetilde{f}$ on a dataset containing n i.i.d. observations. It is desirable that R_n as a function of $n \ge 1$ is non-increasing. As described above, this however may not hold for an arbitrary procedure $\widetilde{f}$. If we have access to R_k for $1 \le k \le n$, then one could just return the predictor obtained by applying the prediction procedure $\widetilde{f}$ on a subset of $k_n^\star$ i.i.d. observations where $k_n^\star = \arg\min\{R_k : 1 \le k \le n\}$. This procedure, (denoted by, say) $\widetilde{f}^{\mathrm{zs}\star}$, essentially returns a predictor whose risk is the largest non-increasing function that is below the risk of $\widetilde{f}$; see Figure 3.2 for an illustration.

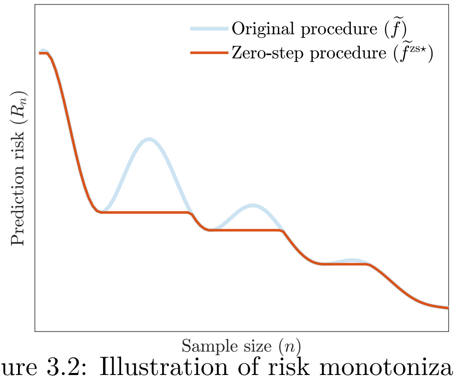


Figure 3.2: Illustration of risk monotonization.

It is trivially true that the risk of the prediction procedure $\widetilde{f}^{\mathrm{zs}\star}$ as a function of $n \ge 1$ is non-decreasing and its risk at the sample size n is given by $\min_{k \le n} R_k$. This procedure $\widetilde{f}^{\mathrm{zs}\star}$ is, however, not actionable in practice because one seldom has access to the true risk R_n of $\widetilde{f}$.

The goal of this section is to develop a prediction procedure $\widehat{f}^{\mathrm{zs}}$ starting with the base prediction procedure $\widetilde{f}$ such that the risk of $\widehat{f}^{\mathrm{zs}}$ is the largest non-increasing function that is below the risk of $\widetilde{f}$ (asymptotically). We achieve this goal by applying Algorithm 1 with the ingredient predictors being the prediction procedure $\widetilde{f}$ applied on the subsets of the original data of varying sample sizes.

Remark 3.3.1 (Conditional versus unconditional risk)**.** There are two versions of the prediction risk R_n that one can consider: conditional (on the dataset $\mathcal{D}_n$) and unconditional/non-stochastic. The conditional risk is not just a function of sample size, but also of the data $\mathcal{D}_n$. Hence, the conditional risk R_k, for $k \le n$, is ill-defined as just a function of the sample size k. Therefore, the motivation above should be considered with respect to a non-stochastic approximation of the conditional risk. See Section 3.3.3 for a precise definition of a non-stochastic approximation of the conditional risk which respect to which we talk of risk monotonization in the sample size.

3.3.2 Formal description

Formally, let the original dataset be denoted by $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$. As in Algorithm 1, consider the training and testing datasets $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$, respectively. Note that our choice of n_{te} as described in Remark 3.2.13 satisfies $n_{\mathrm{te}} = o(n)$, and hence, the risk of $\widetilde{f}$ trained on $\mathcal{D}_{\mathrm{tr}}$ is expected to be asymptotically the same as the risk of $\widetilde{f}$ trained on $\mathcal{D}_n$.

To achieve the goal described in Section 3.3.1, one can define the ingredient predictors required in Algorithm 1 as follows: Let $\mathcal{D}_{\mathrm{tr}}^k$ denote a subset of $\mathcal{D}_{\mathrm{tr}}$ with $n_{\mathrm{tr}} - k$ observations for $1 \le k \le n_{\mathrm{tr}}$. For $\Xi_n = \{1, 2, \dots, n_{\mathrm{tr}} - 1\}$ and $\xi \in \Xi_n$, define $\widetilde{f}^{\xi}(x) = \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^{\xi})$ as the predictor obtained by training $\widetilde{f}$ on $\mathcal{D}_{\mathrm{tr}}^{\xi}$. Proposition 3.2.1 along with Lemmas 3.2.4 and 3.2.5 and Lemmas 3.2.9 and 3.2.10 can be used to imply that $\widehat{f}^{\mathrm{cv}}$ thus obtained has a non-increasing risk as a function of the sample size.

There are two important points to note here:

1. The external randomness of choosing a subset $\mathcal{D}_{\mathrm{tr}}^{\xi} \subseteq \mathcal{D}_n$ of size ξ. Observe that there are $\binom{n_{\mathrm{tr}}}{\xi}$ different subsets each with $n_{\mathrm{tr}} - \xi$ i.i.d. observations. Asymptotically, the prediction risk of $\widetilde{f}$ trained on any of these subsets would be the same. To reduce such external randomness and make use of many different subsets of the same size, we take the ingredient predictor $\widehat{f}^{\xi}$ to be:

$$\widehat{f}^{\xi}(x) = \frac{1}{M} \sum_{j=1}^{M} \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^{\xi, j}), \tag{3.26}$$

   where $\mathcal{D}_{\mathrm{tr}}^{\xi,j}$, $1 \le j \le M$ are M sets drawn independently (with replacement) from the collection of $\binom{n_{\mathrm{tr}}}{\xi}$ [3] subsets of $\mathcal{D}_{\mathrm{tr}}$ of size $n_{\mathrm{tr}} - \xi$. With $M = \infty$, $\widehat{f}^{\xi}$ becomes the average of $\widetilde{f}$ trained on all possible subsets of $\mathcal{D}_{\mathrm{tr}}$ of size $n_{\mathrm{tr}} - \xi$. This choice of M removes any potential external randomness in defining $\widehat{f}^{\xi}$. The choice of $M = 1$ has the largest amount of external randomness. Based on the theory of U-statistics (Serfling, 2009, Chapter 5), we expect the choice $M = \infty$ to yield a predictor with the smallest variance; see (3.63). Observe that the expected value $\widehat{f}^{\xi}(x)$ remains constant as M changes because the distribution of $\mathcal{D}_{\mathrm{tr}}^{\xi,j}$ remains identical across $j \ge 1$. However, the computation of $\widehat{f}^{\xi}$ with $M = \infty$ is infeasible, and hence, we use a finite $M \ge 1$.

2. In the description above, we have n_{tr} predictors to use in Algorithm 1. Note that the risk of a predictor trained on $m + 1$ observations is asymptotically no different from that of a predictor trained on m observations. The same comment holds true for predictors trained on $m + o(m)$ and m observations. For this reason, we can replace $\Xi_n = \{1, 2, \dots, n_{\mathrm{tr}} - 1\}$ with

$$\Xi_n = \left\{1, 2, \dots, \left\lceil \frac{n_{\mathrm{tr}}}{\lfloor n^{\nu} \rfloor} - 2 \right\rceil \right\}^{[4]}, \quad \text{for some } \nu \in (0, 1), \tag{3.27}$$

   and consider predictors obtained by training $\widetilde{f}$ on subsets of sizes $n_{\mathrm{tr}} - \xi \lfloor n^{\nu} \rfloor$ for $\xi \in \Xi_n$. This helps in reducing the computational cost of obtaining $\widehat{f}^{\mathrm{cv}}$ using Algorithm 1. This further helps in the theoretical properties of $\widehat{f}^{\mathrm{cv}}$ in our application of union bound in the results of Section 3.2.

[3] Here, $\binom{n}{r}$ denotes the binomial coefficient representing the number of distinct ways to pick r elements from a set of n elements for positive integers n and r.

[4] The subtraction of 2 in right end point in the definition (3.27) of Ξ_n is for technical reasons.

Taking into account the remarks above, with Ξ as in (3.27), for $\xi \in \Xi_n$, we define $\widehat{f}^{\xi}$ as in (3.26), but with an important change that $\mathcal{D}_{\mathrm{tr}}^{\xi,j}$, $1 \le j \le M$, now represent randomly drawn subsets of $\mathcal{D}_{\mathrm{tr}}$ of size $n_\xi = n_{\mathrm{tr}} - \xi \lfloor n^\nu \rfloor$. The ingredient predictors used in Algorithm 1 are given by $\widehat{f}^{\xi}$, $\xi \in \Xi_n$. We call the resulting predictor obtained from Algorithm 1 as the zero-step predictor based on $\widetilde{f}$ and we denote the corresponding prediction procedure to be $\widehat{f}^{\mathrm{zs}}$. The zero-step procedure is summarized in Algorithm 2.

Algorithm 2 Zero-step procedure

Inputs:

- all inputs of Algorithm 1 other than the index set Ξ;
- a positive integer M.

Output:

- a predictor $\widehat{f}^{\mathrm{zs}}$

Procedure:

1. Let $n_{\mathrm{tr}} = n - n_{\mathrm{te}}$. Construct an index set Ξ_n per (3.27).
2. Construct train and test sets $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$ per Step 1 of Algorithm 1.
3. Let $n_\xi = n_{\mathrm{tr}} - \xi \lfloor n^\nu \rfloor$. For each $\xi \in \Xi_n$ and $j = 1, \dots, M$, draw random subsets $\mathcal{D}_{\mathrm{tr}}^{\xi,j}$ of size n_ξ from $\mathcal{D}_{\mathrm{tr}}$. For each $\xi \in \Xi$, fit predictors $\widehat{f}^{\xi}$ per (3.26) using prediction procedure $\widetilde{f}$ and $\{\mathcal{D}_{\mathrm{tr}}^{\xi,j} : 1 \le j \le M\}$.
4. Run Steps 3–5 of Algorithm 1 using index set $\Xi = \Xi_n$ and set of predictors $\{\widehat{f}^{\xi}, \xi \in \Xi\}$.
5. Return $\widehat{f}^{\mathrm{zs}}$ as the resulting $\widehat{f}^{\mathrm{cv}}$ from Algorithm 1.

3.3.3 Risk behavior of $\widehat{f}^{\mathrm{zs}}$

As alluded to before, in order to talk about risk monotonization, one needs to consider a non-stochastic approximation to the conditional risk that depends only on the prediction procedure, the sample size, and properties of the data distribution. The definition below makes this precise.

Definition 3.3.2 (Deterministic approximation of conditional prediction risk)**.** For any prediction procedure $\widetilde{f}$, we call a map $R^{\det}(\cdot; \widetilde{f}) : \mathbb{N} \to \mathbb{R}_{\ge 0}$ a deterministic (or non-stochastic) approximation of the conditional risk of $\widetilde{f}$ if for all datasets $\mathcal{D}_m$ of m i.i.d. random vectors,

$$\frac{|R(\widetilde{f}(\cdot; \mathcal{D}_m)) - R^{\det}(m; \widetilde{f})|}{R^{\det}(m; \widetilde{f})} = o_p(1), \tag{3.28}$$

as $m \to \infty$. (Recall that $R(\widetilde{f}(\cdot, \mathcal{D}_m)) = \int \ell(y; \widetilde{f}(x; \mathcal{D}_m)) \mathrm{d}P(x, y)$.)

It is important to recognize that $R^{\det}(m; \widehat{f})$ is only a function of the sample size m, the prediction *procedure* $\widetilde{f}$, and the underlying distribution P, and not the dataset $\mathcal{D}_m$. Note that we do not necessarily require $R^{\det}(m; \widetilde{f})$ to be the expected value of $R(\widetilde{f}(\cdot; \mathcal{D}_m))$. Furthermore, a non-asymptotic approximation $R^{\det}(\cdot; \widetilde{f})$ of the conditional risk may not be unique.

Remark 3.3.3 (Relative convergence in Definition 3.3.2)**.** In (3.28), the division by $R^{\det}(m; \widetilde{f})$ ensures that the deterministic approximation to the conditional risk of $\widetilde{f}(\cdot; \mathcal{D}_m)$ is non-trivial (i.e., non-zero) even if the conditional risk converges in probability to zero. If the conditional risk is bounded away from zero, asymptotically, then (3.28) is trivially implied by

$$|R(\widetilde{f}(\cdot;\mathcal{D}_m)) - R^{\mathrm{det}}(m;\widetilde{f})| = o_p(1),$$

as $m \to \infty$. In most settings of overparameterized learning, the conditional prediction risk is asymptotically bounded away from zero (see (3.36), for example).

Because $|\Xi_n| \le n$, the results of Section 3.2 imply that with appropriate choices of `CEN` and η in Algorithm 1 we obtain $\widehat{f}^{\mathrm{zs}}$ that satisfies the following risk bound:

$$R(\widehat{f}^{\mathrm{zs}}) = \begin{cases} \min_{\xi\in\Xi_n} R(\widehat{f}^{\xi}) + O_p(1)\sqrt{\log n/n_{\mathrm{te}}} & \text{if } \widehat{\sigma}_{\Xi} = O_p(1) \\ \min_{\xi\in\Xi_n} R(\widehat{f}^{\xi})\big(1 + O_p(1)\sqrt{\log n/n_{\mathrm{te}}}\big) & \text{if } \widehat{\kappa}_{\xi} = O_p(1). \end{cases} \tag{3.29}$$

Assume now there exists a function $R^{\mathrm{det}} : \mathbb{N} \to \mathbb{R}_{\ge 0}$ such that the following holds:

$$\lim_{n\to\infty} \sup_{\xi_n\in\Xi_n} \mathbb{P}\left(\frac{|R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi_n,j})) - R^{\mathrm{det}}(n_{\xi_n};\widetilde{f})|}{R^{\mathrm{det}}(n_{\xi_n};\widetilde{f})} > \epsilon\right) = 0 \quad \text{for all } \epsilon > 0. \tag{DET}$$

Recall that $\mathcal{D}_{\mathrm{tr}}^{\xi_n,j}$ for $1 \le j \le n$ are identically distributed, and hence, $\widetilde{f}(\cdot,\mathcal{D}_{\mathrm{tr}}^{\xi_n,j})$ are also identically distributed predictors. This implies that assuming (DET) for $j = 1$ is the same as assuming it for all $1 \le j \le M$. Note that (DET) is essentially the same as (3.28), but with a different sequence of sample sizes $\{n_{\xi_n}\}_{n\ge 1}$ with $\xi_n \in \Xi_n$. In accordance with our goal of monotonizing the non-stochastic approximation $R^{\mathrm{det}}(\cdot;\widetilde{f})$ of the prediction procedure $\widetilde{f}$, we aim to show that the zero-step prediction procedure $\widehat{f}^{\mathrm{zs}}$ has its conditional prediction risk approximated by $\min_{\xi\in\Xi_n} R^{\mathrm{det}}(n_\xi;\widetilde{f})$. For notational convenience, set

$$R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f}) \ := \ \min_{\xi\in\Xi_n} R^{\mathrm{det}}(n_\xi;\widetilde{f}) \quad \text{and} \quad \xi_n^\star \ \in \ \operatorname*{arg\,min}_{\xi\in\Xi_n} R^{\mathrm{det}}(n_\xi;\widetilde{f}). \tag{3.30}$$

Note the notation above is meant to reflect that the index $\xi_n^\star$ can be chosen to be any element of the minimizing set. If $\Xi_n = \{1,\dots,n_{\mathrm{tr}}-1\}$, and $\nu = 0$, then $R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f}) = \min\{R^{\mathrm{det}}(k;\widetilde{f}) : 1 \le k \le n_{\mathrm{tr}} - 1\}$. Although it might be tempting to take $\Xi_n = \{1,\dots,n_{\mathrm{tr}}-1\}$ and $\nu = 0$, instead of the one in (3.27), assumption (DET) for all non-stochastic sequences $\{n_{\xi_n}\}_{n\ge1}$ with $\xi_n \in \Xi_n$ becomes almost certainly unreasonable. To see this, observe that $\xi_n = n_{\mathrm{tr}} - 1$ belongs to Ξ_n for every n, and for this choice, $n_{\xi_n} = 1$. Hence, the predictor $\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})$ is computed based on one observation, and cannot satisfy (DET). In the following calculations, however, we only require assumption (DET) for the non-stochastic sequence $\{\xi_n^\star\}_{n\ge1}$. If $n_{\xi_n^\star}$ is known to diverge to ∞ and the distribution of the data stays constant, then assumption (DET) is reasonable and is exactly the same as the existence of a deterministic approximation to the conditional risk of $\widetilde{f}$ in the sense of Definition 3.3.2. In this favorable case of $n_{\xi_n^\star}$ diverging to ∞ with n, one can take $\Xi_n = \{1,\dots,n_{\mathrm{tr}}-1\}$, and $\nu = 0$. Note that with Ξ_n as defined in (3.27), $n_{\xi_n} \to \infty$ for all $\xi_n \in \Xi_n$, and thus in particular $n_{\xi_n^\star} \to \infty$ as $n \to \infty$.

It should be stressed that (DET) is an assumption on the base prediction procedure $\widetilde{f}$ and not on the ingredient predictors $\widehat{f}^{\xi}$. In general, the risk behavior of $\widetilde{f}$ does not necessarily imply that of $\widehat{f}^{\xi}$ which is an average of M predictors obtained from $\widetilde{f}$. However, the risk of $\widehat{f}^{\xi}$ can be bounded in terms of the risk $\widetilde{f}$ for loss functions $\ell(\cdot,\cdot)$ that are convex in the second argument. Observe that

$$R(\widehat{f}^{\xi}) \ = \ R\left(\frac{1}{M}\sum_{j=1}^{M}\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})\right) \ \le \ \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})). \tag{3.31}$$

The inequality (3.31) follows from Jensen's inequality. It becomes an equality if $M = 1$ without the requirement that the loss function is convex.

Inequality (3.31) along with the non-stochastic risk approximation (DET) can be used to control $\min_{\xi\in\Xi_n} R(\widehat{f}^{\xi})$ in (3.29). From (3.30), we obtain

$$\begin{aligned}\min_{\xi\in\Xi_n} R(\widehat{f}^{\xi}) \overset{(a)}{\le} \min_{\xi\in\Xi_n} \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})) &\overset{(b)}{\le} \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi_n^\star,j}))\\ &= R^{\mathrm{det}}(n_{\xi_n^\star};\widetilde{f})\left(1+\frac{1}{M}\sum_{j=1}^{M}\frac{R(\widetilde{f}(\cdot,\mathcal{D}_{\mathrm{tr}}^{\xi_n^\star,j})) - R^{\mathrm{det}}(n_{\xi_n^\star};\widetilde{f})}{R^{\mathrm{det}}(n_{\xi_n^\star};\widetilde{f})}\right)\\ &\overset{(c)}{=} \min_{\xi\in\Xi_n} R^{\mathrm{det}}(n_\xi;\widetilde{f})(1+o_p(1))\\ &= R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})(1+o_p(1)).\end{aligned} \tag{3.32}$$

Inequality (a) in (3.32) follows from using Jensen's inequality. Inequality (b) follows because $\xi_n^\star \in \Xi_n$. Equality (c) follows for any fixed $M \ge 1$ from the non-stochastic risk approximation (DET); this can be seen from the fact that the sum of a finite number of $o_p(1)$ random variables is $o_p(1)$.

All the inequalities in (3.32) can be made equalities for $M = 1$, if instead of (DET) we make the stronger assumption that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\xi_n\in\Xi_n} \frac{|R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi_n,j})) - R^{\mathrm{det}}(n_{\xi_n};\widetilde{f})|}{R^{\mathrm{det}}(n_{\xi_n};\widetilde{f})} > \epsilon\right) = 0 \quad \text{for all } \epsilon > 0. \tag{DET*}$$

This is clearly a stronger assumption than required for (3.32), where we only required such relative convergence for a specific $\xi_n^\star \in \Xi_n$. Under (DET*), we can write

$$\begin{aligned}\min_{\xi\in\Xi_n} \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})) &= \min_{\xi\in\Xi_n} R^{\mathrm{det}}(n_\xi;\widetilde{f})\left(1+\frac{1}{M}\sum_{j=1}^{M}\frac{R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})) - R^{\mathrm{det}}(n_\xi;\widetilde{f})}{R^{\mathrm{det}}(n_\xi;\widetilde{f})}\right)\\ &\lessgtr R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})\left(1 \pm \frac{1}{M}\sum_{j=1}^{M}\sup_{\xi\in\Xi_n}\left|\frac{R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,j})) - R^{\mathrm{det}}(n_\xi;\widetilde{f})}{R^{\mathrm{det}}(n_\xi;\widetilde{f})}\right|\right)\\ &= R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})(1+o_p(1)).\end{aligned}$$

We now conclude that for $M = 1$,

$$\min_{\xi\in\Xi_n} R(\widehat{f}^{\xi}) = \min_{\xi\in\Xi_n} R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi,1})) = R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})(1+o_p(1)). \tag{3.33}$$

This proves that all the inequalities in (3.32) can be made equalities for $M = 1$ under the stronger assumption (DET*). Combined with (3.29), this implies that

$$\begin{aligned}R(\widehat{f}^{\mathrm{zs}}) &= \begin{cases} R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})(1+o_p(1)) + O_p(1)\sqrt{\log n/n_{\mathrm{te}}} & \text{if } \widehat{\sigma}_{\Xi} = O_p(1)\\ R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})(1+o_p(1)) & \text{if } \widehat{\kappa}_{\Xi} = O_p(1)\end{cases}\\ &= R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})\begin{cases} 1+o_p(1) + \sqrt{\log n/n_{\mathrm{te}}}/R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f}) & \text{if } \widehat{\sigma}_{\Xi} = O_p(1)\\ 1+o_p(1) & \text{if } \widehat{\kappa}_{\Xi} = O_p(1).\end{cases}\end{aligned} \tag{3.34}$$

As mentioned before, assumption (DET*) is significantly stronger than (DET). In the absence of (DET*), inequality (3.32) combined with (3.29) implies that (3.34) holds with inequalities instead of equalities. For simplicity, denote:

(O1) $\widehat{\sigma}_{\Xi} = O_p(1)$ and $R^{\mathrm{det}}_{\nearrow}(n;\widetilde{f})\sqrt{n_{\mathrm{te}}/\log n} \to \infty$.

(O2) $\widehat{\kappa}_\Xi = O_p(1)$.

Hence, we have proved the following result:

Theorem 3.3.4 (Monotonization by zero-step procedure)**.** *For $M = 1$, if assumption (DET*) and either (O1) or (O2) hold true, then $R^{\mathrm{det}}_{\nearrow}(\cdot; \widetilde{f})$ is a deterministic approximation of the prediction procedure $\widehat{f}^{\mathrm{zs}}$, i.e.,*

$$\frac{|R(\widehat{f}^{\mathrm{zs}}) - R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})|}{R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})} = o_p(1).$$

For $M \geq 1$, if $\ell(\cdot, \cdot)$ is convex in the second argument, assumption (DET), and either (O1) or (O2) hold true, then

$$\frac{(R(\widehat{f}^{\mathrm{zs}}) - R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f}))_+}{R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})} = o_p(1).$$

Remark 3.3.5 (Choice of Ξ_n)**.** All the calculations presented in this section hold for any set Ξ_n with $|\Xi_n| \leq n$. As long as either (DET) (for $\xi_n = \xi_n^\star$ in (3.30)) or (DET*) holds true, then one can use $\Xi_n = \{1, 2, \dots, n_{\mathrm{tr}} - 1\}$ and $\nu = 0$. For this choice, $R^{\mathrm{det}}_{\nearrow}(\cdot; \widehat{f})$ is the monotonized risk as illustrated in Figure 3.2. With the choice of Ξ_n mentioned in (3.27), $R^{\mathrm{det}}_{\nearrow}(\cdot; \widehat{f})$ is not a complete monotonization but it serves as an approximate monotone risk.

Remark 3.3.6 (Exact risk $\widehat{f}^{\mathrm{zs}}$)**.** For $M = 1$ (under (DET*)), Theorem 3.3.4 essentially implies that the risk of the zero-step procedure closely tracks the monotonized deterministic approximation to the conditional prediction risk of $\widetilde{f}$ trained on $\mathcal{D}_{\mathrm{tr}}$. For $M \geq 1$ (under (DET)), Theorem 3.3.4 does not imply the risk of the zero-step predictor is monotonic or even that that a non-stochastic approximation of the risk exists in the sense of Definition 3.3.2. However, our simulations in limited settings presented in Section 3.3.4 suggest that the risk of the zero-step prediction procedure is monotone even for $M \geq 1$.

Remark 3.3.7 (Verification of assumptions in Theorem 3.3.4)**.** The bound on $\widehat{\sigma}_\Xi$ and $\widehat{\kappa}_\Xi$ in Assumptions (O1) and (O2) can be verified for some common loss functions and predictors as discussed in Section 3.2.3. The verification of assumption (DET) or (DET*) is very much tied to the exact prediction procedure. We verify (DET) in a specific setting in Section 3.3.3.1.

3.3.3.1 Risk behavior of $\widehat{f}^{\mathrm{zs}}$ under proportional asymptotics

In the discussion leading up to Theorem 3.3.4, we have not made a specific reference to the growth or non-growth of the dimension of the features. Technically, Theorem 3.3.4 does allow for the dimension p of the features to change with the sample size n, i.e., one can have $p = p_n$.

Risk monotonization is an interesting phenomenon to study in light of the double (or multiple) descent results in the overparameterized setting where $p_n/n \to \gamma$ as $n \to \infty$. In our previous discussion of non-stochastic approximation of the conditional prediction risk, we did not stress the dependence on the dimension of features. In the following, we consider the implications of Theorem 3.3.4 in the context of overparameterized learning and hence consider the following setting.

Recall that the original dataset $\mathcal{D}_n$ consists of n i.i.d. observations $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$, $1 \leq i \leq n$ from distribution P. In the following as we allow the dimension p of the features to change with the sample size n and assume that $p = p_n$ satisfies

(PA(γ)) $p_n/n \to \gamma \in (0, \infty)$ as $n \to \infty$.

The above asymptotic regime, which is standard in random matrix theory (Bai and Silverstein, 2010), is used in the overparameterized learning literature, where it has been referred to as proportional asymptotics (see e.g., Dobriban and Wager (2018); Hastie et al. (2022); Mei and Montanari (2022); Bartlett et al. (2021)). Note that under assumption (PA(γ)) the underlying distribution P of the observations in $\mathcal{D}_n$ should be indexed by the sample size n. We suppress this dependence for convenience. Under the proportional asymptotics regime for commonly studied prediction procedures, a deterministic approximation to the conditional prediction risk of a subset $\mathcal{D}_m \subseteq \mathcal{D}_n$ depends not on m but on p_n/m, among other properties of the distribution P. For this reason, in any discussion of the deterministic approximation of the conditional prediction risk, we write $R^{\mathrm{det}}(p_n/m; \widetilde{f})$ instead of $R^{\mathrm{det}}(m; \widetilde{f})$. Now the goal of this subsection is to derive the deterministic approximation of the conditional risk of the zero-step predictor under (PA(γ)).

Recall that from the crucial calculation in (3.32) leading to the risk of zero-step predictor, we require

$$\frac{R(\widetilde{f}(\cdot, \mathcal{D}_{\mathrm{tr}}^{\xi_n^\star, j})) - R^{\mathrm{det}}(n_{\xi_n^\star}; \widetilde{f})}{R^{\mathrm{det}}(n_{\xi_n^\star}; \widetilde{f})} = o_p(1), \tag{3.35}$$

with $\xi_n^\star$ defined as in (3.30). Except for (3.35), all the remaining steps in (3.32) hold true even in the overparameterized setting. In the following, we will provide simple sufficient condition for verification of (3.35) under (PA(γ)). As mentioned above, the deterministic risk under (PA(γ)) often depends not only on the sample size alone, but also on the ratio of the number of features to the sample size. Therefore, we find it helpful to rewrite (3.35) as

$$\frac{R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_n^\star, j})) - R^{\mathrm{det}}(p_n/n_{\xi_n^\star}; \widetilde{f})}{R^{\mathrm{det}}(p_n/n_{\xi_n^\star}; \widetilde{f})} = o_p(1), \quad \text{where} \quad \xi_n^\star \in \operatorname*{arg\,min}_{\xi \in \Xi_n} R^{\mathrm{det}}(p_n/n_\xi; \widetilde{f}). \tag{DETPA-0}$$

Note that assumption (PA(γ)) does not imply that $p_n/n_{\xi_n^\star}$ converges to a fixed limit as $n \to \infty$.

Under assumption (DETPA-0), Theorem 3.3.4 readily implies the risk behavior of $\widehat{f}^{\mathrm{zs}}$. However, the possibility that $p_n/n_{\xi_n^\star}$ does not converge to a fixed limit necessitates a closer examination of assumption (DETPA-0). We provide a two-fold reduction of assumption (DETPA-0). Firstly, it suffices to verify that the absolute difference between $R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_n^\star, j}))$ and $R^{\mathrm{det}}(p_n/n_{\xi_n^\star}; \widetilde{f})$ converges to 0 when $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is uniformly bounded away from 0. This is a reasonable assumption in practice because several loss functions under mild conditions on the response have risk lower bounded by the unavoidable error which is strictly positive. For example, assuming the loss ℓ is the squared loss and that $\mathbb{E}[(Y_0 - \mathbb{E}[Y_0 \mid X_0])^2] > 0$, we have for any prediction procedure $\widetilde{f}$ and any training dataset $\mathcal{D}_m$ containing m observation,

$$R(\widetilde{f}(\cdot; \mathcal{D}_m)) \;=\; \mathbb{E}[(Y_0 - \widetilde{f}(X_0; \mathcal{D}_m))^2 \big| \mathcal{D}_m] \;\geq\; \mathbb{E}[(Y_0 - \mathbb{E}[Y_0|X_0])^2] > 0. \tag{3.36}$$

Hence, in this case, if there exists a deterministic function $R^{\mathrm{det}} : (0, \infty] \to [0, \infty]$ such that under (PA(γ)), as $n \to \infty$,

$$R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_n^\star, j})) - R^{\mathrm{det}}(p_n/n_{\xi_n^\star}; \widetilde{f}) = o_p(1), \quad \text{where} \quad \xi_n^\star \in \operatorname*{arg\,min}_{\xi \in \Xi_n} R^{\mathrm{det}}(p_n/n_\xi; \widetilde{f}), \tag{3.37}$$

then (DETPA-0) is satisfied. Secondly, the following lemma shows that under (PA(γ)), (3.37) is satisfied if there exists a deterministic approximation for the conditional risk with datasets having a converging aspect ratio (i.e., datasets for which the ratio of the number of features to the sample size converges to a constant).

For any $\gamma > 0$, define

$$\mathcal{M}_\gamma^{\mathrm{zs}} \;:=\; \operatorname*{arg\,min}_{\zeta : \zeta \geq \gamma} R^{\mathrm{det}}(\zeta; \widetilde{f}).$$

Lemma 3.3.8 (Reduction of (DETPA-0))**.** *Let $\mathcal{D}_{k_m}$ be a dataset with k_m observations and p_m features. Consider a prediction procedure $\widetilde{f}$ trained on $\mathcal{D}_{k_m}$. Assume the loss function ℓ is such that $R(\widetilde{f}(\cdot; \mathcal{D}_{k_m}))$ is uniformly bounded from below by 0. Let $\gamma > 0$ be a real number. Suppose there exists a proper, lower semicontinuous function $R^{\mathrm{det}}(\cdot; \widetilde{f}) : [\gamma, \infty] \to [0, \infty]$ such that*

$$R(\widetilde{f}(\cdot; \mathcal{D}_{k_m})) \;\xrightarrow{\mathrm{p}}\; R^{\mathrm{det}}(\phi; \widetilde{f}), \tag{DETPAR-0}$$

as $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in \mathcal{M}_\gamma^{\mathrm{zs}}$. Further suppose that $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is continuous on the set $\mathcal{M}_\gamma^{\mathrm{zs}}$. Then, (DETPA-0) *is satisfied.*

We prove Lemma 3.3.8 using the real analysis fact that a sequence $\{a_n\}_{n\geq 1}$ converges to 0 if and only if for any subsequence $\{a_{n_k}\}_{k\geq 1}$, there exists a further subsequence $\{a_{n_{k_l}}\}_{l\geq 1}$ that converges to 0 (see, for example, Problem 12 of Royden (1988); also see Lemma C.6.3 for a self-contained proof). We apply this fact to the sequence

$$a_n(\epsilon) = \mathbb{P}\left(\left|R(\widetilde{f}(\cdot;\mathcal{D}_{\mathrm{tr}}^{\xi_n^\star,j})) - R^{\det}(p_n/n_{\xi_n^\star};\widetilde{f})\right| \geq \epsilon\right),$$

for every $\epsilon > 0$. A crucial component in applying this technique is to first produce a subsequence $\{n_{k_l}\}_{l\geq 1}$ such that $p_{n_{k_l}}/n_{\xi^\star_{n_{k_l}}}$ converges to a point in $\arg\min_{\zeta\in[\gamma,\infty]} R^{\det}(\zeta;\widetilde{f})$. A few remarks on the assumptions of Lemma 3.3.8 are in order.

- In most cases, the set of minimizers of $R^{\det}(\cdot;\widetilde{f})$ is a singleton set. For such a scenario, Lemma 3.3.8 only requires the deterministic approximation of the conditional prediction risk for a single limiting aspect ratio (i.e., (DETPAR-0) is only required for a single ϕ). Several commonly studied predictors satisfy (DETPAR-0) as discussed below.

- Assuming lower semicontinuity of $R^{\det}(\cdot;\widetilde{f})$ is a mild assumption. In particular, it does not preclude the possibility that $R^{\det}$ diverges to ∞ at several values in the domain as shown in Proposition 3.3.9. Such risk diverging behavior is a common occurrence for several popular predictors in overparameterized learning, for example, MN2LS, MN1LS, etc. The requirement of the lower semicontinuity stems from the goal of monotonizing $R^{\det}$ from *below*.

Proposition 3.3.9 (Verifying lower semicontinuity for diverging risk profiles)**.** *Suppose $h : [a,c] \to \mathbb{R}$ is continuous on $[a,b)\cup(b,c]$ and $\lim_{x\to b^-} h(x) = \lim_{x\to b^+} h(x) = \infty$. Then, h is lower semicontinuous on $[a,c]$.*

Proposition 3.3.9 implies that if $R^{\det}$ is continuous on a set except for a point where it diverges to ∞, then $R^{\det}$ is lower semicontinuous on that set. In this sense, Proposition 3.3.9 relates the lower semicontinuity assumption of Lemma 3.3.8 to the continuity assumption of the lemma.

- Continuity assumption on $R^{\det}(\cdot;\widetilde{f})$ at the argmin set $\arg\min_{\zeta\in[\gamma,\infty]} R^{\det}(\zeta;\widetilde{f})$ is also mild. Proposition 3.3.10 below shows that (DETPAR-0) holding for ϕ in any open set $\mathcal{I}$ implies continuity of $R^{\det}$ on $\mathcal{I}$. In particular, this implies continuity on the sets of the type $\mathcal{I} = (a,\infty]$. If the set of minimizers of $R^{\det}$ is a singleton set, then (DETPAR-0) itself does not suffice to guarantee the continuity of $R^{\det}$ at the minimizer. Proposition 3.3.10 in such a case requires verifying (DETPAR-0) on an open interval containing the minimizer.

Proposition 3.3.10 (Certifying continuity from continuous convergence)**.** *Let $\mathcal{D}_{k_m}$ be a dataset with k_m observations and p_m features, and consider a prediction procedure $\widetilde{f}$ trained on $\mathcal{D}_{k_m}$. Let $\mathcal{I}$ be an open set in $(0,\infty)$. Suppose there exists a function $R^{\det} : (0,\infty] \to [0,\infty]$ such that*

$$R(\widetilde{f}(\cdot;\mathcal{D}_{k_m})) \xrightarrow{\mathrm{p}} R^{\det}(\phi;\widetilde{f}) \tag{3.38}$$

as $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in \mathcal{I}$. Then, $R^{\det}(\cdot;\widetilde{f})$ is continuous on $\mathcal{I}$.

Combining the results and the discussion above, the verification of (DETPA-0) under (PA(γ)) can proceed with the following two-step program.

(PRG-0-C1) For ϕ such that $R^{\det}(\phi;\widetilde{f}) < \infty$, verify that for all datasets $\mathcal{D}_{k_m}$ with limiting aspect ratio ϕ, $R(\widetilde{f}(\cdot;\mathcal{D}_{k_m})) \xrightarrow{\mathrm{p}} R^{\det}(\phi;\widetilde{f})$.

(PRG-0-C2) Whenever $R^{\det}(\phi;\widetilde{f}) = \infty$,

$$\lim_{\phi'\to\phi^-} R^{\det}(\phi';\widetilde{f}) = \lim_{\phi'\to\phi^+} R^{\det}(\phi';\widetilde{f}) = \infty.$$

The continuity of $R^{\det}$ at points where it is finite follows from (PRG-0-C1) via Proposition 3.3.10. This kind of convergence is verified in the literature for several commonly used prediction procedures, such as ridge regression and MN2LS (Hastie et al., 2022), lasso and MN1LS (Li and Wei, 2021), etc; see Remark 3.3.16 for more details. This combined with (PRG-0-C2) via Proposition 3.3.9 implies lower semicontinuity of $R^{\det}$ on $[\gamma, \infty]$. If there is more than one ϕ at which $R^{\det}$ is ∞, then Proposition 3.3.9 should be applied separately by splitting the domain to only contain one point of divergence. A more general result of this flavour can be found in Proposition 3.4.2 in Section 3.4.3.1.

We will follow these steps to verify (DETPA-0) for the ridge and lasso prediction procedures in Section 3.3.3.2. But first we will complete the derivation of the deterministic approximation to the conditional risk of $\widehat{f}^{\mathrm{zs}}$ under (DETPA-0) following (3.32). Lemma 3.3.8 combined with Theorem 3.3.4 proves that the zero-step prediction procedure approximately monotonizes the risk of the base prediction procedure $\widetilde{f}$ as shown in the following result:

Theorem 3.3.11 (Asymptotic risk profile of zero-step predictor)**.** *For any prediction procedure $\widetilde{f}$, suppose (PA(γ)), either (O1) or (O2), and the assumptions of Lemma 3.3.8 hold true. In addition, if the loss function is convex in the second argument, then for any $M \ge 1$,*

$$\left(R(\widehat{f}^{\mathrm{zs}}; \mathcal{D}_n) - \min_{\zeta \ge \gamma} R^{\det}(\zeta; \widetilde{f})\right)_+ = o_p(1).$$

Remark 3.3.12 (Monotonicity in the limiting aspect ratio and improvement over base procedure)**.** If we replace assumption (DETPA-0) with the stronger version

$$\sup_{\xi \in \Xi_n} \frac{|R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi,j})) - R^{\det}(p_n/n_\xi; \widetilde{f})|}{R^{\det}(p_n/n_\xi; \widetilde{f})} = o_p(1), \tag{DETPA-0*}$$

as $n \to \infty$, then for $M = 1$, the conclusion of Theorem 3.3.11 can be strengthened to

$$\left|R(\widehat{f}^{\mathrm{zs}}; \mathcal{D}_n) - \min_{\zeta \ge \gamma} R^{\det}(\zeta; \widetilde{f})\right| = o_p(1). \tag{3.39}$$

This implies that the risk of the zero-step procedure is monotonically non-decreasing in γ. Under the assumptions of Theorem 3.3.11, one can only conclude that the risk of zero-step procedure is asymptotically bounded above by a monotonically non-decreasing function in γ in general. It is trivially true that $\min_{\zeta \le \gamma} R^{\det}(\zeta; \widetilde{f}) \le R^{\det}(\gamma; \widetilde{f})$. Hence, the asymptotic risk of zero-step procedure is no worse than that of the base procedure.

Remark 3.3.13 (Finiteness of the risk of $\widehat{f}^{\mathrm{zs}}$)**.** Predictors such the MN2LS or MN1LS undergo divergence in the prediction risk. The zero-step prediction procedure does not have such a divergence in the risk under general regularity conditions. In particular, as long as $\mathbb{E}[\ell(y, 0)] < \infty$, then the risk of $\widehat{f}^{\mathrm{zs}}$ is asymptotically bounded by $\mathbb{E}[\ell(y, 0)]$. Observe that $\mathbb{E}[\ell(y, 0)]$ is the risk of the null predictor which always returns 0 as its prediction. By including the zero predictor in Algorithm 1, the risk of $\widehat{f}^{\mathrm{zs}}$ will always be asymptotically bounded by this null risk.

3.3.3.2 Verifying deterministic profile assumption (DETPAR-0)

In the following, we will restrict ourselves to the case of linear predictors and squared error loss, and verify assumption (DETPAR-0) for MN2LS and MN1LS base procedures.

Suppose $\mathcal{D}_{k_m} = \{(X_i, Y_i) \in \mathbb{R}^{p_m} \times \mathbb{R} : 1 \le i \le k_m\}$. Recall the MN2LS and MN1LS predictor procedures defined in Examples 3.2.25 and 3.2.26. It is now well-known that the MN2LS and MN1LS prediction procedures has a non-monotone risk as a function of sample size n (Nakkiran et al., 2021; Hastie et al., 2022; Li and Wei, 2021). The following two results verify assumption (DETPAR-0) for these two procedures under some regularity conditions stated in Hastie et al. (2022); Li and Wei (2021).

Proposition 3.3.14 (Verification of (DETPAR-0) for MN2LS procedure)**.** *Assume the setting of Theorem 3 of Hastie et al. (2022). Then, there exists a function $R^{\det}(\cdot; \widetilde{f}_{\mathrm{mn2}}) : (0, \infty] \to [0, \infty]$ such that (PRG-0-C1) holds for all $\phi \neq 1$ and (PRG-0-C2) holds for $\phi = 1$.*

Proposition 3.3.15 (Verification of (DETPAR-0) for MN1LS procedure)**.** *Assume the setting of Theorem 2 of Li and Wei (2021). Then, there exists a function* $R^{\mathrm{det}}(\cdot; \widetilde{f}_{\mathrm{mn1}}) : (0, \infty] \to [0, \infty]$ *such that* (PRG-0-C1) *holds for all* $\phi \neq 1$ *and* (PRG-0-C2) *holds for* $\phi = 1$.

Remark 3.3.16 (Extending Propositions 3.3.14 and 3.3.15 to other predictors)**.** Theorem 3 of Hastie et al. (2022) only provides the asymptotic behavior of the prediction risk computed conditional only on $\{X_i, 1 \leq i \leq k_m\}$. The proof in Appendix C.3 of Proposition 3.3.14 extends the calculations of of Hastie et al. (2022) for prediction risk conditional on $\mathcal{D}_{k_m}$. These calculations can be further extended in a straightforward manner to cover the case of $\lambda > 0$, i.e., the ridge regression procedure. See Proposition 3.3.14 for more details. Similar comments apply to Proposition 3.3.15 where the proposition can be easily extended to cover the case of $\lambda > 0$, i.e., the lasso prediction procedure.

Additionally, most results in the literature under (PA(γ)) derive the risk behavior as $p_m/k_m \to \phi < \infty$. Propositions 3.3.14 and 3.3.15 also extend the existing results to the case when $p_m/k_m \to \infty$ as $m \to \infty$.

We present Propositions 3.3.14 and 3.3.15 as example results to show the verification of our assumptions follow rather easily from the existing asymptotic profile results in the literature. In the proportional asymptotic regime, the risk profiles have been characterized for various other prediction procedures including, high dimensional robust M-estimator (Karoui, 2013, 2018; Donoho and Montanari, 2016), the Lasso estimator (Miolane and Montanari, 2021; Celentano et al., 2020), and various classification procedures (Montanari et al., 2019a; Liang and Sur, 2020a; Sur et al., 2019). Our results can be suitably extended to verify (DETPA-0) for these other predictors. Note that for our results, we only need to know that the asymptotic risk exists, which can potentially hold true under weaker assumptions.

3.3.4 Numerical illustrations

In this section, we provide numerical illustration of the risk monotonization of zero-step prediction procedure in the overparameterized setting, when the base prediction procedures are minimum ℓ_2-norm least squares (MN2LS) and minimum ℓ_1-norm least squares (MN1LS). In order to illustrate risk monotonization as in Theorem 3.3.11, we need to show the risk behavior of $\widehat{f}^{\mathrm{zs}}$ at different aspect ratios. We use the following simulation setups for the two predictors.

Minimum ℓ_2-norm least squares (MN2LS). We fix $n = 1000$ and vary the dimension p of the features from 100 to 10000 (for a total of 20 values of $\gamma = p/n$ logarithmically spaced between 0.1 to 10). This will show the risk behavior of zero-step procedure for aspect ratios between 0.1 to 10. For every pair of sample size $n = 1000$ and dimension p, we generate 100 independent datasets each with n i.i.d. observations from the linear model $Y_i = X_i^\top \beta_0 + \varepsilon_i$, where $X_i \sim \mathcal{N}(0_p, I_p)$, $\beta_0 \sim \mathcal{N}(0_p, \rho^2/pI_p)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ drawn independently of X_i. The model represents a dense signal regime with average signal energy ρ^2. We define the signal-to-noise ratio (SNR) to be ρ^2/σ^2. On each dataset, we apply the MN2LS baseline procedure as well as the zero-step procedure.

In each run, we additionally generate independent test datasets each with 10000 i.i.d. observations from the same $p + 1$ dimensional distribution described above in order to approximate the true risk of the zero-step and the base prediction procedure. Figure 3.3 shows the risks of the baseline MN2LS procedure and the zero-step prediction procedure for high (left, SNR = 4) and low (right, SNR = 1) SNR regimes; we take $\sigma^2 = 1$ and $\rho^2 = \mathrm{SNR}$. We also present the null risk ($\rho^2 + \sigma^2$), i.e., the risk of the zero predictor as a baseline in both the plots. We observe from the figure that the risk of the zero-step procedure for every $M \geq 1$ is non-decreasing in γ. Theorem 3.3.11 implies that the risk of the zero-step prediction procedure for every $M \geq 1$ is *asymptotically* bounded by the risk of the base prediction procedure at each aspect ratio (γ). Although this is somewhat evident from Figure 3.3, it is not satisfied for all γ, especially for $M = 1$. This primarily stems from the smaller sample size at hand and the fact that we are comparing MN2LS trained on full data ($n = 1000$) to the zero-step predictor computed on the train data ($n_{\mathrm{tr}} = 900$). With an increased sample size (to say, $n = 2500$), this finite-sample discrepancy vanishes.

Figure 3.3 shows that the zero-step procedure with $M = 1$ attains risk monotonization in a precise sense that its risk is the largest non-increasing function (of γ) below the risk of the MN2LS predictor. For $M > 1$, our results do not characterize the risk of zero-step predictor, but Figure 3.3 shows that averaging

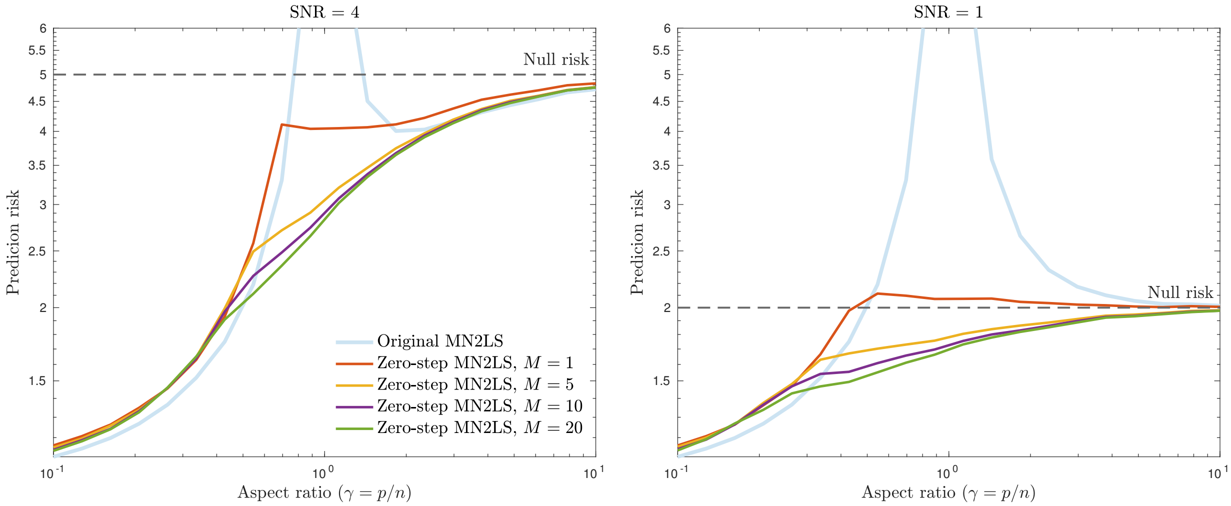


Figure 3.3: Illustration of the zero-step prediction procedure with MN2LS as the base predictor with varying M. The left panel shows a high SNR regime (SNR = 4), while the right panel shows a low SNR regime (SNR = 1). Here, $n = 1000$, $n_{\text{tr}} = 900$, $n_{\text{te}} = 100$, $n^{\nu} = 50$. The features are drawn from an isotropic Gaussian distribution, the response follows a linear model. The figure show averaged risk over 100 dataset repetitions.

has a significant effect in further reducing the risk. As mentioned before, this is expected from the theory of U-statistics as U-statistics are UMVUE's of their expectations (see, e.g., Chapter 5 of Serfling (2009)). All these comments hold for both low and high SNR alike.

Note that the base predictor has unbounded risk near $\gamma = 1$. The risk of the zero-step procedure, on the other hand, is always bounded for all $M \geq 1$ and all γ. In this sense, the zero-step procedure can also be used as a general procedure for mitigating the surprising descent behavior in the prediction risk.

Minimum ℓ_1-norm least squares (MN1LS). We fix $n = 500$ and vary the dimension p of the features from 50 to 50000 (for a total of 30 values of $\gamma = p/n$ logarithmically spaced between 0.1 to 100). This will show risk behavior of zero-step procedure for aspect ratios between 0.1 and 100. For every pair of sample size $n = 500$ and dimension p, we generate 250 independent dataset each with n i.i.d. observations from the linear model $Y_i = X_i^\top \beta_0 + \varepsilon_i$, where $X_i \in \mathcal{N}(0_p, I_p)$, β_0 has coordinates generated i.i.d. from the distribution $B\delta_{r/\sqrt{p\pi}} + (1 - B)\delta_0$, where $B \sim \text{Bernoulli}(\pi = 0.005)$ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent of X_i. The model represents a sparse signal regime (with linear sparsity level π) with average signal energy ρ^2. We again define SNR to be ρ^2/σ^2. On each dataset, we apply the MN1LS baseline procedure as well as the zero-step procedure.

In each run, we additionally generate independent test datasets each with 10000 i.i.d. observations from the same $p + 1$ dimensional distribution described above in order to approximate the true risk of the zero-step and the base prediction procedure. Figure 3.4 shows the risks of the baseline MN1LS procedure and the zero-step procedure for high (left, SNR = 4) and low (right, SNR = 1) SNR regimes. We take $\sigma^2 = 1$ and ρ^2=SNR. We also present the null risk ($\rho^2 + \sigma^2$), i.e., the risk of the zero predictor as a baseline in both the plots. We again observe that the risk of the zero-step procedure for every $M \geq 1$ is non-decreasing in γ.

Similar to Figure 3.3, we observe in Figure 3.4 that the zero-step procedure with $M = 1$ attains precise risk monotonization while zero-step with $M > 1$ improves significantly upon the $M = 1$ when γ is near one. All these comments hold for both low and high SNR alike.

As with Figure 3.3, note that the base predictor MN2LS has unbounded risk near $\gamma = 1$ in Figure 3.4. The risk of the zero-step procedure, on the other hand, is always bounded for all $M \geq 1$ and all γ.

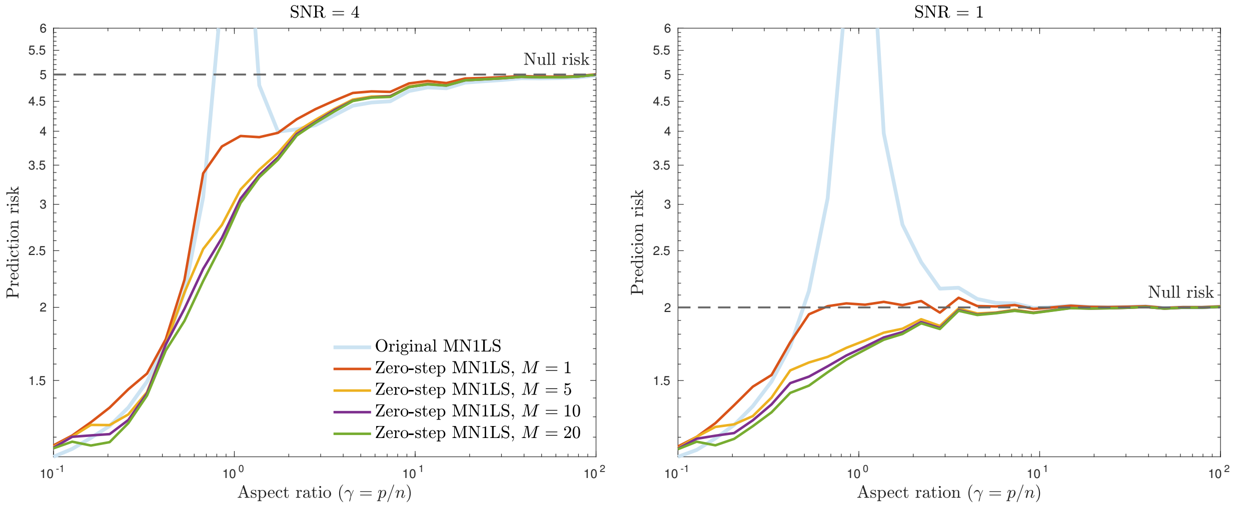


Figure 3.4: Illustration of the zero-step prediction procedure with MN1LS as the base predictor with varying M. The left panel shows a high SNR regime (SNR = 4), while the right panel shows a low SNR regime (SNR = 1). Here, $n = 500$, $n_{\mathrm{tr}} = 420$, $n_{\mathrm{te}} = 80$, $n^{\nu} = 42$. The features are drawn from an isotropic Gaussian distribution, the response follows a linear model with sparse signal (sparsity level = 0.005). The risks are averaged over 250 dataset repetitions.

3.4 Application 2: One-step prediction procedure

3.4.1 Motivation

The zero-step procedure introduced in Section 3.3 provides the desired asymptotic monotonization of the conditional prediction risk under certain regularity conditions. It takes advantage of the fact that we can train our predictors on a smaller subset of the data when it is appropriate. In addition, it uses repeated sampling and averaging in order to remove the external randomness in the choice of the subset.

In this section, we introduce a variant of the zero-step procedure motivated by the classical statistical idea of one-step estimation (see, e.g., Section 5.7 of Van der Vaart, 2000). In the simplest case of linear regression where the feature dimension is fixed, the idea of one-step estimation is that we can start with an arbitrary linear predictor and add to it an adjustment computed based on the residuals of the initial linear predictor. More precisely, starting with any initial estimator $\widetilde{\beta}^{\mathrm{init}}$ and the associated linear predictor $\widetilde{f}(x) = x^{\top}\widetilde{\beta}^{\mathrm{init}}$, we have

$$\underbrace{X^{\top}\widetilde{\beta}^{\mathrm{init}}}_{\text{initial predictor}} + \underbrace{X^{\top}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i^{\top}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i (Y_i - X_i^{\top}\widetilde{\beta}^{\mathrm{init}})\right)}_{\text{one-step adjustment}} = X^{\top}\widetilde{\beta}^{\mathrm{ols}}, \tag{3.40}$$

where the final resulting predictor corresponds to the ordinary least squares (OLS) estimator $\widetilde{\beta}^{\mathrm{ols}}$ that enjoys $n^{-1/2}$ rate and risk optimality under a well-specified linear model.

This idea of one-step estimation is not specific to ordinary least squares. It can be generalized to other estimators that are solutions to estimating equation $\Psi_n(\beta) = 0$ where $\Psi_n : \mathbb{R}^p \to \mathbb{R}^p$. The general idea is to solve a linear approximation to the estimating equation, i.e., given an initial estimator $\widetilde{\beta}^{\mathrm{init}}$, the one-step estimator is the solution (in β) to the linearized estimating equation (around $\widetilde{\beta}^{\mathrm{init}}$)

$$\Psi_n(\widetilde{\beta}^{\mathrm{init}}) + \nabla\Psi_n(\widetilde{\beta}^{\mathrm{init}})(\beta - \widetilde{\beta}^{\mathrm{init}}) = 0.$$

The solution can be expressed as

$$\widetilde{\beta} = \underbrace{\widetilde{\beta}^{\text{init}}}_{\text{initial estimator}} - \underbrace{(\nabla\Psi(\widetilde{\beta}^{\text{init}}))^{-1}\Psi(\widetilde{\beta}^{\text{init}})}_{\text{one-step adjustment}}. \tag{3.41}$$

Here $\nabla\Psi : \mathbb{R}^p \to \mathbb{R}^p \times \mathbb{R}^p$ denotes the Jacobian of Ψ.

One can also view the one-step estimator from the point of view of the Newton's algorithm. The classical one-step estimator starts at an initial estimator $\widetilde{\beta}^{\text{init}}$ and takes a Newton's step on the empirical risk minimization problem. For a parametric predictor $f(\cdot;\widetilde{\beta}^{\text{init}})$, starting with a base estimator $\widetilde{\beta}^{\text{init}}$, we can define the corresponding one-step predictor as $f(\cdot;\widetilde{\beta})$, where $\widetilde{\beta}$ is the Newton's step update starting with $\widetilde{\beta}^{\text{init}}$ given by

$$\widetilde{\beta} = \underbrace{\widetilde{\beta}^{\text{init}}}_{\text{initial estimator}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(Y_i, f(X_i;\widetilde{\beta}^{\text{init}}))\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla\ell(Y_i, f(X_i;\widetilde{\beta}^{\text{init}}))\right)}_{\text{Newton's step}}. \tag{3.42}$$

Here, for $1 \le i \le n$, $\nabla\ell(Y_i, f(X_i;\cdot)) : \mathbb{R}^p \to \mathbb{R}^p$ denotes the gradient of the prediction loss function $\ell(Y_i, f(X_i;\beta))$ with respect to β, and $\nabla^2\ell(Y_i, f(X_i;\cdot)) : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ denotes the Hessian of the prediction loss function with respect to β. In the special case of a linear predictor, where $f(x;\beta) = x^T\beta$, the one-step estimator becomes

$$\widetilde{\beta} = \widetilde{\beta}^{\text{init}} - \left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i^T\ell''(Y_i, X_i^T\widetilde{\beta}^{\text{init}})\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\ell'(Y_i, X_i^T\widetilde{\beta}^{\text{init}})\right),$$

where $\ell' : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the first derivative of the loss function $\ell(\cdot,\cdot)$ in the second coordinate, and $\ell'' : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the second derivative of the loss function in the second coordinate.

Our goal in this section is to build upon this idea of one-step estimation towards risk-monotonization and improve on the zero-step procedure. We will restrict ourselves to one-step adjustment with respect to the square error loss and linear predictors (per (3.40)). We leave extension to a more general one-step adjustment (per (3.41) or (3.42)) for future work. For more discussion, see Section 3.5.

There are two points to note when defining (3.40).

1. The inverse of the sample covariance matrix $\sum_{i=1}^{n} X_iX_i^\top/n$ in (3.40) need not always exist. In particular, when the feature dimension $p > n$, the sample covariance matrix is guaranteed to be rank deficient.

2. In the overparameterized regime, the residuals $Y_i - X_i^\top\widetilde{\beta}^{\text{init}}$ for $i = 1,\dots,n$ in (3.40) are identically zero for several commonly used estimators such MN2LS or MN1LS, if $\widetilde{\beta}^{\text{init}}$ and the residuals are computed on the same dataset.

In order to overcome these two limitations, we consider a variant of the idea of one-step estimation, in which we make the following changes:

1′. We use a Moore-Penrose pseudo-inverse in place of regular matrix inverse. Note that this is the same as adding a MN2LS component fitted on the residuals $Y_i - X_i^\top\widetilde{\beta}^{\text{init}}$.

2′. We split the training data and use one part to compute $\widetilde{\beta}^{\text{init}}$ and use the other part to compute the residuals $Y_i - X_i^\top\widetilde{\beta}^{\text{init}}$. This ensures that the residuals are not identically zero in the overparameterized regime.

In summary, to construct the one-step predictor, we start with a base predictor computed on a subset of data, evaluate the residuals of this predictor on a different subset of data, and add to the base predictor a MN2LS fit on the residuals. We formalize this construction next.

3.4.2 Formal description

As before, let the original dataset be denoted by $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and let $\widetilde{f}$ be a base prediction procedure. As per Algorithm 1, let the train and test datasets be $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$, respectively. We define the ingredient predictors to be used in Algorithm 1 constructed using the one-step methodology as follows: Define the index set Ξ_n as

$$\Xi_n := \Big\{(\xi_1, \xi_2) \ : \ \xi_1 \in \{0, 1, \dots, n_{\mathrm{tr}} - 1\}, \xi_2 \in \{0, 1, \dots, \xi_1 - 1\}\Big\}.$$

Let $\mathcal{D}_{\mathrm{tr}}^{\xi_1}$ and $\mathcal{D}_{\mathrm{tr}}^{\xi_2}$ be disjoint subsets of $\mathcal{D}_{\mathrm{tr}}$ with $n_{\mathrm{tr}} - \xi_1$ (for $0 \le \xi_1 \le n_{\mathrm{tr}} - 1$) and ξ_2 (for $0 \le \xi_2 \le \xi_1$) observations, respectively. Let $\mathcal{I}_{\mathrm{tr}}^{\xi_1}$ and $\mathcal{I}_{\mathrm{tr}}^{\xi_2}$ denote the corresponding index sets of $\mathcal{D}_{\mathrm{tr}}^{\xi_1}$ and $\mathcal{D}_{\mathrm{tr}}^{\xi_2}$, respectively. For each index $\xi = (\xi_1, \xi_2) \in \Xi_n$, define the ingredient predictor $\widetilde{f}^{\xi}$ to be used in Algorithm 1 in three steps:

1. Fit a base prediction procedure $\widetilde{f}$ on $\mathcal{D}_{\mathrm{tr}}^{\xi_1}$. Call this $\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_1})$.

2. Compute the residuals of predictor $\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_1})$ on $\mathcal{D}_{\mathrm{tr}}^{\xi_2}$, i.e., $r_j = Y_j - \widetilde{f}(X_j; \mathcal{D}_{\mathrm{tr}}^{\xi_1})$ for $j \in \mathcal{I}_{\mathrm{tr}}^{\xi_2}$.

3. Fit the MN2LS predictor on $\{(X_j, r_j) : j \in \mathcal{I}_{\mathrm{tr}}^{\xi_2}\}$. This is the one-step adjustment.

The final ingredient predictor $\widetilde{f}^{\xi}$ is given by

$$\widetilde{f}^{\xi}(x; \mathcal{D}_{\mathrm{tr}}^{\xi_1}, \mathcal{D}_{\mathrm{tr}}^{\xi_2}) \ := \ \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^{\xi_1}) + x^{\top} \left(\sum_{j \in \mathcal{I}_{\mathrm{tr}}^{\xi_2}} X_j X_j^{\top} \right)^{\dagger} \left(\sum_{j \in \mathcal{I}_{\mathrm{tr}}^{\xi_2}} X_j r_j \right).$$

If $\xi_2 = 0$, then $\mathcal{I}_{\mathrm{tr}}^{\xi_2}$ is an empty set and there are no residuals r_j computed. In this case, we adopt the convention that there is no one-step adjustment. Therefore, the ingredient predictors for our one-step procedure includes the ingredient predictors for the zero-step procedure. As with the zero-step procedure, two remarks are in order:

- There is external randomness in choosing subsets $\mathcal{D}_{\mathrm{tr}}^{\xi_1}$ and $\mathcal{D}_{\mathrm{tr}}^{\xi_2}$ of sizes $n_{\mathrm{tr}} - \xi_1$ and ξ_2, respectively. To reduce such randomness, we make use of many different subsets of the same sizes and average such different one-step predictors. More precisely, for each $\xi = (\xi_1, \xi_2) \in \Xi$, draw m disjoint pairs of sets $(\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j}), \dots, (\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$ from $\mathcal{D}_{\mathrm{tr}}$. Formally, for $1 \le j \le m$, we randomly draw a subset $\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}$ from $\mathcal{D}_{\mathrm{tr}}$ of size $n_{\mathrm{tr}} - \xi_1$ and a subset $\mathcal{D}_{\mathrm{tr}}^{\xi_2, j}$ from $\mathcal{D}_{\mathrm{tr}} \setminus \mathcal{D}_{\mathrm{tr}}^{\xi_1, j}$ of size ξ_2. We then fit different one-step predictors $\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_i, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$ on $(\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$ for $1 \le j \le M$, and take the final ingredient predictor $\widehat{f}^{\xi}$ to be the average of M such predictors:

$$\widehat{f}^{\xi}(x) = \frac{1}{M} \sum_{j=1}^{M} \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j}). \tag{3.43}$$

  As before, when $M = \infty$, $\widehat{f}^{\xi}$ becomes the average of all possible pairs of disjoints subsets $\mathcal{D}_{\mathrm{tr}}$ of sizes $n_{\mathrm{tr}} - \xi_1$ and ξ_2, while the case of $M = 1$ has the largest amount of external randomness. Based on the theory of U-statistics, we again expect the choice of $M = \infty$ to provide a predictor with the smallest variance. For computational reasons, we use a finite value of $M \ge 1$.

- In the description above, we have $n_{\mathrm{tr}}(n_{\mathrm{tr}} + 1)/2$ predictors to use in Algorithm 1. Similar to the zero-step procedure, we replace Ξ_n with

$$\Xi_n := \left\{(\xi_1, \xi_2) \ : \ \xi_1 \in \left\{2, \dots, \left\lceil \frac{n_{\mathrm{tr}}}{\lfloor n^{\nu} \rfloor} - 2 \right\rceil \right\}, \xi_2 \in \{1, \dots, \xi_1 - 1\} \right\}, \quad \text{for some } \nu \in (0, 1), \tag{3.44}$$

  and consider predictors obtained by training components of $\widetilde{f}$ on subsets of sizes $n_{\mathrm{tr}} - \xi_1 \lfloor n^{\nu} \rfloor$ and $\xi_2 \lfloor n^{\nu} \rfloor$. Such a change helps in reducing the cost of computing $\widehat{f}^{\mathrm{cv}}$ using Algorithm 1. In addition, this also helps in the statistical properties of $\widehat{f}^{\mathrm{cv}}$ when applying the union bound in the results of Section 3.2.

With these two modifications, with Ξ_n as defined in (3.44), for $\xi \in \Xi_n$, we define $\widehat{f}^{\xi}$ as in (3.43) with the subsets $\mathcal{D}_{\mathrm{tr}}^{\xi_1,j}$, $\mathcal{D}_{\mathrm{tr}}^{\xi_2,j}$ (for $1 \le j \le M$) now representing disjoints subsets of sizes $n_{\mathrm{tr}} - \xi_1 \lfloor n^{\nu} \rfloor$ and $\xi_2 \lfloor n^{\nu} \rfloor$, respectively. The ingredients predictors to be used in Algorithm 1 are given by $\widehat{f}^{\xi}$, $\xi \in \Xi_n$. We call the resulting predictor obtained from Algorithm 1 as the one-step predictor based on $\widetilde{f}$, and we denote the corresponding prediction procedure to be $\widehat{f}^{\mathrm{os}}$. The one-step procedure is summarized in Algorithm 3.

```
Algorithm 3 One-step procedure
Inputs:

  – all inputs of Algorithm 1 other than the index set $\Xi$;
  – a positive integer $M$.

Output:

  – a predictor $\widehat{f}^{\mathrm{os}}$

Procedure:

  1. Let $n_{\mathrm{tr}} = n - n_{\mathrm{te}}$. Construct an index set $\Xi_n$ per (3.44).
  2. Construct train and test sets $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$ per Step 1 of Algorithm 1.
  3. Let $n_{1,\xi_1} = n_{\mathrm{tr}} - \xi_1 \lfloor n^{\nu} \rfloor$ and $n_{2,\xi_2} = \xi_2 \lfloor n^{\nu} \rfloor$. For each $(\xi_1, \xi_2) \in \Xi_n$ and $j = 1, \dots, M$, draw random pairs of disjoint subsets $(\mathcal{D}_{\mathrm{tr}}^{\xi_1,j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2,j})$ of sizes $n_{1,\xi_1}$ and $n_{2,\xi_2}$ from $\mathcal{D}_{\mathrm{tr}}$, respectively. For each $(\xi_1, \xi_2) \in \Xi_n$, fit predictors $\widehat{f}^{\xi}$ as described by (3.43) using prediction procedure $\widetilde{f}$ and $\{(\mathcal{D}_{\mathrm{tr}}^{\xi_1,j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2,j}) : 1 \le j \le M\}$.
  4. Run Steps 3–5 of Algorithm 1 using index set $\Xi = \Xi_n$ and set of predictors $\{\widehat{f}^{\xi}, \xi \in \Xi\}$.
  5. Return $\widehat{f}^{\mathrm{os}}$ as the resulting $\widehat{f}^{\mathrm{cv}}$ from Algorithm 1.
```

3.4.3 Risk behavior of $\widehat{f}^{\mathrm{os}}$

In this section, we examine the risk behavior of one-step predictor $\widehat{f}^{\mathrm{os}}$. Similar treatment as done for the zero-step procedure in Section 3.3.3 applies in general. To avoid repetition, we will primarily restrict ourselves to the proportional asymptotics regime in this section.

3.4.3.1 Risk behavior of $\widehat{f}^{\mathrm{os}}$ under proportional asymptotics

Define $n_{1,\xi_1} = n_{\mathrm{tr}} - \xi_1 \lfloor n^{\nu} \rfloor$ and $n_{2,\xi_2} = \xi_2 \lfloor n^{\nu} \rfloor$. Assume that there exists a deterministic profile $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f}) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of $\widetilde{f}$ such that the following holds:

$$\left| R\big(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi^\star_{1,n},j}, \mathcal{D}_{\mathrm{tr}}^{\xi^\star_{2,n},j})\big) - R^{\mathrm{det}}\left(\frac{p}{n_{1,\xi^\star_{1,n}}}, \frac{p}{n_{2,\xi^\star_{2,n}}}; \widetilde{f}\right) \right| = o_p(1) R^{\mathrm{det}}\left(\frac{p}{n_{1,\xi^\star_{1,n}}}, \frac{p}{n_{2,\xi^\star_{2,n}}}; \widetilde{f}\right), \tag{DETPA-1}$$

where $(\xi^\star_{1,n}, \xi^\star_{2,n})$ are the indices that minimize the deterministic profile $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$:

$$(\xi^\star_{1,n}, \xi^\star_{2,n}) \in \mathop{\arg\min}_{(\xi_1, \xi_2) \in \Xi_n} R^{\mathrm{det}}\left(\frac{p}{n_{1,\xi_1}}, \frac{p}{n_{2,\xi_2}}; \widetilde{f}\right). \tag{3.45}$$

Because $\log(|\Xi_n|) \le 2\log(n)$, following the arguments in Section 3.3.3, we conclude that if (DETPA-1) and either (O1)[5] or (O2) hold, then

$$\left(R(\widehat{f}^{\mathrm{os}}) - \min_{(\xi_1,\xi_2)\in\Xi_n} R^{\mathrm{det}}\left(\frac{p}{n_{1,\xi_1}}, \frac{p}{n_{2,\xi_2}}; \widetilde{f}\right)\right)_+ = o_p(1) \cdot \min_{(\xi_1,\xi_2)\in\Xi_n} R^{\mathrm{det}}\left(\frac{p}{n_{1,\xi_1}}, \frac{p}{n_{2,\xi_2}}; \widetilde{f}\right). \tag{3.46}$$

Just as we reduced verification of (DETPA-0) to (DETPAR-0), we state below a reduction of the verification of (DETPA-1) that only considers non-deterministic sequences for which the aspect ratios of the split datasets for the constituent one-step predictors converge.

For any $\gamma > 0$, define

$$\mathcal{M}_{\gamma}^{\mathrm{os}} := \operatorname*{arg\,min}_{(\zeta_1,\zeta_2):\zeta_1^{-1}+\zeta_2^{-1}\le\gamma^{-1}} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}).$$

Lemma 3.4.1 (Reduction of (DETPA-1))**.** *Suppose $\mathcal{D}_{k_{1,m}}$ and $\mathcal{D}_{k_{2,m}}$ are dataset with $k_{1,m}$ and $k_{2,m}$ observations and p_m features. Assume the loss function ℓ is such that $R(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}}))$ is uniformly bounded away from 0. Let $\gamma > 0$ be a real number. Suppose there exists a proper, lower semicontinuous function $R^{\mathrm{det}} : [\gamma,\infty] \times [\gamma,\infty] \to [0,\infty]$ such that the following holds true:*

$$R\big(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})\big) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi_1,\phi_2;\widetilde{f}) \tag{DETPAR-1}$$

as $k_{1,m}, k_{2,m}, p_m \to \infty$ and $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_1,\phi_2) \in \mathcal{M}_{\gamma}^{\mathrm{os}}$. Furthermore, suppose that $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ is continuous on the set $\mathcal{M}_{\gamma}^{\mathrm{os}}$. Then, (DETPA-1) is satisfied.

The proof of Lemma 3.4.1 follows analogously to that of Lemma 3.3.8 where we show that even though the sequence $\{\mathbf{\Phi}_n = (p_n/n_{1,\xi^\star_{1,n}}, p_n/n_{2,\xi^\star_{2,n}})\}_{n\ge1}$ may not converge, there exists a subsequence $\{\mathbf{\Phi}_{n_{k_l}}\}_{l\ge1}$ that converges to some $(\phi_1,\phi_2) \in \mathcal{M}_{\gamma}^{\mathrm{os}}$. Below we provide some commentary on the assumptions of Lemma 3.4.1.

- We note that assuming lower semicontinuity of $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ is a mild assumption. In particular, it does not preclude the possibility that R^{det} diverges to ∞ at several values in the domain as shown in Proposition 3.4.2. For example, the proposition implies that if $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ is continuous on a set except for when $\phi_1 = 1$ or $\phi_2 = 1$, then R^{det} is lower semicontinuous, provided R^{det} diverges to ∞ when either ϕ_1 or ϕ_2 converges to 1. The condition of lower semicontinuous deterministic approximation $R^{\mathrm{det}}(\cdot;\cdot;\widetilde{f})$ follows from the continuity of the domain of $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ (i.e., points of finite function value). This is similar to Proposition 3.3.9 discussed in the context of the zero-step predictor. The formal statement for the one-step predictor is as follows.

Proposition 3.4.2 (Verifying lower semicontinuity for diverging risk profiles)**.** *Let (M,d) be a metric space. Let C be a closed set. Suppose $h : M \to \overline{\mathbb{R}}$ is a function such that $h(x) < \infty$ for $x \in M \setminus C$, and $h(x) = \infty$ for $x \in C$. In addition, if h restricted to $M \setminus C$ (denoted by $h|_{M\setminus C}(\cdot)$) is continuous, and for any sequence $\{x_n\}_{n\ge1}$ that converges to a point in C, $\{h(x_n)\}_{n\ge1}$ converges to ∞. Then, h is lower semicontinuous on M.*

- Continuity assumption on $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ at the argmin set $\mathcal{M}_{\gamma}^{\mathrm{os}}$ is also mild. Proposition 3.4.3 below shows that (DETPAR-0) holding for (ϕ_1,ϕ_2) in any open set $\mathcal{I}$ implies continuity of R^{det} on $\mathcal{I}$.

Proposition 3.4.3 (Certifying continuity from continuous convergence)**.** *Let $\mathcal{D}_{k_{1,m}}$ and $\mathcal{D}_{k_{2,m}}$ be datasets with $k_{1,m}$ and $k_{2,m}$ observations and p_m features, and consider one-step ingredient prediction procedure $\widetilde{f}$ trained on $\mathcal{D}_{k_{1,m}}$ and $\mathcal{D}_{k_{2,m}}$. Fix a open set $\mathcal{I} \subseteq (0,\infty] \times (0,\infty]$. Suppose there exists a function $R^{\mathrm{det}} : (0,\infty] \times (0,\infty] \to [0,\infty]$ such that*

$$R(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi_1,\phi_2;\widetilde{f}) \tag{3.47}$$

as $k_{1,m}, k_{2,m}, p_m \to \infty$ and $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_1,\phi_2) \in \mathcal{I}$. Then, $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f})$ is continuous on $\mathcal{I}$.

[5] Here, we need (O1) with $R^{\mathrm{det}}_{\nearrow}(n,\widetilde{f})$ replaced with the minimum appearing in (3.46).

Combining the results and the discussion above, the verification of (DETPAR-1) under (PA(γ)) can proceed the following three-point program:

(PRG-1-C1) For (ϕ_1, ϕ_2) such that $R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f}) < \infty$, verify that for all datasets $\mathcal{D}_{k_{1,m}}$ and $\mathcal{D}_{k_{2,m}}$ with limiting aspect ratios (ϕ_1, ϕ_2), $R(\widetilde{f}(\cdot, \cdot; \mathcal{D}_{k_{1,m}, \mathcal{D}_{k_{2,m}}})) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f})$.

(PRG-1-C2) Whenever $R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f}) = \infty$, it obeys that

$$\lim_{(\phi_1', \phi_2') \to (\phi_1, \phi_2)} R^{\mathrm{det}}(\phi_1', \phi_2'; \widetilde{f}) = \infty.$$

(PRG-1-C3) The set of all points where $R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f}) = \infty$ is a closed set.

We will follow these steps to verify (DETPAR-1) for the MN2LS and MN1LS prediction procedures in Section 3.4.3.2. But we will first complete the derivation of the deterministic approximation to the conditional risk of $\widehat{f}^{\mathrm{os}}$ under (DETPAR-1). Following similar arguments as those in Section 3.3.3 for the zero-step procedure, Lemma 3.4.1 along with (3.46) provides the following monotonization result for the one-step procedure:

Theorem 3.4.4 (Asymptotic risk profile of one-step predictor)**.** *For any prediction procedure $\widetilde{f}$ suppose* (PA(γ)), *either* (O1) *or* (O2), *and the assumptions of Lemma 3.4.1 hold true. In addition, if the loss function is convex in the second argument, then for any $M \ge 1$,*

$$\left(R(\widehat{f}^{\mathrm{os}}; \mathcal{D}_n) - \min_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) \right)_+ = o_p(1). \tag{3.48}$$

Theorem 3.4.4 hinges on (DETPA-1) and continuity of $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$ which we will verify below in a specific model setting. Before doing that, let us briefly remark about the extensions and implications of (3.48).

Remark 3.4.5 (Exact risk of $\widehat{f}^{\mathrm{os}}$)**.** For $M = 1$ under (DETPA-1), (3.48) only guarantees that the risk of $\widehat{f}^{\mathrm{os}}$ is bounded above by the minimum in (3.48). Considering a stricter version (DETPA-1*) of (DETPA-1) that requires the $o_p(1)$ in (DETPA-1) to be uniform over all $(\xi_{1,n}, \xi_{2,n}) \in \Xi_n$, conclusion (3.48) can be extended to imply for $M = 1$ that

$$\left| R(\widehat{f}^{\mathrm{os}}; \mathcal{D}_n) - \min_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) \right| = o_p(1). \tag{3.49}$$

This shows that the risk of the one-step procedure with $M = 1$ under the stricter assumption of (DETPA-1*) is exactly the same as the minimum in the display above. This is the characterization of the risk of the one-step procedure in the same vein as (3.39) is the characterization of the risk of the zero-step procedure.

Remark 3.4.6 (Monotonicity in the limiting aspect ratio)**.** Observe that the following map

$$\gamma \mapsto \min_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f})$$

is non-decreasing in γ. This is because

$$\{(\zeta_1, \zeta_2) : 1/\zeta_1 + 1/\zeta_2 \le 1/\gamma_u\} \subseteq \{(\zeta_1, \zeta_2) : 1/\zeta_1 + 1/\zeta_2 \le 1/\gamma_l\} \quad \text{for } \gamma_l \le \gamma_u,$$

and hence the minimum can only be larger as γ increases. This implies that the risk of the one-step procedure in asymptotically bounded above by a monotonically non-decreasing function in γ under the assumptions of Theorem 3.4.4.

Remark 3.4.7 (Comparison with $\widehat{f}^{\mathrm{zs}}$)**.** Observe that

$$\min_{1/\zeta_1 + 1/\zeta_2 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) \;\le\; \min_{1/\zeta_1 \le 1/\gamma} R^{\mathrm{det}}(\zeta_1; \widetilde{f}), \tag{3.50}$$

where the left hand side is the asymptotic risk of $\widehat{f}^{\mathrm{os}}$ (with $M = 1$ and under (DETPA-1*)), the right hand side is the asymptotic risk of $\widehat{f}^{\mathrm{zs}}$ (with $M = 1$ under (DETPA-0*)). Hence, under some regularity conditions, the one-step procedure is as good as the zero-step procedure if not better. See Remark 3.4.12 for more details. For $M > 1$ such a comparison is not readily plausible from our results.

3.4.3.2 Verification of (DETPAR-1)

We now verify the assumption (DETPAR-1) in a specific model setting when the base prediction procedure is either MN2LS or MN1LS. But first, we provide a general result describing the asymptotic risk profile of $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}))$ when the base prediction procedure is linear.

Let $\widetilde{f}$ be a linear base prediction procedure given by $\widetilde{f}(x; \mathcal{D}_{k_{1,m}}) = x^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}})$, for some $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) \in \mathbb{R}^p$ computed on $\mathcal{D}_{k_{1,m}}$. If $\mathcal{D}_{k_{2,m}} = \{(X_i, Y_i) : 1 \le i \le k_{2,m}\}$, the ingredient predictor $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ for the one-step prediction procedure is given by

$$\widetilde{f}(x; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) = x^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + x^\top \widetilde{\beta}_{\mathrm{mn2}}(\{(X_i, Y_i - X_i^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}})) : 1 \le i \le k_{2,m}\}). \tag{3.51}$$

The following result characterizes the conditional prediction risk of $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ for the squared error loss in terms of the risk behavior of $\widetilde{\beta}(\mathcal{D}_{k_{1,m}})$. This is possible because the one-step adjustment is fixed to be the MN2LS prediction procedure and its risk behavior can be completely characterized as done in Section 3.3.3.1.

Consider the setting of Proposition 3.3.14. Let $\Sigma = WRW^\top$ denote the eigenvalue decomposition of the covariance matrix $\Sigma = \mathrm{Cov}(X_0)$, where $R \in \mathbb{R}^{p_m \times p_m}$ is a diagonal matrix containing eigenvalues $r_1 \ge r_2 \ge \cdots \ge r_{p_m} \ge 0$, and $W \in \mathbb{R}^{p_m \times p_m}$ is an orthonormal matrix containing the corresponding eigenvectors $w_1, w_2, \ldots, w_{p_m} \in \mathbb{R}^{p_m}$. In preparation for the statement to follow, define the following (random) probability distribution on $\mathbb{R}_{\ge 0}$:

$$\widehat{Q}_n(r) := \frac{1}{R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})) - \sigma^2} \sum_{i=1}^{p_m} ((\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top w_i)^2 r_i \mathbb{I}\{r_i \le r\}. \tag{3.52}$$

Let H_{p_m} denote the empirical spectral distribution of Σ, whose value at any $r \in \mathbb{R}$ is given by

$$H_{p_m}(r) = \frac{1}{p_m} \sum_{i=1}^{p_m} \mathbb{I}_{\{r_i \le r\}}, \tag{3.53}$$

and let H denote the corresponding limiting spectral distribution, i.e., $H_{p_m} \xrightarrow{\mathrm{d}} H$ as $p_m \to \infty$. See (ℓ_2A5) in the proof of Proposition 3.3.14 for more details.

Lemma 3.4.8 (Continuous convergence of squared risk for one-step procedure)**.** *Let $\widetilde{f}$ be any linear prediction procedure, and assume the setting of Proposition 3.3.14. Let $k_{1,m}, k_{2,m}, p_m \to \infty$ such that $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_1, \phi_2)$. Suppose there exists a deterministic approximation $R^{\det}(\phi_1; \widetilde{f})$ to the conditional squared prediction risk of $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})$ such that $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})) \xrightarrow{\mathrm{p}} R^{\det}(\phi_1; \widetilde{f})$ for ϕ_1 that satisfy $R^{\det}(\phi_1; \widetilde{f}) < \infty$. Assume the distribution $\widehat{Q}_n$ as defined in (3.52) converges weakly to a fixed distribution Q, in probability. Then, for $\phi_2 \in (0, 1) \cup (1, \infty]$, we have $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})) \xrightarrow{\mathrm{p}} R^{\det}(\phi_1, \phi_2; \widetilde{f})$, where $R^{\det}(\phi_1, \phi_2; \widetilde{f})$ is given by*

$$R^{\det}(\phi_1, \phi_2; \widetilde{f}) = \begin{cases} R^{\det}(\phi_1; \widetilde{f}) & \text{if } \phi_2 = \infty \\ R^{\det}(\phi_1; \widetilde{f}) \Upsilon_b(\phi_1, \phi_2) + \sigma^2(1 - \Upsilon_b(\phi_1, \phi_2)) + \sigma^2 \widetilde{v}_g(0; \phi_2) & \text{if } \phi_2 \in (1, \infty) \\ \sigma^2 \left(\dfrac{1}{1 - \phi_2} \right) & \text{if } \phi_2 \in (0, 1). \end{cases} \tag{3.54}$$

Here, the scalars $v(0; \phi_2)$, $\widetilde{v}(0; \phi_2)$, $\widetilde{v}_g(0; \phi_2)$, and $\Upsilon_b(\phi_1, \phi_2)$, for $\phi_2 \in (1, \infty)$, are defined as follows:

- *$v(0; \phi_2)$ is the unique solution to the fixed-point equation:*

$$v(0; \phi_2) = \left(\phi_2 \int \frac{r}{v(0; \phi_2) r + 1} \, \mathrm{d}H(r) \right)^{-1}, \tag{3.55}$$

– *$\widetilde{v}(0;\phi_2)$ is defined in terms of $v(0;\phi_2)$ by the equation:*

$$\widetilde{v}(0;\phi_2) = \left(\frac{1}{v(0;\phi_2)^2} - \phi_2 \int \frac{r^2}{(v(0;\phi_2)r+1)^2}\,\mathrm{d}H(r)\right)^{-1}, \tag{3.56}$$

– *$\widetilde{v}_g(0;\phi_2)$ is defined in terms of $v(0;\phi_2)$ and $\widetilde{v}(0;\phi_2)$ by the equation:*

$$\widetilde{v}_g(0;\phi_2) = \widetilde{v}(0;\phi_2)\phi_2 \int \frac{r^2}{(v(0;\phi_2)r+1)^2}\,\mathrm{d}H(r), \tag{3.57}$$

– *$\Upsilon_b(\phi_1,\phi_2)$ is defined in terms of $v(0;\phi_2)$ and $\widetilde{v}_g(0;\phi_2)$ by the equation:*

$$\Upsilon_b(\phi_1,\phi_2) = (1+\widetilde{v}_g(0;\phi_2)) \int \frac{1}{(v(0;\phi_2)r+1)^2}\,\mathrm{d}Q(r). \tag{3.58}$$

Lemma 3.4.8 provides a deterministic risk approximation for the ingredient one-step predictor $\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})$ in terms of the deterministic risk approximation of the base prediction procedure $\widetilde{f}$. In case of isotropic covariates, i.e., $\Sigma = I_{p_m}$, the distribution H is degenerate at 1, and $R^{\det}(\phi_1,\phi_2;\widetilde{f})$ can be simplified because $\Upsilon_b(\phi_1,\phi_2) = (1-1/\phi_2)$, and $\widetilde{v}_g(0;\phi_2) = 1/(\phi_2-1)$. See the proof of Proposition 3.4.11 for more details.

Note that the assumed limiting distribution Q in general depends on ϕ_1, ϕ_2, and hence $\Upsilon_b(\phi_1,\phi_2)$ is in general a function of ϕ_1, ϕ_2, and the distribution of the data. On the other hand, $v(0;\phi_2)$ defined in (3.55), is a function of ϕ_2 alone, and hence $\widetilde{v}_g(0;\phi_2)$ is just a function of ϕ_2. Furthermore, it can be verified that $\widetilde{v}_g(0;\cdot)$ is a continuous function on $(1,\infty)$ and $\lim_{\phi_2\to 1^+}\widetilde{v}_g(0;\phi_2) = \infty$; see Lemma C.6.13 (4). This implies that $R^{\det}(\phi_1,\phi_2;\widetilde{f})$ satisfies (PRG-1-C1)–(PRG-1-C3), if the base prediction procedure satisfies (PRG-0-C2). Hence, any prediction procedure that can be used for zero-step can also be used for one-step as long as the convergence assumption on $\widehat{Q}_n$ is satisfied. We make this precise in the following result.

Corollary 3.4.9 (Verification of one-step deterministic profile program)**.** *Assume the setting of Lemma 3.4.8. In addition, suppose $R^{\det}(\phi_1;\widetilde{f})$ satisfies* (PRG-0-C2)*. Then, $\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})$ satisfies* (PRG-1-C1)–(PRG-1-C3) *and hence satisfies* (DETPAR-1)*.*

Therefore, the prediction procedures mentioned in Remark 3.3.16 can be easily shown to satisfy (DETPAR-1). Although we assume that $\widehat{Q}_n$ converges weakly to Q in probability, we only need in probability convergence of $\int f(r)\,\mathrm{d}\widehat{Q}_n(r)$ to $\int f(r)\,\mathrm{d}Q(r)$ for $f(r) = r/(v(0;\phi_2)r+1)^2$, which is a weaker requirement. Intuitively, this assumption comes from the representation of $\widetilde{f}(x;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})$ in (3.51) as $\widetilde{f}(x;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}}) = x^\top \widehat{A}\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + x^\top \widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_{2,m}})$ for some random matrix $\widehat{A}$; see Lemma C.5.1. Hence, the risk of $\widetilde{f}$ can be written in terms of a weighted prediction error of $\widetilde{\beta}(\mathcal{D}_{k_{1,m}})$ with the weights depending on $f(\cdot)$; see (C.69).

Proposition 3.4.10 (Verification of (DETPAR-1) for the MN2LS base procedure)**.** *Assume the setting of Proposition 3.3.14. Then, the one-step ingredient predictor constructed from the MN2LS base prediction procedure satisfies* (DETPAR-1)*.*

Proposition 3.4.11 (Verification of (DETPAR-1) for the MN1LS base procedure)**.** *Assume the setting of Proposition 3.3.15. Then, the one-step ingredient predictor constructed from the MN1LS base prediction procedure satisfies* (DETPAR-1)*.*

Remark 3.4.12 (Comparison of zero and one-step procedure for isotropic covariance)**.** In order to get an intuition about the risk of one-step procedure, consider the case of isotropic features. In this case, $R^{\det}(\phi_1,\phi_2;\widetilde{f})$ simplifies to

$$R^{\det}(\phi_1,\phi_2;\widetilde{f}) = \begin{cases} R^{\det}(\phi_1;\widetilde{f}) & \text{if } \phi_2 = \infty \\ R^{\det}(\phi_1;\widetilde{f})\left(1-\dfrac{1}{\phi_2}\right) + \sigma^2\left(\dfrac{1}{\phi_2}+\dfrac{1}{\phi_2-1}\right) & \text{if } \phi_2 \in (1,\infty) \\ \sigma^2\left(\dfrac{1}{1-\phi_2}\right) & \text{if } \phi_2 \in (0,1). \end{cases} \tag{3.59}$$

Note that $\phi_2 = \infty$ corresponds to simply using the base predictor without any one-step residual adjustment. This is the same as the ingredient predictor used in the zero-step prediction procedure. The one-step prediction procedure would minimize the expression shown in (3.59), over ϕ_1 and ϕ_2 satisfying $\phi_1^{-1} + \phi_2^{-1} \le \gamma^{-1}$. If the optimal ϕ_2 turned out to be ∞, then one-step predictor and the zero-step predictor become the same, and the resulting limiting risk is $R^{\mathrm{det}}(\phi_1; \widetilde{f})$. From (3.59), the risk for $\phi_2 \in (1, \infty)$ can be decomposed as

$$R^{\mathrm{det}}(\phi_1; \widetilde{f}) + \left(\frac{\sigma^2}{\phi_2} + \frac{\sigma^2}{\phi_2 - 1} - \frac{R^{\mathrm{det}}(\phi_1; \widetilde{f})}{\phi_2} \right).$$

If the quantity in the parenthesis is negative for some (ϕ_1, ϕ_2) satisfying the condition $\phi_1^{-1} + \phi_2^{-1} \le \gamma^{-1}$, then the one-step prediction procedure will yield a strictly better risk than the zero-step prediction procedure (for $M = 1$).

One can gain more insight into how one-step procedure improves on the zero-step by considering the case of isotropic covariance and MN2LS base prediction procedure. The intriguing finding in this case is that the one-step prediction procedure with base MN2LS procedure is effectively the same as applying MN2LS on new data with reduced signal energy and with a larger limiting aspect ratio.

Formally, under isotropic covariance with MN2LS base procedure, R^{det} can be written as follows. Recall ρ^2 denotes the limit of $\|\beta_0\|_2^2$ and σ^2 is the noise variance. Then, one has

$$\begin{aligned} & R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f}_{\mathrm{mn2}}) \\ &= \begin{cases} \left[\rho^2 \left(1 - \dfrac{1}{\phi_1} \right) + \sigma^2 \left(\dfrac{1}{\phi_1 - 1} \right) \right] \left(1 - \dfrac{1}{\phi_2} \right) + \sigma^2 \left(\dfrac{1}{\phi_2 - 1} \right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (1, \infty] \times (1, \infty] \\ \left[\sigma^2 \left(\dfrac{\phi_1}{1 - \phi_1} \right) \right] \left(1 - \dfrac{1}{\phi_2} \right) + \sigma^2 \left(\dfrac{1}{\phi_2 - 1} \right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (0, 1) \times (1, \infty) \\ \sigma^2 \left(\dfrac{\phi_2}{1 - \phi_2} \right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (0, \infty) \times (0, 1). \end{cases} \end{aligned}$$

Here, we treat $1/x$ and $1/(x-1)$ to be 0 when $x = \infty$.

Let $R^{\mathrm{det}}_{\mathrm{mn2}}(\phi; \rho^2, \sigma^2)$ denote the asymptotic risk profile of the MN2LS predictor at aspect ratio ϕ, signal energy ρ^2, and noise energy σ^2; from the proof of Proposition 3.3.14 (see also Hastie et al., 2022, Theorem 1), we have

$$R^{\mathrm{det}}_{\mathrm{mn2}}(\phi; \rho^2, \sigma^2) = \begin{cases} \rho^2 \left(1 - \frac{1}{\phi} \right) + \sigma^2 \left(\frac{1}{\phi - 1} \right) + \sigma^2 & \text{if } \phi \in (1, \infty] \\ \sigma^2 \left(\frac{\phi}{1 - \phi} \right) + \sigma^2 & \text{if } \phi \in (0, 1). \end{cases}$$

Let $R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1, \phi_2; \rho^2, \sigma^2)$ denote the asymptotic risk profile of the one-step ingredient predictor with MN2LS base predictor with signal and noise energy ρ^2 and σ^2, respectively – which above we have denoted with $R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f}_{\mathrm{mn2}})$. Then, we can write

$$R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1, \phi_2; \rho^2, \sigma^2) = R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_2; R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1; \rho^2, \sigma^2) - \sigma^2, \sigma^2). \tag{3.60}$$

Thus, the limiting risk of the one-step predictor computed on a data with limiting aspect ratio γ is given by

$$R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_2(\gamma); R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1(\gamma); \rho^2, \sigma^2) - \sigma^2, \sigma^2), \tag{3.61}$$

where $(\phi_1(\gamma), \phi_2(\gamma))$ represents the minimizer of $R^{\mathrm{det}}_{\mathrm{mn2}}(\zeta_1, \zeta_2; \rho^2, \sigma^2)$ over $\zeta_1^{-1} + \zeta_2^{-1} \le \gamma^{-1}$. Now the risk expression (3.61) can be interpreted as follows: The one-step prediction procedure with base MN2LS procedure is effectively the same as applying MN2LS on new data with reduced signal energy (because $R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1(\gamma); \rho^2, \sigma^2) < \rho^2 + \sigma^2$) and with a larger limiting aspect ratio $\phi_2(\gamma) > \gamma$. Note that reducing the signal energy reduces the risk for MN2LS due to a reduction in the estimation bias; see Figure C.6 and Lemma C.6.18 (5). Recall that the effect of the zero-step procedure would just be applying MN2LS on a data set with a large limiting aspect ratio, but with the original signal energy ρ^2. Hence, the improvement of the one-step procedure over the zero-step procedure (which only takes place in the overparametrized

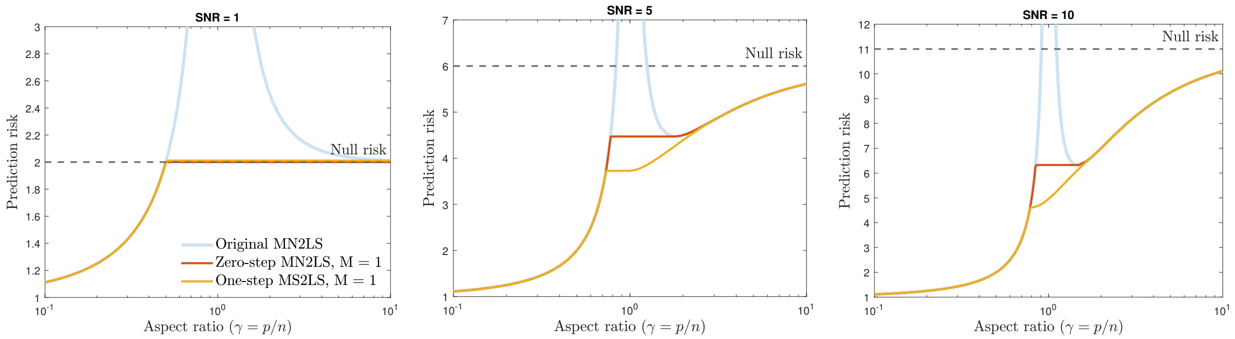


Figure 3.5: Comparison of zero-step and one-step procedures with MN2LS base procedures under isotropic feature covariance, and low, moderate, and high SNR regimes. Observe that for SNR $= 1$, zero-step and one-step both have the same risk profile with $M = 1$. This holds true even for SNR ≤ 1, as shown in Theorem C.6.16. For SNR > 1, there exists a range of γ for which one-step is strictly better than zero-step. See Theorem C.6.16 for more details.

regime) essentially stems from reducing the signal energy and thus the bias, which "boosts" the asymptotic risk.

In this case, we can also explicitly carry out the optimization of minimizing $R^{\det}(\zeta_1, \zeta_2; \widetilde{f})$ subject to the constraint $\zeta_1^{-1} + \zeta_2^{-1} \le \gamma^{-1}$. See Appendix C.6.7 for the details. See Figure 3.5 for an illustration of the comparison the limiting risk of the one-step prediction procedure with the the zero-step prediction procedure.

Finally, we comment that for base predictors other than the MN2LS, the risk of one-step procedure may not have as nice an interpretation as "boosting" the asymptotic risk by reducing the signal energy in addition to increasing aspect ratio. However, the message is that the one-step procedure adds another knob to the zero-step procedure which leads to an improved risk.

3.4.4 Numerical illustrations

In this section, we provide numerical illustration of the risk monotonization of one-step prediction procedure in the proportional asymptotic regime, when the base prediction procedures are MN2LS and MN1LS prediction procedures, and the one-step adjustment is *always* performed via MN2LS. In order to illustrate risk monotonization as in Theorem 3.4.4, we need to show the risk behavior of $\widehat{f}^{\mathrm{os}}$ at different aspect ratios. We use the same simulation settings used for the illustration of the zero-step procedure in Section 3.3.4. Figures 3.6 and 3.7 present our simulation results. The conclusions are essentially the same as those stated for the zero-step procedure in Section 3.3.4.

Minimum ℓ_2-norm least squares (MN2LS). Figure 3.6 shows the risks of the baseline MN2LS procedure and the one-step prediction procedure with MN2LS as the base prediction procedure for high and low SNR regimes (left: SNR $= 4$; right: SNR $= 1$); we take $\sigma^2 = 1$, so that ρ^2=SNR. We also present the null risk ($\rho^2 + \sigma^2$), i.e., the risk of the zero predictor as a baseline in both the plots.

Similar to the behavior of the zero-step procedure we observe that the risk of the one-step procedure is non-decreasing in γ for every $M \ge 1$. Although the risk of the one-step procedure is close to being below the risk of the base procedure, Figure 6 shows the effects of working with a finite sample. (The risk of one-step for $M = 1$ is sometimes above the risk of the base procedure.)

Figure 3.6 also shows that the one-step prediction procedure can be strictly better than the zero-step prediction procedure. In particular, the left panel of Figure 6 shows that around the interpolation threshold of 1, the risk of one-step prediction procedure is not flat. It is strictly increasing. The risk of one-step procedure for $M > 1$ is once again seen to be a strict improvement over $M = 1$.

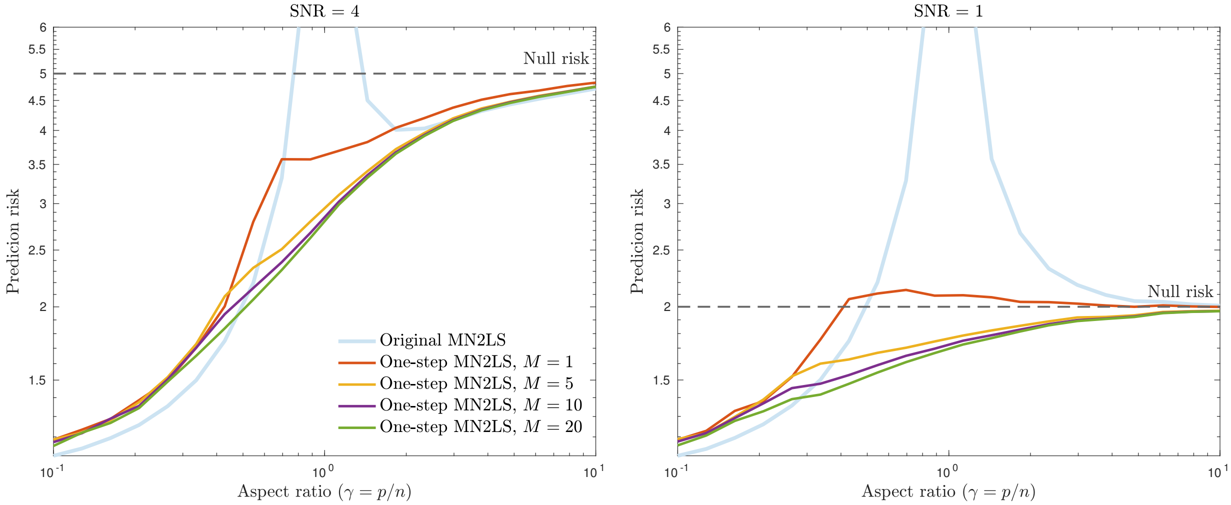


Figure 3.6: Illustration of the one-step procedure with the MN2LS as the base predictor and MN2LS one-step adjustment with varying M. The left panel shows a high SNR setting (SNR $= 4$), while the right panel shows a low SNR setting (SNR $= 1$). The setup has $n = 1000$, $n_{\text{tr}} = 900$, $n_{\text{te}} = 100$, $n^{\nu} = 50$. The features are drawn from an isotropic Gaussian distribution, the response follows a linear model with dense signal. The risks are averaged over 100 dataset repetitions.

Minimum ℓ_1-norm least squares (MN1LS). Figure 3.7 shows the risks of the baseline MN1LS procedure and the one-step procedure with MN1LS as the base prediction procedure for high (left, SNR $=$ 4) and low (right, SNR $= 1$) SNR regimes. We take $\sigma^2 = 1$ and $\rho^2 =$ SNR. We also present the null risk ($\rho^2 + \sigma^2$), i.e., the risk of the zero predictor as a baseline in both the plots. We again observe that the risk of the one-step procedure for every $M \geq 1$ is non-decreasing in γ. As before, once again we observe in Figure 3.7 that the one-step procedure with $M = 1$ attains precise risk monotonization while zero-step with $M > 1$ improves significantly upon the $M = 1$ case when γ is near one. All these comments hold for both low and high SNR regimes.

3.5 Discussion

In this work, we have proposed a generic cross-validation framework to monotonize any given prediction procedure in terms of the sample size. We studied two concrete methodologies: zero-step and one-step prediction procedures. The ingredient predictors for the zero-step prediction procedure is the base procedure applied on a subset of the data. The ingredient predictor for the one-step prediction procedure can be thought of as boosting applied to the base procedure learned on a subset of data (Schapire and Freund (2013)). In both cases, we also introduced averaging over the subsets of the data (via the parameter M). This particular averaging step can be seen as bagging, which is known to have a variance reduction effect.

We have analyzed the properties of zero-step and one-step prediction procedures in a model-free setting under mild regularity assumptions. This is in contrast to many other works in this literature that require strong distributional assumptions. In part this is possible because we assume the existence of the limiting risk and monotonize it (in a data-driven way) without requiring the knowledge/form of the risk.

Monotonization of asymptotic risk also has implications for minimax risk. If the base prediction procedure has a finite asymptotic risk $\underline{R}$ and $\overline{R}$, respectively, at the limiting aspect ratios of 0 and ∞, then both zero-step and one-step prediction procedures applied to such a base procedure yield predictors whose asymptotic risk lies between $[\underline{R}, \overline{R}]$ for all limiting aspect ratios. For example, for the squared error loss and a linear model, the MN1LS and MN2LS predictors have $\underline{R} = \sigma^2$ and $\overline{R} = \|\beta_0\|_\Sigma^2 + \sigma^2$, where σ^2 is the noise energy, which is also the unavoidable prediction risk, and $\|\beta_0\|_\Sigma^2$ is the effective signal energy. Because σ^2 is the unavoidable prediction risk, and hence a minimax lower bound, the zero-step and

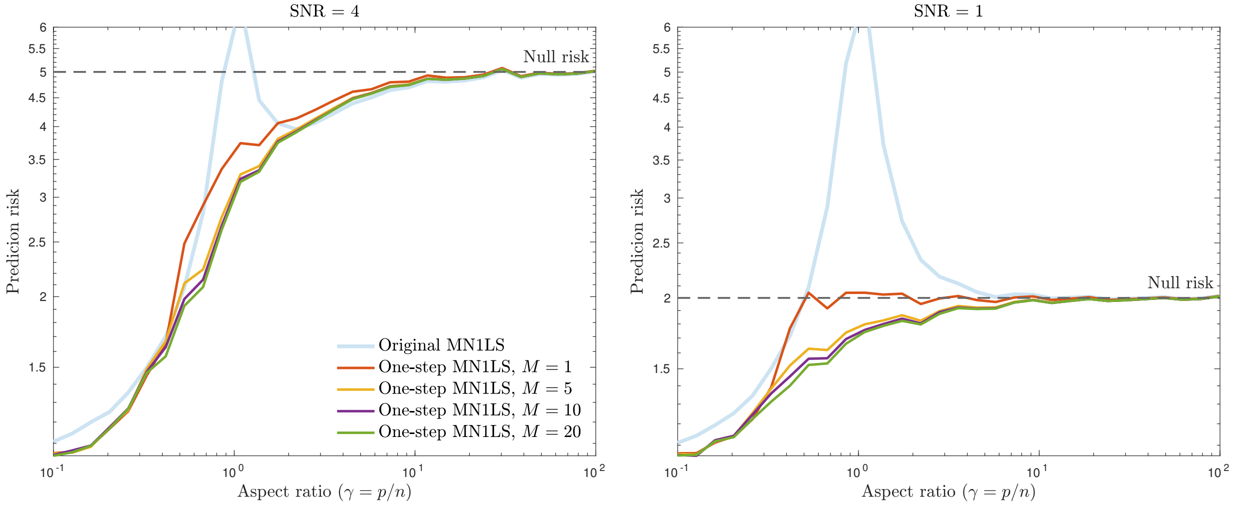


Figure 3.7: Illustration of the one-step procedure with MN1LS as the base procedure and MN2LS one-step adjustment with varying M. The left panel shows a high SNR setting (SNR = 4), while the right panel shows a low SNR setting (SNR = 1). In the setup, $n = 500$, $n_{\text{tr}} = 420$, $n_{\text{te}} = 80$, $n^{\nu} = 42$. The features are drawn from an isotropic Gaussian distribution, the response follows a linear model with sparse signal (sparsity level = 0.0005). The risks are averaged over 100 dataset repetitions.

one-step predictors based on MN1LS and MN2LS are minimax optimal up to a multiplicative factor of $1 + \text{SNR} = 1 + \|\beta_0\|_\Sigma^2/\sigma^2$ over all aspect ratios ranging from 0 to ∞. Any base prediction procedure that leads to the null predictor (i.e., $\widehat{f}(x) = 0$ for all x) for the limiting aspect ratio of ∞ also has the same property. (Most reasonable prediction procedures would yield the null predictor as the limiting aspect ratio tends to ∞.) Furthermore, for every procedure, there exists another procedure (such as the zero-step) whose risk is at least as good and is monotone. Thus, the minimax risk is a monotone function of the limiting aspect ratio. To our knowledge, the minimax risk in the proportional asymptotics regime under generic signal structure is not available in the literature.

Although the focus of the current work is exclusively on choosing optimal sample size, one could apply the cross-validation framework proposed for selecting optimal predictors from any collection. In particular, one can use our methodology to find optimal penalty parameter for ridge regression or lasso. It can also be used to select the number of random features in random features regression or kernel features in kernel regression, or more generally, the number of parameters in a neural network. In the latter case, our procedures will yield model-wise monotonicity (Nakkiran et al., 2019).

There are several interesting future directions that one can pursue. We will discuss three specific directions below.

Theoretical characterization of the effect of bagging. We have only characterized the risk of the zero-step and one-step with $M = 1$ in terms of the limiting risk of the base procedure. In this sense, we did not fully analyze the effect of bagging ($M > 1$) for both zero-step and one-step procedures. It is of interest to characterize the effect of bagging:

> What is the limiting risk of the zero-step and one-step procedures when $M > 1$?

From the theory of U-statistics, it is expected that the risk for $M > 1$ is non-increasing in M. It is hard to however argue that the risk of zero/one-step predictors is monotone in the limiting aspect ratio when $M > 1$. The main difficulty lies in proving that the ingredient predictors for the zero-step procedure have an asymptotic risk profile for $M \geq 1$. Once this is guaranteed, the theory developed in Section 3.3.3.1 will readily imply that the zero-step procedure with $M > 1$ has an asymptotic monotonic risk profile. We now briefly mention the difficulty in proving the existence of the asymptotic risk profile for the ingredient predictor when $M > 1$.

For concreteness, consider the ingredient predictor of the zero-step prediction procedure with $M > 1$ that uses $k_n \le n$ observations. This is given by

$$\widetilde{f}_M(x) = \frac{1}{M}\sum_{j=1}^{M} \widetilde{f}(x; \mathcal{D}_{\mathrm{tr}}^j) \quad \text{with} \quad |\mathcal{D}_{\mathrm{tr}}^j| = k_n.$$

Note that we take subsets $\mathcal{D}_{\mathrm{tr}}^j$ as independent and identically distributed subsets of size k_n from the data and hence for $M = \infty$, we get

$$\widetilde{f}_\infty(x; \mathcal{D}_{\mathrm{tr}}) = \frac{1}{\binom{n}{k_n}} \sum_{1 \le i_1 < \ldots < i_{k_n} \le n_{\mathrm{tr}}} \widetilde{f}(x; \{(X_{i_j}, Y_{i_j}) : 1 \le j \le k_n\}). \tag{3.62}$$

This is a U-statistics of order k_n for every fixed x in terms of the training data. If $R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^j)) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi)$ whenever $p/k_n \to \phi$, then from the theory developed in Section 3.3.3.1, it follows that $R(\widehat{f}_M^{\mathrm{zs}}) \xrightarrow{\mathrm{p}} \min_{\zeta \ge \gamma} R^{\mathrm{det}}(\zeta)$ under (PA(γ)). Hence, the main difficulty in characterizing the effect of bagging lies in proving the existence of limit of $R(\widetilde{f})$. For the squared error loss, it can be proved that (see Appendix C.6.11)

$$R(\widetilde{f}_M) = R(\widetilde{f}_\infty(\cdot; \mathcal{D}_{\mathrm{tr}})) + \frac{1}{M}\frac{1}{\binom{n}{k_n}} \sum_{i_1, \ldots, i_{k_n}} \int \Big(\widetilde{f}(x; \{(X_{i_j}, Y_{i_j}) : 1 \le j \le k_n\}) - \widetilde{f}_\infty(x; \mathcal{D}_{\mathrm{tr}})\Big)^2 \, \mathrm{d}P_{X_0}(x). \tag{3.63}$$

It is interesting to note that the risk of $\widetilde{f}_M$ only depends on M as a linear function of $1/M$. If the base predictor $\widetilde{f}$ is non-zero almost surely, then the risk of $\widetilde{f}_M$ is a strictly decreasing function of M. Observe that (3.63) holds true even for $M = 1$ and from our results, we know that the right hand side with $M = 1$ has a finite deterministic approximation. This implies that each of the components in (3.63) is asymptotically bounded. Hence, as $M \to \infty$, we can conclude that $R(\widetilde{f}_M) - R(\widetilde{f}_\infty) \xrightarrow{\mathrm{p}} 0$.

Because $k_n \to \infty$ and $p/k_n \to \phi$, the second term in (3.63) above could be analyzed using deterministic representation for $\widetilde{f}(X_0; \{(X_{i_j}, Y_{i_j}) : 1 \le j \le k_n\})$ (e.g., Theorem 1 of Liu and Dobriban (2019) for ridge regression) and the theory of U-statistics. On the other hand, $R(\widetilde{f}_\infty)$ could also be similarly analyzed using deterministic representations and the theory of U-statistics. We leave this for future work.

Other variants of boosting. In our empirical studies, we found that the one-step predictor (for $M = 1$) which is a boosted version of the subsampled predictor has a much better performance than the zero-step predictor (with $M = 1$), especially around the interpolation threshold. For reasons unclear to us currently, the performance of one-step predictor (for $M = 1$) can be matched, at least in shape, by a zero-step predictor with some $M > 1$. In this sense, the effect of one iterate boosting can be matched by the effect of multi-subsample bagging. Furthermore, as M increases, both zero-step and one-step seem to approach the same limit in our empirical studies. The interesting aspect is that the work done by M subsample bagging is achieved by one boosting iterate. This begs the question: is there a better boosting mechanism that can match zero-step predictors performance at $M = \infty$. In particular:

> What are the other choices of one-step residual adjustments? And what is the “best” choice?

We have only analyzed the one-step residual adjustment done via MN2LS. Other choices are certainly possible: for instance, one could do MN1LS or minimum ℓ_p-norm least squares or minimum ℓ_2 robust least squares in the context of linear regression. It seems cumbersome to analyze each one of these residuals adjustments case-by-case and find the best choice. For general models, one can think of the residuals adjustment we proposed as a variant of Newton’s step for the squared error loss under homoscedasticity as mentioned in (3.41). The discussion of the “best” choice of the residual adjustment very much hinges on the question of what is the best predictor in a given model in the proportional asymptotics regime. Although we do not know the answer to this question, one can potentially target the question of deriving a residual adjustment that yields an asymptotic risk performance similar to that of the zero-step predictor with $M = \infty$. For any given predictor, is there a one iterate boosted version (i.e., one-step predictor with $M = 1$) that achieves the same asymptotic performance as the M-subsample bagging with $M = \infty$?

Similar to the one-step predictor one can develop a k-step predictor by splitting the data into potentially $(k+1)$ batches and optimizing over the number of observations in each batch. This is analogues to k-iterate boosting as our one-step procedure (with $M = 1$) is analogues to the one iterate boosting. This gets computationally intensive very quickly as k increases. Furthermore, we believe that k-step predictor combined with bagging would yield the same asymptotic risk profile as the zero- and one-step predictors with $M = \infty$. In this sense, it seems a worth problem to investigate a better one iterate booster than to investigate the k-step predictor precisely.

Comparison with other regularization strategies. On the surface, zero-step and one-step procedures might seem to use only a subset of the data, and hence might appear sub-optimal. Along the same lines, one might also wonder why not employ regularization techniques and optimize over the regularization parameter. To the first point, note that we make use of the whole data in estimating the risk and comparing predictors at different sample sizes, and hence make use of the full data. To the second point, it is somewhat surprising to report that optimally-regularized procedures such as ridge regression with optimal choice of penalty need not have monotone risk (in the limiting aspect ratio); see, for example, Figure 1 of Hastie et al. (2022). But our procedure will always lead to a monotone risk and hence makes better use of the data compared to optimum regularization procedures in general. Irrespective, it is still interesting to consider the relation between zero-step and one-step, and the optimum regularization procedures in cases where the latter has a monotone risk. In our empirical studies we found that in a well-specified linear model, zero-step and one-step procedures (with the MN2LS base procedure) with a large enough M have asymptotic risk very close to the risk of the optimum ridge regression procedure. See the left panel of Figure 3.8. In a sparse linear regression model, zero-step and one-step procedures (with the MN1LS base procedure) with a large enough M has asymptotic risk very close to the risk of the optimum lasso regression. It is also interesting to observe that the risk is monotone for optimally tuned lasso. See the right panel of Figure 3.8. The effect of both bagging and boosting with large M in this case appears to be similar. In other words, thinking of the base procedures MN2LS and MN1LS as ridge and lasso, respectively, with zero penalty parameter, the zero- and one-step predictors with M large attaining the same asymptotic risk as optimum ridge or lasso can be considered as finding optimal regularization for these procedures. Without explicitly formalizing the regularization predictor, zero- and one-step perform "optimal" implicit regularization. To what extent such similarity extends to other settings is an interesting future direction:

> Under what conditions, do zero- and one-step predictors with MN2LS/MN1LS base predictor match the asymptotic risk profile of optimized regularization of ridge/lasso regression? What other base predictors (and corresponding classes of regularized predictors) does this phenomenon extend to?

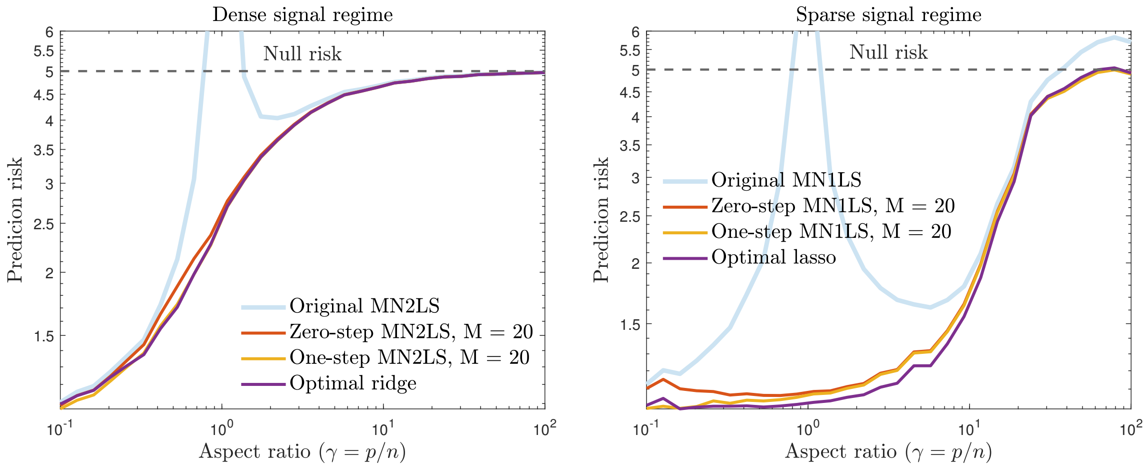


Figure 3.8: Comparison of different regularization strategies of zero-step, one-step, optimal ridge, and optimal lasso. The left panel shows a dense signal regime and the right panel shows a sparse signal regime. The setup has $n = 100$, SNR = 4. The features are drawn from an isotropic Gaussian distribution, the response follows a linear model with dense (left panel) and sparse signal (right panel, sparsity level = 0.0005). The risks are averaged over 100 dataset repetitions.

Chapter 4

Analyzing bagging

4.1 Introduction

Modern machine learning models succeed in various tasks, such as classification and regression, utilizing numerous parameters compared to the number of observations. Several commonly used estimators in such regime exhibit peculiar risk behavior, which is referred to as double or multiple descent in the risk profile (Belkin et al., 2019a; Zhang et al., 2017, 2021). The precise nature of the double or multiple descent behavior in the generalization error has been studied for various estimators: e.g., linear regression (Belkin et al., 2020; Muthukumar et al., 2020; Hastie et al., 2022), logistic regression (Deng et al., 2022), random features regression (Mei and Montanari, 2022), kernel regression (Liu et al., 2021), to name a few. We refer the readers to a survey paper by Bartlett et al. (2021) for a more comprehensive review. See also Belkin (2021); Dar et al. (2021) for other related references. In these cases, the asymptotic predictive risk behavior is often studied as a function of the data aspect ratio (the ratio of the number of parameters/features to the number of observations). The double descent behavior refers to the phenomenon where a predictor's (asymptotic) risk first increases as a function of the aspect ratio, reaches a peak (or blows up to infinity) at a point, and then decreases as a function of the aspect ratio. From a traditional statistical point of view, desirable behavior as a function of aspect ratio is not obvious. However, we can reformulate the behavior as a function of $\phi = p/n$ in terms of the observation size n with a fixed p; think of a large but fixed p and n changing from 1 to ∞. In this reformulation, the double descent behavior translates to a pattern in which the risk first decreases as n increases, then increases, reaches a peak at a point, and then decreases again with n. This is a rather counter-intuitive and sub-optimal behavior for a prediction procedure. The least one would expect from a good prediction procedure is that it yields better and better performance with more information (i.e., more data). The works mentioned above show that many commonly used predictors may not exhibit such "good" behavior. Simply put, the non-monotonicity of the asymptotic risk as a function of the number of observations or the limiting aspect ratio implies that more data may hurt (Nakkiran, 2019).

Several ad hoc regularization techniques have been proposed in the literature to mitigate the double/multiple descent behaviors. Most of these methods are trial-and-error in the sense that they do not directly target monotonizing the asymptotic risk but try a modification and check that it yields monotonic risk. Recently, Patil et al. (2022a) proposed a generic cross-validation framework that directly addresses the problem and yields a modification of any given prediction procedure that provably monotonizes the risk. In a nutshell, the method works by training the predictor on subsets of the full data (with different subset sizes) and picking the optimal subset size based on the estimated prediction risk computed using testing data. Intuitively, it is clear that this yields a prediction procedure whose risk is a decreasing function of the observation size. In the proportional asymptotics regime, where $p/n \to \phi$ as $n \to \infty$, Patil et al. (2022a) prove that this strategy returns a prediction procedure whose asymptotic risk is monotonically increasing in ϕ. In that paper, the authors have only analyzed the case where only one subset is used for each subset size, but illustrated via numerical simulations that using multiple subsets of the data of the same size (i.e., subsampling) can yield better prediction performance in addition to monotonizing the risk. Note that averaging a predictor computed on M different subsets of the data of the same size is referred in the literature as subagging, a variant of the classical bagging (bootstrap aggregation) proposed by Breiman (1996). The focus of the current work is to analyze the properties of bagged predictors in two directions (in the proportional asymptotics regime): (1) what is the asymptotic predictive risk of the bagged predictors with M bags as a function of M?, and (2) does the cross-validated bagged predictor provably yield improvements over the predictor computed on full data and further does it have a monotone risk profile (i.e., asymptotic predictive risk as a function of ϕ)? These questions are left as an open future direction in Patil et al. (2022a).

Establishing interesting connections to the simple random sampling results from survey sampling, we consider different variants of bagging that include subagging as a special case (which is also the variant considered in Patil et al. (2022a)). The second variant of bagging, which we call splagging (that stands for **spl**it-**agg**regat**ing**), is the same as the divide-and-conquer or the data-splitting approach (Rosenblatt and Nadler, 2016; Banerjee et al., 2019). The divide-and-conquer approach does not usually appear in the bagging literature, but is popularly considered in distributed learning (Dobriban and Sheng, 2020, 2021; Mücke et al., 2022). Formally, splagging is defined as a procedure that splits the data into non-overlapping parts of equal size and averages the predictors trained on these non-overlapping parts. We refer to the equal size of each part of the data as subsample size. We use the same terminology for subagging also, for simplicity. Using classical results from survey sampling and some simple lemmas about almost sure convergence, we are able to analyze the behavior of subagged and splagged predictors[1] with M bags for arbitrary prediction procedures and general $M \geq 1$. In fact, we show that the asymptotic risk of bagged predictors for general $M \geq 1$ (or simply, M-bagged predictor) can be written in terms of the asymptotic risks of bagged predictors with $M = 1$ and $M = 2$. Rather interestingly, we are able to prove that the finite sample predictive risk of the M-bagged predictor is close to its asymptotic limit uniformly over all $M \geq 1$. These results are proved in a model-agnostic setting and do not require the proportional asymptotics regime. Deriving the asymptotic risk behavior of bagged predictors with $M = 1$ and $M = 2$ has to be done on a case-by-case basis, which we do for ridge and ridgeless prediction procedures. In the context of bagging for general predictors, we further analyze the cross-validation procedure with M-bagged predictors for arbitrary $M \geq 1$ to select the "best" subsample size for both subagging and splagging. These results show that subagging and splagging for any $M \geq 1$ are better than the predictor computed on the full data. We further present conditions under which the cross-validated predictor with M-bagged predictors has an asymptotic risk monotone in the aspect ratio. Specializing these results to the ridge and ridgeless predictors, leads to somewhat surprising results connecting subagging to optimal ridge regression as well as the benefits of interpolation.

Before we proceed to give the summary of main contributions, below we present the two most significant take-away messages from our work (that hold under a well-specified linear model with an arbitrary covariance matrix for features and an arbitrary signal vector, subject to certain bounded norm constraints).

(T1) Subagging and splagging (the data-splitting approach) of the ridge and ridgeless predictors, when properly tuned, can yield significantly better prediction risks than these predictors trained on the full data. This improvement is most pronounced near the interpolation threshold. Moreover, subagging always outperforms splagging. See the left panel of Figure 4.1 for a numerical illustration and Proposition 4.5.6 for a formal statement of this result.

(T2) A model-agnostic algorithm exists to tune the subsample size for subagging that provides a predictor whose risk matches that of the oracle-tuned subagged predictor. The oracle-tuned subsample size for the ridgeless predictor is always smaller than the number of features. Consequently, subagged ridgeless *interpolators* always outperform subagged least squares, even when the full data has more observations than the number of features. The same holds true for splagging whenever it helps. See the right panel of Figure 4.1 for numerical illustrations and Proposition 4.5.7 for formal statements of this result.

[1] A note on terminology used in this work: when referring to subagging and splagging together, we use the generic term bagging. Similarly, when referring to subagged and splagged predictors together, we simply say bagged predictors.

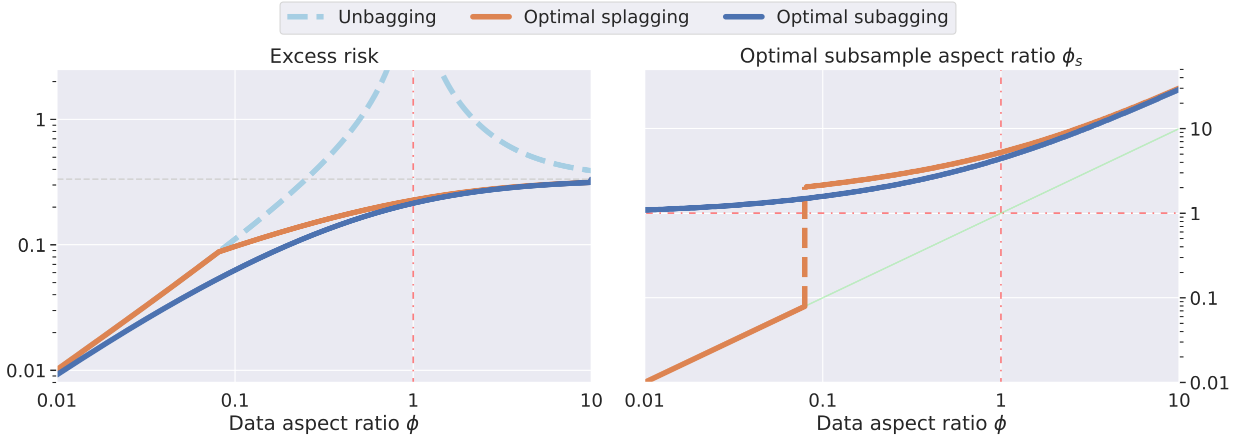


Figure 4.1: Overview of optimal bagging over both the subsample aspect ratio and the number of bags. (a) Optimal asymptotic excess risk curves for ridgeless predictors with and without bagging, under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$. The excess risk is the difference between the prediction risk and the noise level σ^2. The risk for the unbagged ridgeless predictor is represented by a blue dash line and the null risk is marked as a gray dotted line. (b) The corresponding optimal limiting subsample aspect ratio $\phi_s = p/k$ versus the data aspect ratio $\phi = p/n$ for bagged ridgeless predictors. The line $\phi_s = \phi$ is colored in green. The optimal subsample aspect ratios are larger than 1 (above the horizontal red dashed line).

Intuitively, even though bagging may induce bias because of subsampling, it can substantially reduce the prediction risk by deflating the variance for a suitably chosen subsample size that is less than the feature size. This tradeoff is possible because of the different rates at which the bias and variance of the ridgeless predictor grow near the interpolation threshold. This advantage of *interpolation* or *overparameterization* is distinct from other benefits noted in the literature, such as self-induced regularization (Bartlett et al., 2021).

4.1.1 Summary of main results

Below we provide a summary of the main results of this work.

1. **General predictors.** For the squared error loss, we measure the performance of a predictor by *data conditional risk*, which is the expected squared error on a future data point, conditional on the full data. In Section 4.3, we formulate a generic strategy for analyzing the limiting data conditional risk of general M-bagged predictors, showing that the existence of the limiting risk for $M = 1$ and $M = 2$ implies the existence of the limiting risk for every $M \geq 1$. Moreover, we show that the limiting risk of the M-bagged predictor can be written as a linear combination of the limiting risks of M-bagged predictors with $M = 1$ and $M = 2$. Interestingly, the same strategy also works for analyzing the limit of the *subsample conditional risk*, which considers conditioning on both the full data and the randomly drawn subsamples. See Theorem 4.3.9 for a formal statement. In this general context, Theorem 4.3.9 implies that both the data conditional and subsample conditional risks are asymptotically monotone in the number of bags M. Additionally, for general strongly convex and smooth loss functions, we can sandwich the risk between numbers of the form $\mathfrak{C}_1 + \mathfrak{C}_2/M$, for some fixed random variables $\mathfrak{C}_1$ and $\mathfrak{C}_2$ (Proposition 4.3.6).

2. **Ridge and ridgeless predictors.** In Section 4.4, we specialize the general strategy described above to characterize the data conditional and subsample conditional risks of M-bagged ridge and ridgeless predictors. The results are formalized in Theorem 4.4.1 for subagging with and without replacement, and Theorem 4.4.6 for splagging without replacement. All these results assume a well-specified linear model, with an arbitrary covariance matrix for the features and an arbitrary signal vector

(i.e., we assume neither Gaussian features nor isotropic features nor a randomly generated signal). These results indicate that for the three bagging strategies mentioned above, the bias and variance components are non-increasing in M.

3. **General cross-validation.** In Section 4.5, we develop a generic cross-validation strategy to select the optimal subsample or split size (or equivalently, the subsample aspect ratio) and provide a general result to understand the limiting risks of cross-validated predictors. The theoretical results serve as a way to verify monotonicity of the limiting risk of the cross-validated predictor in terms of the limiting data aspect ratio ϕ (Theorem 4.5.1).

4. **Cross-validation for bagged ridge and ridgeless predictors.** In Section 4.5.2, we specialize in the cross-validated ridge and ridgeless predictors to obtain the optimal subsample aspect ratio for every M (Theorem 4.5.5). Moreover, when optimizing over both the subsample aspect ratio and the number of bags, we show that optimal subagging is better than optimal splagging (Proposition 4.5.6). Rather surprisingly, in our investigation of the oracle choice of the subsample size for optimal subagging with $M = \infty$, we find that the subsample ratio is always large than 1 (Proposition 4.5.7). For practical data analysis, this indicates that it is always better to bag a suitably chosen ridgeless interpolator with a large M, even when the full data has more observations than features and the ordinary least squares are well-defined (Remark 4.5.8). When considering isotropic features, we find that optimally subagging the ridgeless predictor yields the same prediction risk as the optimal ridge predictor (Theorem 4.6.3).

5. **Proof techniques.** On the technical side, in the course of our risk analysis of the bagged ridge and ridgeless predictors, we derive novel deterministic equivalents for ridge resolvents with random Tikhonov-type regularization building on ideas of conditional deterministic equivalents and related calculus that may be of independent interest. See Appendix D.8 for more details.

4.1.2 Related work

Non-monotonicty of the limiting risk of commonly used predictors has been well documented in the literature. For example, recent line of work by Belkin et al. (2019a); Viering et al. (2019); Nakkiran (2019); Loog et al. (2020) exemplify the non-monotonic risk behavior of several prediction procedures. This being the suboptimal use of the data, several methods have been proposed that modify a given (class of) prediction procedure(s) so as to construct a new prediction procedure that has a monotonic risk profile. In particular, Nakkiran et al. (2021) investigates the role of optimal tuning in the context of ridge regression, and demonstrates that the optimally-tuned ℓ_2 regularization achieves monotonic generalization performance for a class of linear models under isotropic design. Mhammedi (2021) provides an algorithm to monotonize the risk profile for bounded loss functions. Patil et al. (2022a) propose a general framework to monotonize the prediction risk for general predictors and for both bounded and unbounded loss functions by cross-validation. The paper further empirically shows that bagging can further improve the performance of the predictors, along with monotonizing the risk profile. In this work, we characterize the risk behavior of bagging, which was left as an open direction in Patil et al. (2022a). Below we provide a brief overview of the literature related to bagging and its relation to our current work.

Ensemble methods combine several weak predictors to produce one powerful predictor, which is commonly used in machine learning and statistics. One important class of ensemble methods is bagging (Breiman, 1996; Bühlmann and Yu, 2002) and its variants such as subagging (Bühlmann and Yu, 2002) that average predictors trained on independent subsamples of the data. Bagging has been found to yield significant improvements in predictive performance in several empirical studies (Breiman, 1996; Strobl et al., 2009; Fernández-Delgado et al., 2014). The theoretical study of bagging has been mostly limited to smooth predictors (predictors that are smooth functions of the empirical data distribution); see Buja and Stuetzle (2006); Friedman and Hall (2007). For some work on bagging for non-parametric estimators, see Hall and Samworth (2005); Samworth (2012); Wu et al. (2021); Bühlmann and Yu (2002); Athey et al. (2019). Besides sample-wise bagging, bagging over linear combinations of features has also been considered in Lopes et al. (2011); Srivastava et al. (2016); Cannings and Samworth (2017).

Bagging in the proportional asymptotic regime has also been considered in the literature. LeJeune et al. (2020) consider subagging of both features and observations, and derive the limiting risk of the resulting subagged predictor. Dobriban and Sheng (2020, 2021); Mücke et al. (2022) consider the divide-and-conquer approach, or splagging, and study their properties. These works are set in the context of distributed learning. Under proportional asymptotics, Dobriban and Sheng (2020) derive the limiting mean squared error of the distributed ridge estimator in the underparameterized regime, while Mücke et al. (2022) provide finite-sample upper bounds on the prediction risk for ridgeless regression in the overparameterized regime.

The closest works to ours are that of LeJeune et al. (2020) and Mücke et al. (2022). LeJeune et al. (2020) consider bagged least squares predictor obtained by subsampling both the observations and features in a Gaussian isotropic design. The authors restrict subsampling in a way that the final subsampled data always has more observations than the features (so that ordinary least squares is well-defined). They also study monotonicity of the asymptotic expected squared risk with respect to the number of bags, similar to ours. They further study the best subsampling ratios for optimal asymptotic risk, but do not consider the question of how to pick the best subsample size. The most crucial difference between their work and ours is that we subsample observations (not features) while they consider feature subsampling which is not appropriate without isotopic covariance. On the other hand, Mücke et al. (2022) consider splagging and provide finite-sample upper bounds on the bias and variance components of the squared prediction risk under the assumption of sub-Gaussian features. In contrast, our results do not assume sub-Gaussianity for either feature or the noise structure in the response, and only impose minimal bounded moment assumptions.

4.1.3 Organization

- In Section 4.2, we collect basic background and results from simple random sampling that we use in our subsequent analysis of bagged predictors.
- In Section 4.3, we provide risk decompositions conditional on both the full dataset and subsampled datasets for different bagging variants for general predictors. Based on the form of decompositions, we provide a series of reductions and a generic strategy for analyzing the squared prediction risk of general bagged predictors.
- In Section 4.4, we give risk characterizations for bagging ridge and ridgeless predictors. We give results for both subagging with and without replacement, and splagging without replacement, and show monotonicities of the bias and variance components in the number of bags.
- In Section 4.5, we prescribe a framework for monotonizing the risk profile of any given predictor based on cross-validation over subsample size. The result is then specialized to the ridge and ridgeless predictors. Furthermore, we compare the monotonized risk profiles of bagged ridgeless and ridge predictors.
- In Section 4.6, we specialize our results for isotropic features and provide explicit analytic expressions for the risks of bagged ridgeless regression. In addition, we present analysis of the optimal subsample size and the corresponding optimal bagged risk.
- In Section 4.7, we conclude by providing related questions for future work.

In the supplement, we give proofs of all the results, and present additional numerical illustrations. The organization structure for the Supplement is provided in the first section of the Supplement, which also gives an overview of the general notations used in this work.

4.2 Background

In Patil et al. (2022a), the authors have proposed a generic algorithm to improve the risk monotonicity behavior of general predictors using strategies analogous to bagging and boosting. Specifically in the context of bagging, they have considered bagged predictors obtained by averaging ingredient predictors trained on M subsets of the original data. In that paper, the authors have characterized the risk of the bagged predictor for $M = 1$, and crudely bounded the risk of M-bagged predictor with that of $M = 1$. However, the numerical simulations presented therein indicate that the M-bagged predictor for $M > 1$ can have significantly better performance compared that for $M = 1$, especially around the interpolation threshold. In this work, our main goal is to develop tools that can help characterize the risk of the M-bagged predictor for any $M \geq 1$, and also to analyze the corresponding risk monotonization procedure. The M-bagged predictor considered in the aforementioned work is obtained by simple random sampling with replacement. In this work, we also extend our analysis to other versions of bagged predictors to be described shortly. The discussion in this section and that follows (Section 4.3) pertains to the development of different versions of general bagged predictors and their risk characterization.

Because the M-bagged predictor is defined through simple random sampling with replacement, the well-known results from survey sampling (Chaudhuri, 2014) are insightful for understanding its risk behavior. They also provide other versions of the bagged predictors that one can consider. For this reason, we find it useful to collect and summarize some results from survey sampling about simple random sampling with and without replacement from appropriate finite population. We briefly mention *simple random sampling with replacement* (SRSWR) and *simple random sampling without replacement* (SRSWOR) on an abstract finite population below.

Sampling with replacement. Suppose we have N numbers $\mathcal{N} = \{a_1, a_2, \dots, a_N\}$, a finite population. Let the set of indices be $\mathcal{I} := \{1, \dots, N\}$, a finite population. An SRSWR of size M from $\mathcal{I}$ is an i.i.d. draw from $\mathcal{I}$ with the uniform distribution. An unbiased estimator of the average of elements in the finite population $\mathcal{N}$ is given by

$$\widehat{\mu}^{\texttt{WR}}_{M,\mathcal{I}} = \frac{1}{M}\sum_{\ell=1}^{M} a_{I_\ell},$$

where $\{I_1, I_2, \dots, I_M\}$ is an SRSWR sample of size M from $\mathcal{I}$. It is very important to stress here that $a_1, \dots, a_N$ are all fixed numbers and only $I_1, \dots, I_M$ are random. Standard results from survey sampling (Chaudhuri, 2014, Section 2.5) imply that

$$\mathbb{E}[\widehat{\mu}^{\texttt{WR}}_{M,\mathcal{I}}] = \frac{1}{N}\sum_{\ell=1}^{N} a_\ell =: \mu, \quad \text{and} \quad \mathrm{Var}(\widehat{\mu}^{\texttt{WR}}_{M,\mathcal{I}}) = \frac{1}{M}\left(\frac{1}{N}\sum_{\ell=1}^{N}(a_\ell - \mu)^2\right). \tag{4.1}$$

Sampling without replacement. An SRSWOR of size M from $\mathcal{I}$ is a sample drawn without replacement from $\mathcal{I}$, i.e., I_1 is drawn from $\mathcal{I}$ with each element of $\mathcal{I}$ being equally likely, I_2 is drawn from $\mathcal{I} \setminus \{I_1\}$ with each element being equally likely, and so on. Define

$$\widehat{\mu}^{\texttt{WOR}}_{M,\mathcal{I}} = \frac{1}{M}\sum_{\ell=1}^{M} a_{I_\ell},$$

where $I_1, \dots, I_M$ are drawn sequentially without replacement from $\mathcal{I}$. Once again the only randomness in $\widehat{\mu}^{\texttt{WOR}}_{M,\mathcal{I}}$ stems from the randomness of $I_1, \dots, I_M$. The results from Chaudhuri (2014, Section 2.5) imply that

$$\mathbb{E}[\widehat{\mu}^{\texttt{WOR}}_{M,\mathcal{I}}] = \frac{1}{N}\sum_{\ell=1}^{N} a_\ell =: \mu, \quad \text{and} \quad \mathrm{Var}(\widehat{\mu}^{\texttt{WOR}}_{M,\mathcal{I}}) = \frac{N-M}{N-1}\frac{1}{M}\left(\frac{1}{N}\sum_{\ell=1}^{N}(a_\ell - \mu)^2\right). \tag{4.2}$$

Comparing formulas (4.1) and (4.2), one can note that both the averages are unbiased estimators of the mean of elements in the finite population, the variance of SRSWOR estimator is smaller than SRSWR (whenever $M > 1$)[2], and that the variance of SRSWOR estimator becomes zero if $M = N$. The last fact can be understood by noting that if we draw N elements without replacement from a set of N elements,

[2] Note that $(N - M)/(N - 1) = 1 - (M - 1)/(N - 1) < 1$, if $M > 1$.

then we end up with the whole set and there is no randomness left. Note $\widehat{\mu}^{\mathtt{WOR}}_{M,\mathcal{I}}$ is not well-defined if $M > N$. This particular restriction of $M > N$ becomes notationally cumbersome in the context of bagging and risk monotonization to be discussed subsequently. To avoid this, we define $\widehat{\mu}^{\mathtt{WOR}}_{M,\mathcal{I}} = \widehat{\mu}^{\mathtt{WOR}}_{N,\mathcal{I}}$ for $M > N$. This is natural in the sense that for $M \ge N$, when sampling without replacement, there is no randomness left in the estimator $\widehat{\mu}^{\mathtt{WOR}}_{N,\mathcal{I}}$. By definition, the variance of the estimator $\widehat{\mu}^{\mathtt{WOR}}_{N,\mathcal{I}}$ is 0.

4.3 Bagging general predictors

Equipped with the basics from Section 4.2, we are now ready to describe different versions of subagged predictors. Before this, let us define the index sets relevant for our study. Fix any $k \in \{1, 2, \dots, n\}$ and any permutation $\pi : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$. Define

$$\begin{aligned} \mathcal{I}_k &:= \{\{i_1, i_2, \dots, i_k\} : 1 \le i_1 < i_2 < \ldots < i_k \le n\}, \\ \mathcal{I}_k^\pi &:= \left\{ \{\pi((j-1)k+1), \pi((j-1)k+2), \dots, \pi(jk)\} : 1 \le j \le \left\lfloor \frac{n}{k} \right\rfloor \right\}. \end{aligned} \tag{4.3}$$

The set $\mathcal{I}_k$ represents the set of all k subset choices from $\{1, 2, \dots, n\}$ and there are $\binom{n}{k}$ many of them. The set $\mathcal{I}_k^\pi$ on the other hand represent the set of indices in a non-overlapping split of $\{1, 2, \dots, n\}$ into blocks of size k. If we split $\{1, 2, \dots, n\}$ randomly into different non-overlapping blocks each of size k, then this corresponds to choosing a permutation π randomly from the set of all permutations and splitting them in order. Finally, note that $\mathcal{I}_k^\pi \subseteq \mathcal{I}_k$ for any permutation π and $\cup_\pi \mathcal{I}_k^\pi = \mathcal{I}_k$.

4.3.1 Conditional risk decompositions

Suppose $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ represents a dataset with random vectors from $\mathbb{R}^p \times \mathbb{R}$. A prediction procedure $\widetilde{f}(\cdot;\cdot)$ is defined as a map from $\mathbb{R}^p \times \mathscr{P}(\mathcal{D}_n) \to \mathbb{R}$, where $\mathscr{P}(A)$ for any set A represents the power set of A. For any $I \in \mathcal{I}_k$ (or $I \in \mathcal{I}_k^\pi$), let $\mathcal{D}_I$ and the corresponding subsampled predictor be defined as

$$\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}, \quad \text{and} \quad \widehat{f}(\boldsymbol{x}; \mathcal{D}_I) = \widehat{f}(\boldsymbol{x}; \{(\boldsymbol{x}_j, y_j) : j \in I\}).$$

Given two sets of indices and two types of simple random samplings one can draw, we get four different versions of subagged predictors, as follows:

$$\left.\begin{aligned} \widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) &= \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}), \quad I_1, \dots, I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k, \\ \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) &= \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}), \quad I_1, \dots, I_M \overset{\mathtt{SRSWOR}}{\sim} \mathcal{I}_k, \end{aligned}\right\} \quad \text{Subagging} \tag{4.4}$$

$$\left.\begin{aligned} \widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) &= \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}), \quad I_1, \dots, I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k^\pi, \\ \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) &= \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}), \quad I_1, \dots, I_M \overset{\mathtt{SRSWOR}}{\sim} \mathcal{I}_k^\pi. \end{aligned}\right\} \quad \text{Splagging} \tag{4.5}$$

Traditionally, bagging (as in **b**ootstrap-**agg**regat**ing**) refers to computing predictors multiple times based on bootstrapped data (Breiman, 1996), which can involve repeated observations. In this work, we do not allow for repeated observations and consider only the four versions of bagging mentioned in (4.4)-(4.5). Bühlmann and Yu (2002, Section 3.2) call $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ as subagging (as in **sub**sample-**agg**regat**ing**). Given that SRSWOR mean estimator has smaller mean squared error than SRSWR mean estimator, we also consider the variant $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}$ of subagging. Because for any fixed M, the expectation and variance of $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}$ are the same as $N \to \infty$, the asymptotic risk behavior of $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}$ is same if $|\mathcal{I}_k| = \binom{n}{k} \to \infty$ (which holds, for example, if $1 \le k \le n-1$ and $n \to \infty$). Given this equivalence and relative popularity of subagging (i.e., $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$), in Section 4.4.2, we focus our results on $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ although we indicate the implications for $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$. In what follows, we refer to $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ as *subagging with and without replacement*, respectively.

In contrast, the predictors $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ are not usually considered in the bagging literature. They often appear in distributed learning where the predictors are trained on different parts of the data and averaged to obtain a final predictor. We call these versions as "splagging" (as in **spl**it-**agg**regat**ing**). In this case, the without replacement predictor $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ is more popular (Dobriban and Sheng, 2020; Mücke et al., 2022). Because of this and also because SRSWOR is superior to SRSWR in general, in Section 4.4.3, we focus our results on $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$. In what follows, we refer to $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ as *splagging with and without replacement.* Recall from our discussion at the end of Section 4.2 that $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ is defined as $\widetilde{f}^{\texttt{WOR}}_{\lfloor n/k\rfloor,\mathcal{I}_k^\pi}$ if $M > \lfloor n/k\rfloor$. Effectively, we are replacing M with $\min\{M, \lfloor n/k\rfloor\}$.

The results to be discussed below are general and apply to all the four versions of the bagged predictors in (4.4) and (4.5). Formulas (4.1) and (4.2) provide bias and variance of these subagged predictors conditional on the data for each $\boldsymbol{x}$. Here the finite population can be thought of as either $\{\widehat{f}(\boldsymbol{x};\mathcal{D}_I) : I \in \mathcal{I}_k\}$ or $\{\widehat{f}(\boldsymbol{x};\mathcal{D}_I) : I \in \mathcal{I}_k^\pi\}$, but with the data $\mathcal{D}_n$ treated as fixed (non-stochastic). From the bias formulas, we know that $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$ has the same expectation, given by

$$\widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}) = \frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\widehat{f}(\boldsymbol{x};\mathcal{D}_I).$$

But the variance is smaller for $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$. Using bias and variance formulas (4.1)–(4.2), the following result can be derived for the subagged predictors.

Proposition 4.3.1 (Conditional risk decomposition)**.** *Without any assumptions on the data and the prediction procedure $\widehat{f}(\cdot;\cdot)$, we have for every $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$,*

$$\begin{aligned}
\mathbb{E}[(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2 \mid \mathcal{D}_n] &= \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y) + \frac{1}{M}\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y),\\
\mathbb{E}[(y - \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2 \mid \mathcal{D}_n] &= \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y) + \frac{|\mathcal{I}_k| - M}{|\mathcal{I}_k| - 1}\frac{1}{M}\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y),
\end{aligned} \tag{4.6}$$

where

$$\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y) = (y - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2 \quad \textit{and} \quad \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y) = \frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x};\mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2. \tag{4.7}$$

The results still hold by replacing $\mathcal{I}_k$ with $\mathcal{I}_k^\pi$. Here in (4.6) the expectation is with respect to the randomness of $I_1, \ldots, I_M$ only.

In line with the traditional thinking of prediction, we care about the performance of our predictors computed on $\mathcal{D}_n$ on future data from the same distribution P. Because we only observe one dataset $\mathcal{D}_n$, we consider the behavior of the predictors in terms of the conditional (on $\mathcal{D}_n$) risk. More specifically, for a predictor $\widehat{f}$ fitted on $\mathcal{D}_n$ and its subagged predictor $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ fitted on $\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M$, with $I_1, \ldots, I_M$ being M samples of size k from $\mathcal{I}_k$, the conditional (on $\mathcal{D}_n$) risks are defined as:

$$\begin{aligned}
R(\widehat{f};\mathcal{D}_n) &:= \int (y - \widehat{f}(\boldsymbol{x};\mathcal{D}_n))^2 \,\mathrm{d}P(\boldsymbol{x},y),\\
R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot;\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M);\mathcal{D}_n) &:= \int \mathbb{E}\left[\left(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \,\middle|\, \mathcal{D}_n\right] \mathrm{d}P(\boldsymbol{x},y).
\end{aligned} \tag{4.8}$$

The conditional (on $\mathcal{D}_n$) risk of $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot;\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$ is defined similarly, and so are the conditional risks for the splagged predictors with and without replacement from $\mathcal{I}_k^\pi$ for a fixed permutation π. Observe that the conditional (on $\mathcal{D}_n$) risk of the subagged predictor $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot;\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$ integrates over the randomness of the future observation $(\boldsymbol{x}, y)$ as well as the randomness due the simple random sampling of I_ℓ, $\ell = 1, \ldots, M$.

As with the observation of a single dataset $\mathcal{D}_n$, in practice, one would only draw one simple random sample I_ℓ, $\ell = 1, \dots, M$, and hence, one might also be interested in considering another version of the conditional risk that ignores the expectation over the simple random sample:

$$R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M); \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) := \int \left(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \mathrm{d}P(\boldsymbol{x}, y). \tag{4.9}$$

We call the former conditional (on $\mathcal{D}_n$) risk *data conditional risk* and the latter *subsample conditional risk* (on $\mathcal{D}_n$ and $\{I_\ell\}_{\ell=1}^M$).

Proposition 4.3.1 implies that the data conditional risks of the predictors $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot)$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot)$ can be written as

$$\begin{aligned} R(\widetilde{f}_M; \mathcal{D}_n) &= \int \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y)\, \mathrm{d}P(\boldsymbol{x}, y) + K_{|\mathcal{I}_k|,M} \frac{1}{M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y)\, \mathrm{d}P(\boldsymbol{x}, y) \\ &= R(\widetilde{f}_\infty; \mathcal{D}_n) + \frac{K_{|\mathcal{I}_k|,M}}{M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y)\, \mathrm{d}P(\boldsymbol{x}, y) \end{aligned} \tag{4.10}$$

where for $N \geq 1$, $K_{N,M}$ is defined as

$$K_{N,M} = \begin{cases} 1 & \text{if } \widetilde{f} = \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k} \\ (N - M)_+/(N - 1) & \text{if } \widetilde{f} = \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}. \end{cases} \tag{4.11}$$

The advantage of the representation (4.10) for the data conditional risk of $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot)$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot)$ is that it allows us to obtain the limiting behavior of their risks for any $M \geq 1$ by just studying their limiting risk behavior for $M = 1$ and $M = 2$. This is trivially shown by solving a system of linear equations in two variables and is formalized in the following result.

Proposition 4.3.2 (Data conditional risk for arbitrary M)**.** *For $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}\}$, suppose there exists non-stochastic numbers a_1 and a_2 such that as $n \to \infty$,*

$$|R(\widetilde{f}_M; \mathcal{D}_n) - a_M| \xrightarrow{\text{a.s.}} 0, \quad \textit{for} \quad M = 1, 2. \tag{4.12}$$

Then, we have[3]

$$\sup_{M \in \mathbb{N}} \left| R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right] \right| \xrightarrow{\text{a.s.}} 0. \tag{4.13}$$

Observe that from Proposition 4.3.1, we have $a_1 \geq a_2$, irrespective of what prediction procedure one uses. In Proposition 4.3.2, if $a_1 > a_2$ (instead of just $a_1 \geq a_2$), then the asymptotic approximations of the conditional risk $R(\widetilde{f}_M; \mathcal{D}_n)$ is strictly decreasing in M. Similarly, we can also derive the asymptotic subsample conditional risk of subagged predictors with an arbitrary number of bags M if we know the limiting risk for $M = 1$ and $M = 2$, as summarized in Proposition 4.3.3.

Proposition 4.3.3 (Subsample conditional risk for arbitrary M)**.** *For $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}\}$, suppose there exist non-stochastic numbers b_1 and b_2 such that*

$$|R(\widetilde{f}_1; \mathcal{D}_n, I) - b_1| \xrightarrow{\text{a.s.}} 0, \quad \textit{for all } I \in \mathcal{I}_k \textit{ or } \mathcal{I}_k^\pi, \tag{4.14}$$

and

$$|R(\widetilde{f}_2; \mathcal{D}_n, \{I_1, I_2\}) - b_2| \xrightarrow{\text{a.s.}} 0, \quad \textit{for random samples } I_1, I_2\text{[4]}. \tag{4.15}$$

For any $M \in \mathbb{N}$, suppose $\{I_\ell\}_{\ell=1}^M$ is a simple random sample according to the definition of $\widetilde{f}_M$. Then, we have

$$\sup_{M \in \mathbb{N}} \left| R(\widetilde{f}_M; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left[(2b_2 - b_1) + \frac{2(b_1 - b_2)}{M}\right] \right| \xrightarrow{\text{p}} 0. \tag{4.16}$$

[3] For SRSWOR, supremum over $M \in \mathbb{N}$ should be understood as either $M \leq |\mathcal{I}_k|$ or $M \leq |\mathcal{I}_k^\pi|$ depending on whether $\widetilde{f}_M$ is $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ or $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$. The same convention is used for all the other results in this section.

[4] According to (4.4) and (4.5), I_1 and I_2 are drawn using SRSWR if $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}\}$ and SRSWOR if $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}\}$.

We make a couple of remarks on the assumption of Proposition 4.3.3 below.

Remark 4.3.4 (On the requirement (4.14))**.** Requirement (4.14) might on surface seem stronger as it requires almost sure convergence to hold for all $I \in \mathcal{I}_k$. However, recall that, for any fixed $I \in \mathcal{I}_k$, $\widetilde{f}_{1,\mathcal{I}_k}(\cdot; \mathcal{D}_I)$ is the same as the prediction procedure $\widehat{f}$ computed on the subset $\mathcal{D}_I$ with cardinality k. This implies that if the original prediction procedure satisfies almost sure convergence as the sample size on which it is trained goes to ∞, then as $k \to \infty$, the requirement (4.14) is satisfied for every fixed $I \in \mathcal{I}_k$.

Remark 4.3.5 (Role of squared loss)**.** In Propositions 4.3.2 and 4.3.3, we observed that only the limiting risks for $M = 1$ and $M = 2$ matter. This is because the data conditional risk can be decomposed as

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) = -\left(1 - \frac{2}{M}\right) R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) + 2\left(1 - \frac{1}{M}\right) R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n).$$

The subsample conditional risk admits similar decomposition as well. See Appendix D.2 for the derivations for both of them. Essentially, the interaction of subsampled datasets is only up to order two. For other loss functions, this may not be true. However, a simple monotonicity property and bounds can be obtained for a large class of loss functions as shown in the next proposition. It is also worth mentioning that while Propositions 4.3.2 and 4.3.3 are derived under the assumption that the distribution of the out-of-sample test point $(\boldsymbol{x}, y)$, $P(\boldsymbol{x}, y)$, is the same as the distribution of the training data, it is not difficult to see that the same conclusions hold for a test point sampled from any arbitrary distribution.

Proposition 4.3.6 (Convex, strongly-convex, and smooth loss functions)**.** *For any loss function* $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$*, every* $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$*, and for* $\widetilde{f}_M \in \{\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}\}$*, define*

$$R(\widetilde{f}_M; \mathcal{D}_n) = \int \mathbb{E}[L(y, \widetilde{f}_M(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)) \mid \mathcal{D}_n] \, \mathrm{d}P(\boldsymbol{x}, y).$$

If $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is convex in the second argument*[5]*, then* $R(\widetilde{f}_M, \mathcal{D}_n)$ *is non-increasing in* $M \geq 1$*, i.e.,* $R(\widetilde{f}_{M+1}; \mathcal{D}_n) \leq R(\widetilde{f}_M; \mathcal{D}_n)$*. Alternatively, if there exists* $\underline{m}, \overline{m} \in \mathbb{R}$ *such that* $L(\cdot, \cdot)$ *is* $\underline{m}$*-strongly convex and* $\overline{m}$*-smooth in the second argument*[6]*, then for* $\widetilde{f}_M \in \{\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}\}$*,*

$$\frac{\underline{m} K_{|\mathcal{I}_k|,M}}{2M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \, \mathrm{d}P(\boldsymbol{x}, y) \;\leq\; R(\widetilde{f}_M; \mathcal{D}_n) - R(\widetilde{f}_\infty; \mathcal{D}_n) \;\leq\; \frac{\overline{m} K_{|\mathcal{I}_k|,M}}{2M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \, \mathrm{d}P(\boldsymbol{x}, y) \quad (4.17)$$

with $K_{N,M}$ *defined in* (4.11)*. The inequalities in* (4.17) *continue to hold for* $\widetilde{f}_M \in \{\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}\}$*, with* $K_{|\mathcal{I}_k|,M}$ *and* $\mathscr{V}_{\mathcal{I}_k}$ *replaced with* $K_{|\mathcal{I}_k^\pi|,M}$ *and* $\mathscr{V}_{\mathcal{I}_k^\pi}$*, respectively.*

Remark 4.3.7 (Comparison with squared risk)**.** In Proposition 4.3.6, $R(\widetilde{f}_\infty; \mathcal{D}_n)$ is defined with respect to a general loss function L. Note that the upper and lower bounds of (4.17) do not depend on the loss function and are of the same form as the second term on the right-hand side of (4.10), except for constant multiples of $\underline{m}/2$ and $\overline{m}/2$. Furthermore, even when the loss function is $\overline{m}$-smooth but not convex in the second argument, the data conditional risk of $\widetilde{f}_M$ can be sandwiched between $c_1 - \overline{m}c_2/M$ and $c_1 + \overline{m}c_2/M$ for two data-dependent quantities c_1 and c_2. Beyond the squared error loss, several popular loss functions used in learning satisfy the conditions of Proposition 4.3.6; for example, the Logistic loss, the Huber loss, among others.

The next lemma connects the data conditional risk with the subsample conditional risk for $M = 1, 2$. In practice, the ingredient predictor is fitted on the subsampled datasets, on which the subsample conditional risk is evaluated. By Lemma 4.3.8, we are able to infer the data conditional risk based on the subsample conditional risk for the simple case when $M = 1, 2$.

[5] Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is convex if $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$ for all $x_1, x_2 \in \mathbb{R}$ and $t \in [0, 1]$.

[6] A function $g : \mathbb{R} \to \mathbb{R}$ is said to be λ_1-strongly convex if $x \mapsto f(x) - \lambda_1/2x^2$ is convex. It is called a λ_2-smooth function if the derivative of f is λ_2-Lipschitz (i.e., $|f'(x_1) - f'(x_2)| \leq \lambda_2|x_1 - x_2|$ for all x_1, x_2).

Lemma 4.3.8 (Transferring from subsample conditional to data conditional risk for $M = 1, 2$)**.** *Suppose the conditions in Proposition 4.3.3 hold, then* (4.12) *holds with* $a_M = b_M$ *for* $M = 1, 2$*. Consequently, the conclusions of Proposition 4.3.2 hold.*

It is worth noting that in the proof of Lemma 4.3.8, we only use the convexity of the square loss function. Therefore, analogous results can be obtained for other convex loss functions as long as the limiting subsample conditional risks exist for $M = 1, 2$.

4.3.2 General reduction strategy

Combining Proposition 4.3.2, Proposition 4.3.3, and Lemma 4.3.8 yields a general strategy to obtain both limiting subsample and data conditional risks for an arbitrary number of bag M, as presented in Theorem 4.3.9. Theorem 4.3.9 states that it is sufficient to obtain the limiting subsample conditional risks for $M = 1, 2$, as shown in Figure 4.2.

Theorem 4.3.9 (Transferring from subsample conditional to data conditional for general M)**.** *Suppose the conditions* (4.14) *and* (4.15) *hold, then the conclusions in Propositions 4.3.2 and 4.3.3 hold.*

For general predictors, both the data conditional risk and the subsample conditional risk for $M = 1$ may be available from known results. In such cases, it remains to first derive limiting subsample conditional risk for $M = 2$ depending on the sampling strategies, and then verify the properties of the limiting conditional risks required in Theorem 4.3.9. In this work, we specialize the asymptotic risk characterization to the ridge and ridgeless predictors that we will discuss next.

$$\begin{tikzcd}
\boxed{R(\widetilde{f}_M; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M),\ M = 1, 2} \arrow[r, "\text{Proposition 4.3.3}"] \arrow[d, "\text{Lemma 4.3.8}"] \arrow[dr, "\text{Theorem 4.3.9}"] & \boxed{R(\widetilde{f}_M; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M),\ M \in \mathbb{N}} \\
\boxed{R(\widetilde{f}_M; \mathcal{D}_n),\ M = 1, 2} \arrow[r, "\text{Proposition 4.3.2}"'] & \boxed{R(\widetilde{f}_M; \mathcal{D}_n),\ M \in \mathbb{N}}
\end{tikzcd}$$

Figure 4.2: General reduction strategy for obtaining limiting risks of subagged predictors with M bags.

4.4 Bagging ridge and ridgeless predictors

In this section, we follow the reduction strategy proposed in Section 4.3 to characterize the risk of subagged ridge and ridgeless predictors. In Section 4.4.1, we formally define these predictors and state the assumptions that we impose on the dataset for our results. The risk characterizations for subagging and splagging are then presented in Section 4.4.2 and Section 4.4.3, respectively.

4.4.1 Predictors and assumptions

Consider a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ consisting of random vectors in $\mathbb{R}^p \times \mathbb{R}$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the corresponding feature matrix whose j-th row contains $\boldsymbol{x}_j^\top$, and let $\boldsymbol{y} \in \mathbb{R}^n$ denote the corresponding response vector whose j-th entry contains y_j. For any index set $I \subseteq \{1, 2, \dots, n\}$, let $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ be a subsampled dataset, and let $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ denote a diagonal matrix such that $L_{jj} = 1$ if $j \in I$.

Recall that the *ridge* estimator with regularization parameter $\lambda > 0$ fitted on $\mathcal{D}_I$ is defined as

$$\widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I) = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{|I|} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

$$= (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{X}/|\mathcal{D}_I| + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{y}/|\mathcal{D}_I|).$$

The associated ridge predictor is given by $\widehat{f}_\lambda(\boldsymbol{x}; \mathcal{D}_I) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)$. The *ridgeless* estimator is the limiting estimator $\widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)$ as $\lambda \to 0^+$. When $|\mathcal{D}_I| \geq p$, and assuming that the p feature vectors are linearly independent in $\mathbb{R}^p$, it is simply the least squares estimator:

$$\widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I) = (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{X}/|\mathcal{D}_I|)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{Y}/|\mathcal{D}_I|).$$

When $|\mathcal{D}_I| < p$, it is the minimum ℓ_2-norm least squares estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I) &= \operatorname*{arg\,min}_{\boldsymbol{\beta}' \in \mathbb{R}^p} \left\{ \|\boldsymbol{\beta}'\|_2 \;\middle|\; \boldsymbol{\beta}' \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 \right\} \\ &= (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{X}/|\mathcal{D}_I|)^{+}(\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{y}/|\mathcal{D}_I|),\end{aligned}$$

where $\boldsymbol{A}^+$ denotes the Moore-Penrose inverse of matrix $\boldsymbol{A}$. Assuming that $\mathcal{D}_I$ has $|\mathcal{D}_I|$ linearly independent observation vectors in $\mathbb{R}^n$, this estimator also interpolates the data, i.e., we have $y_j = \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I)$ for $j \in I$, and has the minimum ℓ_2-norm among all interpolators. The associated ridgeless predictor is again given by $\widehat{f}_0(\boldsymbol{x}; \mathcal{D}_n) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_0(\mathcal{D}_n)$.

Given their relevance to the subagged predictors studied in the literature, we will primarily focus on only two of the four subagged predictors as defined in (4.4) and (4.5), although the implications for the other two can be trivially obtained. For $\lambda \geq 0$, the subagged and slagged predictors respectively are defined as

$$\begin{aligned}\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \mathcal{D}_n) &= \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M), \qquad I_1, \dots, I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k, \\ \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}(\boldsymbol{x}; \mathcal{D}_n) &= \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M), \qquad I_1, \dots, I_M \overset{\mathtt{SRSWOR}}{\sim} \mathcal{I}_k^\pi,\end{aligned} \tag{4.18}$$

where $\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = M^{-1} \sum_{\ell=1}^M \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_{I_\ell})$. For $M > |\mathcal{I}_k^\pi|$, the splagged predictor is defined to be the predictor with $M = |\mathcal{I}_k^\pi|$. When $\lambda = 0$, the base predictors become the ridgeless predictors.

We consider Assumptions 4.1-4.5 on the dataset $\mathcal{D}_n$ to characterize the risk, which are standard in the study of ridge and ridgeless regression under proportional asymptotics; see, e.g., Hastie et al. (2022).

Assumption 4.1 (Feature model)**.** The feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \dots, n$, multiplicatively decompose as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix and $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4 + \delta$ for some $\delta > 0$.

Assumption 4.2 (Response model)**.** The response variables $y_i \in \mathbb{R}$, $i = 1, \dots, n$, additively decompose as $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$, where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown signal vector and ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4 + \delta$ for some $\delta > 0$.

Assumption 4.3 (Signal norm)**.** The ℓ_2-norm of the signal vector $\|\boldsymbol{\beta}_0\|_2$ is uniformly bounded in p, and $\lim_{p \to \infty} \|\boldsymbol{\beta}_0\|_2^2 = \rho^2 < \infty$.

Assumption 4.4 (Covariance norm)**.** There exist real numbers $r_{\min}$ and $r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min} \boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max} \boldsymbol{I}_p$.

Assumption 4.5 (Limiting covariance and signal-weighted spectrums)**.** Let $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ denote the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ is a diagonal matrix containing eigenvalues (in non-increasing order) $r_1 \geq r_2 \geq \cdots \geq r_p \geq 0$, and $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ is an orthonormal matrix containing the associated eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_p \in \mathbb{R}^p$. Let H_p denote the empirical spectral distribution of $\boldsymbol{\Sigma}$ (supposed on $\mathbb{R}_{>0}$) whose value at any $r \in \mathbb{R}$ is given by

$$H_p(r) = \frac{1}{p} \sum_{i=1}^p \mathbb{1}_{\{r_i \leq r\}}.$$

Let G_p denote a certain distribution (supported on $\mathbb{R}_{>0}$) that encodes the components of the signal vector $\boldsymbol{\beta}_0$ in the eigenbasis of $\boldsymbol{\Sigma}$ via the distribution of (squared) projection of $\boldsymbol{\beta}_0$ along the eigenvectors $\boldsymbol{w}_j, 1 \le j \le p$, whose value at any $r \in \mathbb{R}$ is given by

$$G_p(r) = \frac{1}{\|\boldsymbol{\beta}_0\|_2^2} \sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \ \mathbb{1}_{\{r_i \le r\}} .$$

Assume there exist fixed distributions H and G (supported on $\mathbb{R}_{>0}$) such that $H_p \xrightarrow{\mathrm{d}} H$ and $G_p \xrightarrow{\mathrm{d}} G$ as $p \to \infty$.

4.4.2 Subagging with replacement

In this section, we consider the risk asymptotics and properties for subagging. In Section 4.4.2.1, we provide exact risk characterization of subagged ridge and ridgeless predictors. The monotonicity properties of asymptotic bias and variance components of the risk are presented in Section 4.4.2.2.

4.4.2.1 Risk characterization

In preparation for our first result on the risk characterization of subagged ridge and ridgeless predictors, let us introduce some notations. We will analyze the subagged predictors (with M bags) in the proportional asymptotics regime, where the original data aspect ratio (p/n) converges to $\phi \in (0, \infty)$ as $n, p \to \infty$, and the subsample data aspect ratio (p/k) converges to ϕ_s as $k, p \to \infty$. Because $k \le n$, ϕ_s is always no less than ϕ.

One of the key quantities that appears throughout our analysis of subagged ridge predictors is defined through a fixed-point equation. Such fixed point equations have appeared in the literature before in the context of risk analysis of regularized estimators under proportional asymptotics regime; see, e.g., Dobriban and Wager (2018); Hastie et al. (2022); Mei and Montanari (2022) in the context of ridge regression; and more generally, for other M-estimators, see Thrampoulidis et al. (2015, 2018), Sur et al. (2019), Karoui (2013); El Karoui (2018), Miolane and Montanari (2021), among others. For any $\lambda > 0$ and $\theta > 0$, define $v(-\lambda; \theta)$ as the unique nonnegative solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\theta)} = \lambda + \theta \int \frac{r}{1 + v(-\lambda;\theta) r} \,\mathrm{d}H(r), \tag{4.19}$$

and for $\lambda = 0$, $\theta > 1$, we define

$$v(0;\theta) = \begin{cases} \lim\limits_{\lambda \to 0^+} v(-\lambda;\theta), & \text{if } \theta > 1 \\ +\infty, & \text{if } \theta \in (0,1]. \end{cases} \tag{4.20}$$

The fact that the fixed-point equation (4.19) has a unique nonnegative solution follows from Patil et al. (2022a, Lemma S.6.14); for completeness, we also provide a proof in Appendix D.8.3. The existence of the limit of $v(-\lambda;\theta)$ as $\lambda \to 0^+$ follows because $v(-\lambda;\theta)$ is monotonically decreasing in $\lambda > 0$ (Patil et al., 2022a, Lemma S.6.15 (4)). Further, we define non-negative constants $\widetilde{v}(-\lambda;\vartheta,\theta)$ and $\widetilde{c}(-\lambda;\theta)$ via the following equations:

$$\widetilde{v}(-\lambda;\vartheta,\theta) = \frac{\vartheta \int r^2 (1 + v(-\lambda;\theta) r)^{-2} \,\mathrm{d}H(r)}{v(-\lambda;\theta)^{-2} - \vartheta \int r^2 (1 + v(-\lambda;\theta) r)^{-2} \,\mathrm{d}H(r)}, \quad \widetilde{c}(-\lambda;\theta) = \int \frac{r}{(1 + v(-\lambda;\theta) r)^2} \,\mathrm{d}G(r). \tag{4.21}$$

Theorem 4.4.1 (Risk characterization of subagged ridge and ridgeless predictors)**.** *Let $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ be the predictor as defined in* (4.18) *for $\lambda \ge 0$. Suppose Assumptions 4.1-4.5 hold for the dataset $\mathcal{D}_n$. Then,* *as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty]$ (and $\phi_s \neq 1$ if $\lambda = 0$), there exist deterministic functions $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ for $M \in \mathbb{N}$, such that for $I_1, \ldots, I_M \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\begin{aligned} \sup_{M \in \mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)| &\xrightarrow{\text{p}} 0, \\ \sup_{M \in \mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}; \mathcal{D}_n) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)| &\xrightarrow{\text{a.s.}} 0. \end{aligned} \tag{4.22}$$

The guarantee (4.22) also holds true if $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ is replaced by $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$. Furthermore, the function $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ decomposes as

$$\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) + \mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s), \tag{4.23}$$

where the bias and variance terms are given by

$$\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = M^{-1} B_\lambda(\phi_s, \phi_s) + (1 - M^{-1}) B_\lambda(\phi, \phi_s), \tag{4.24}$$

$$\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = M^{-1} V_\lambda(\phi_s, \phi_s) + (1 - M^{-1}) V_\lambda(\phi, \phi_s), \tag{4.25}$$

and the functions $B_\lambda(\cdot, \cdot)$ and $V_\lambda(\cdot, \cdot)$ are defined as

$$B_\lambda(\vartheta, \theta) = \rho^2 (1 + \widetilde{v}(-\lambda; \vartheta, \theta)) \widetilde{c}(-\lambda; \theta), \qquad V_\lambda(\vartheta, \theta) = \sigma^2 \widetilde{v}(-\lambda; \vartheta, \theta), \qquad \theta \in (0, \infty], \, \vartheta \leq \theta. \tag{4.26}$$

Theorem 4.4.1 provides exact asymptotics for the data conditional as well as the subsample conditional risks of subagged ridge and ridgeless predictors. Furthermore, we have derived the bias-variance decomposition for the asymptotic risk in (4.23). Interestingly, the individual bias term is a convex combination of $B_\lambda(\phi_s, \phi_s)$ and $B_\lambda(\phi, \phi_s)$, which correspond to the biases for $M = 1$ and $M = \infty$, respectively. Analogous conclusion also holds for the variance term. Even though the risk behavior for $M = 1$ has been studied by Patil et al. (2022a), the one for general (data-dependent) M is novel. As we shall see later in Section 4.6, the risk behavior for $M = \infty$ is drastically different from the one for $M = 1$.

Note that when $\theta > 1$, the parameter $v(0; \theta)$ defined in (4.20) can also be seen as the unique nonnegative solution to the following fixed-point equation (Patil et al., 2022a, Lemma S.6.14):

$$\frac{1}{v(0; \theta)} = \theta \int \frac{r}{1 + v(0; \theta) r} \, \mathrm{d}H(r). \tag{4.27}$$

When $\theta \in (0, 1]$, since $\lim_{\lambda \to 0^+} v(-\lambda; \theta) = \infty$, we have that $\lim_{\lambda \to 0^+} \widetilde{c}(-\lambda; \theta) = 0$ and $\lim_{\lambda \to 0^+} \widetilde{v}(-\lambda; \vartheta, \theta) = \vartheta(1 - \vartheta)^{-1}$. Therefore, the bias and variance functions in (4.26) for $\vartheta \leq \theta$ reduce to

$$B_0(\vartheta, \theta) = \begin{cases} 0 & \theta \in (0, 1] \\ \rho^2 (1 + \widetilde{v}(0; \vartheta, \theta)) \widetilde{c}(0; \theta) & \theta \in (1, \infty] \end{cases}, \qquad V_0(\vartheta, \theta) = \begin{cases} \sigma^2 \dfrac{\vartheta}{1 - \vartheta} & \theta \in (0, 1) \\ \infty & \theta = 1 \\ \sigma^2 \widetilde{v}(0; \vartheta, \theta) & \theta \in (1, \infty]. \end{cases} \tag{4.28}$$

As a sanity check when $\vartheta = \theta$, it is easy to see that the bias and variance components collapse to that of the minimum ℓ_2-norm least squares estimator with limiting aspect ratio θ.

A few additional remarks on Theorem 4.4.1 follow.

Remark 4.4.2 (Data conditional versus subsample conditional risks)**.** Theorem 4.4.1 shows that the data conditional risk and the subsample conditional risk both converge to the same deterministic limit. Intuitively, this is expected because the data conditional risk is the average subsample conditional risks over all subsamples.

Remark 4.4.3 (Extending theorem to negative regularization)**.** For $\lambda < 0$, the fixed-point equation (4.19) may have more than one solution. However, there still exists a solution to (4.19) with which Theorem 4.4.1 holds whenever $\lambda > -(1 - \sqrt{\phi})^2 r_{\min}$ where $r_{\min}$ is the uniform lower bound on the smallest eigenvalue of $\mathbf{\Sigma}$. In this work, for simplicity we restrict to the case when $\lambda \geq 0$.

Remark 4.4.4 (The requirement of $\phi_s \neq 1$ for $\lambda = 0$)**.** When $\lambda = 0$, the base predictors are ridgeless predictors. In this case, the variance function $\theta \mapsto \mathscr{V}_{0,M}(\vartheta, \theta)$ is unbounded if M is finite and $\theta \to 1$ because $V_0(\theta, \theta)$ in (4.28) diverges as $\theta \to 1$. Empirically, this can be explained by the singularity of the empirical covariance matrices with aspect ratios close to 1. However, the asymptotic risk for $M = \infty$ is always bounded.

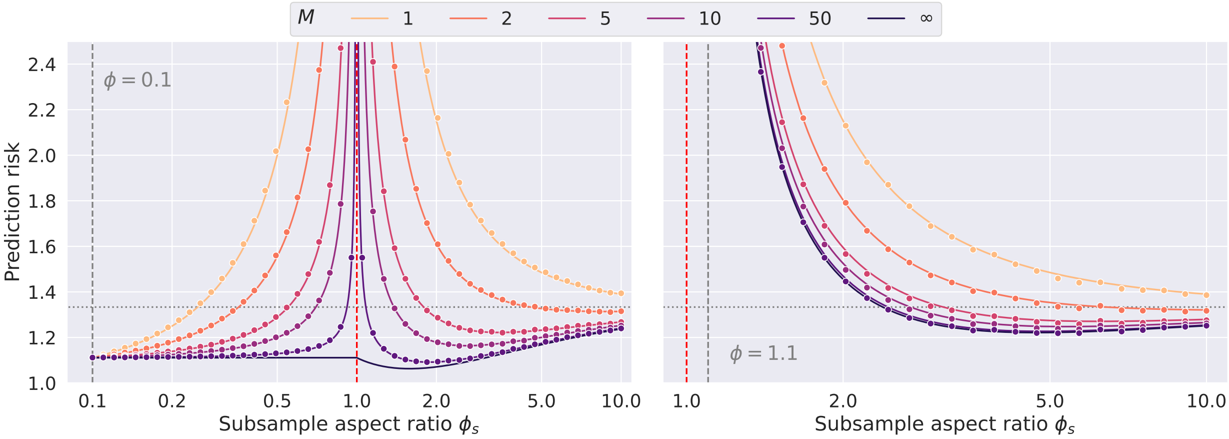


Figure 4.3: Asymptotic prediction risk curves in (4.23) for subagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = \lfloor p\phi \rfloor$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

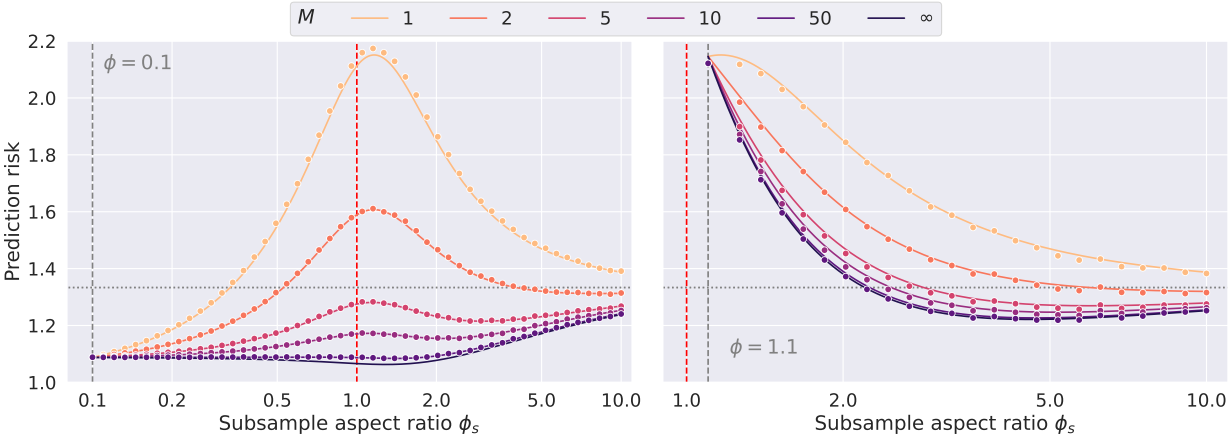


Figure 4.4: Asymptotic prediction risk curves in (4.23) for subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = \lfloor p\phi \rfloor$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

Illustration of Theorem 4.4.1. Before we present the proof outline for Theorem 4.4.1, we first provide some numerical illustrations under the AR(1) data model. The covariance matrix of an auto-regressive process of order 1 (AR(1)) is given by $\boldsymbol{\Sigma}_{\text{ar1}}$, where $(\boldsymbol{\Sigma}_{\text{ar1}})_{ij} = \rho_{\text{ar1}}^{|i-j|}$ for some parameter $\rho_{\text{ar1}} \in (0, 1)$, and the AR(1) data model is defined as

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathrm{ar1}}), \quad \boldsymbol{\beta}_0 = \frac{1}{5}\sum_{j=1}^{5} \boldsymbol{w}_{(j)}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{M-AR1-LI}$$

where $\boldsymbol{w}_{(j)}$ is the eigenvector of $\boldsymbol{\Sigma}_{\mathrm{ar1}}$ associated with the top jth eigenvalue $r_{(j)}$. From Grenander and Szegö (1958, pp. 69-70), the top j-th eigenvalue can be written as $r_{(j)} = (1-\rho_{\mathrm{ar1}}^2)/(1-2\rho_{\mathrm{ar1}}\cos\theta_{jp} + \rho_{\mathrm{ar1}}^2)$ for some $\theta_{jp} \in ((j-1)\pi/(p+1), j\pi/(p+1))$. Then, under model (M-AR1-LI), the signal strength ρ^2 defined in Assumption 4.3 is $5^{-1}(1-\rho_{\mathrm{ar1}}^2)/(1-\rho_{\mathrm{ar1}})^2$, which is the limit of $25^{-1}\sum_{j=1}^{5} r_{(j)}$. Thus model (M-AR1-LI) parameterized by two parameters ρ_{ar1} and σ^2 satisfies Assumption 4.1-4.5. Figures 4.3 and 4.4 show the limiting risk for the subagged ridgeless predictor and subagged ridge predictor, respectively, with the number of bags M varying from 1 to ∞. In the plots, the limiting aspect ratio ϕ of the full data is fixed to be either 0.1 or 1.1, corresponding to the cases when $n > p$ and $n < p$, respectively. For each case, the limiting aspect ratio ϕ_s of each bag takes values in (ϕ, ∞).

We observe that the empirical risks match the deterministic approximations for both cases, and they are more concentrated around the deterministic approximations as M increases, which is expected as the variance of the subagged predictors reduces with M. Furthermore, for any fixed ϕ_s, the asymptotic risk decreases as M increases.

Because of the non-monotonic risk behavior of the underlying ridge and ridgeless predictors, Figures 4.3 and 4.4 show that the best subsample aspect ratio (ϕ_s) in terms of prediction risk might be strictly larger than ϕ. This is true for any choice of $M \geq 1$. The case of $M = 1$ was already mentioned in Patil et al. (2022a). This observation is interesting as it indicates it is better to bag predictors that use even less number of observations than the original data. Similar phenomena are also observed in our simulations with varying signal-to-noise ratios (see Appendix D.10.1). Finding the optimal choice of ϕ_s via an actionable algorithm in practice is discussed in Section 4.5.

Proof outline of Theorem 4.4.1. The proof of Theorem 4.4.1 uses the reduction strategy discussed in Section 3. In particular, we apply Theorem 4.3.9 (subsample conditional for $M = 1$ and $M = 2$ to subsample and data subsample for any M) to prove the theorem.

1. The deterministic risk approximation to the subsample conditional risk for $M = 1$ can be obtained from the results of Patil et al. (2022a) that build on Hastie et al. (2022).

2. Under the linear model, to analyze the subsample conditional risk for $M = 2$, we first decompose it as:

$$\begin{aligned} &R(\widetilde{f}_2; \mathcal{D}_n, \{I_1, I_2\}) \\ &= \sigma^2 + \frac{1}{4}\sum_{i=1}^{2} (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_i}))^\top \Sigma (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_i})) + \frac{1}{2}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top \Sigma (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2})) \\ &= \frac{\sigma^2}{2} + \frac{R(\widetilde{f}_1; \mathcal{D}_n, I_1) + R(\widetilde{f}_2; \mathcal{D}_n, I_2)}{4} + \frac{(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top \Sigma (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))}{2}. \end{aligned} \tag{4.29}$$

The first term in the display above is non-random. The asymptotic risk approximation for the second term follows from the asymptotics of the subsample conditional risk for $M = 1$. The challenging part is the analysis of the final cross term $(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top \Sigma (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))$, because of the non-trivial dependence implied by the overlap between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$. Our strategy to obtain a deterministic approximation for such a term is to write $h(\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}), \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2})) = h(\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1'} \cup \mathcal{D}_{I_0}), \widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2'} \cup \mathcal{D}_{I_0}))$ for any univariate function h. Here, $I_0 = I_1 \cap I_2$ denotes the indices of the overlap and $I_j' = I_j \setminus I_0$ for $j = 1, 2$ are the indices of non-overlapping observations. Observe that conditioning on $\mathcal{D}_{I_0}$, $\mathcal{D}_{I_1'}$ and $\mathcal{D}_{I_2'}$ are independent datasets. This conditional independence, along with the closed-form expression of the ridge predictor, forms a crucial piece in our argument. To carry out this program, we derive conditional deterministic equivalence results for ridge resolvents.

3. In order to prove the results for the ridgeless predictor, we essentially take the limit as $\lambda \to 0^+$ of the deterministic risk approximation for the ridge predictor with regularization λ. This requires appealing to a uniformity argument in λ (see Appendix D.4 for more details).

4.4.2.2 Monotonicity of bias and variance in number of bags

Monotonicity in the number of bags M for both the data conditional risk $R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}; \mathcal{D}_n)$ and the subsample conditional risk $R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ follow from (4.10). In the classical literature of bagging and subagging, however, it has been of interest to better understand the effect of aggregation on not just the risk, but also on the bias and variance. In this section, we show for the ridge and ridgeless predictors, subagging reduces both the bias and the variance. Monotonicity of the risk proved in Theorem 4.4.1, does not imply the monotonicity of asymptotic bias and variance components. Fortunately, from the risk decomposition derived in Theorem 4.4.1, both asymptotic bias and variance components are monotonic in M as shown below.

Proposition 4.4.5 (Improvement due to subagging)**.** *For all $M = 1, 2, \ldots$ and $\lambda \in [0, \infty)$, it holds that*

$$\mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) \leq \mathscr{B}^{\texttt{sub}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) \tag{4.30}$$

$$\mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) \leq \mathscr{V}^{\texttt{sub}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s). \tag{4.31}$$

The inequalities in (4.30) are strict whenever $\rho^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$), while the inequalities in (4.31) are strict when $\sigma^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$). Thus, the asymptotic risk is monotonically decreasing in M: $\mathscr{R}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) \leq \mathscr{R}^{\texttt{sub}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$.

The monotonicity property mentioned in Proposition 4.4.5 does not immediately follow from the decomposition of $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ in (4.24) and (4.25). All that is implied by (4.24) and (4.25) is that $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ either monotonically increase or decrease in $M \geq 1$. Proposition 4.4.5 confirms that they are both decreasing in M by proving that $\mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) \geq \mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) \geq \mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s)$. Further, it explicitly distinguishes the cases of non-increasing and strict decreasing of the bias and variance components.

We visualize the bias and variance components for subagged ridgeless predictors in Figure 4.5 under model (M-AR1-LI). Figure 4.5 validates the monotonicity properties claimed in Proposition 4.4.5. For a similar illustration for subagged ridge predictors, see Figure D.7 in the appendix.

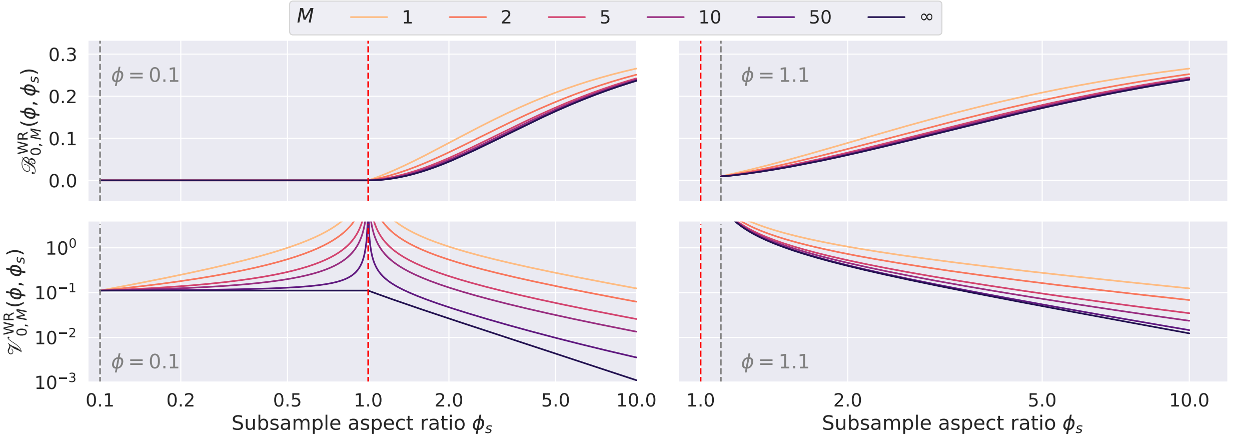


Figure 4.5: Asymptotic bias and variance curves in (4.26) for subagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 0.25$, for varying subsample aspect ratio ϕ_s and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\texttt{sub}}_{0,M}(\phi, \phi_s)$ are shown on a log-10 scale.

4.4.3 Splagging without replacement

In this section, we consider the risk asymptotics and properties for splagging. More formally, we consider the risk asymptotics of the splagged predictor obtained by averaging the predictors computed on M non-overlapping subsets of the data each of size k. This is precisely the splagged predictor $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$. In all the asymptotics below, we consider the permutation π to be fixed. Because the limiting risk below does not depend on the permutation π, the conclusions continue to hold true even when the data or subsample conditional risk is averaged over all permutations π; note that this is not the same as the data conditional risk of the splagged predictor averaged over all permutations π. In Section 4.4.3.1, we provide exact risk characterization of splagging without replacement for both ridge and ridgeless predictors. The monotonicity properties of asymptotic bias and variance are presented in Section 4.4.3.2.

4.4.3.1 Risk characterization

Recall our convention is defining the splagged predictor $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ as $\widetilde{f}^{\texttt{WOR}}_{\min\{M,\lfloor n/k\rfloor\},\mathcal{I}_k^\pi}$, so that the splagged predictor is well defined for all $M \in \mathbb{N}$.

Theorem 4.4.6 (Risk characterization for splagged ridge predictor without replacement)**.** *Let $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ be the predictor as defined in* (4.18) *for $\lambda \geq 0$. Suppose Assumptions 4.1-4.5 hold for the dataset $\mathcal{D}_n$. Then as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$, $p/k \to \phi_s \in [\phi, \infty]$ (and $\phi_s \neq 1$ for $\lambda = 0$), there exist deterministic functions $\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$ for all $M \in \mathbb{N}$, and $\phi_s \geq \phi$, such that for $I_1, \dots, I_M \overset{\texttt{SRSWOR}}{\sim} \mathcal{I}_k^\pi$,*

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)| \xrightarrow{\text{p}} 0,$$
$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}; \mathcal{D}_n) - \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)| \xrightarrow{\text{a.s.}} 0,$$

where $\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = \mathscr{R}^{\texttt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s)$ for $M \geq \lfloor\phi_s/\phi\rfloor$, and for $M \leq \lfloor\phi_s/\phi\rfloor$, the function $\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$ decomposes as

$$\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) + \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) \tag{4.32}$$

where $\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = M^{-1}B_\lambda(\phi_s, \phi_s) + (1 - M^{-1})C_\lambda(\phi_s)$, $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = M^{-1}V_\lambda(\phi_s, \phi_s)$, $C_\lambda(\phi_s) = \rho^2\widetilde{c}(-\lambda; \phi_s)$, with $B_\lambda(\phi_s, \phi_s)$ and $V_\lambda(\phi_s, \phi_s)$ defined in Theorem 4.4.1.

Remark 4.4.7. For every pair (ϕ, ϕ_s) satisfying $\phi_s \geq \phi$, note that the splagged predictor and the risks are defined non-trivially only for $M = 1, \dots, \lfloor\phi_s/\phi\rfloor$, and is defined as a constant for $M > \lfloor\phi_s/\phi\rfloor$. In particular, for a fixed pair (ϕ, ϕ_s), the sequence of risks as M changes looks like

$$\mathscr{R}^{\texttt{spl}}_{\lambda,1}(\phi, \phi_s),\, \mathscr{R}^{\texttt{spl}}_{\lambda,2}(\phi, \phi_s),\, \dots,\, \mathscr{R}^{\texttt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s),\, \mathscr{R}^{\texttt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s),\, \dots.$$

Remark 4.4.8 (Dependence on data and subsample aspect ratios)**.** Even though splagging does not formally involve repeated observations like bootstrapping, we will still refer to $\phi_s = p/k$ as the subsample aspect ratio, where k is the number of observations in each split part of the full dataset. In Theorem 4.4.1 for the subagged predictor with replacement, the asymptotic risk depends on both the data aspect ratio ϕ as well as the subsample aspect ratio ϕ_s. In contrast, the asymptotic risk for the splagged predictor without replacement in Theorem 4.4.6 does not depend on the data aspect ratio ϕ. This can be seen from the expressions for $\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$. However, it is interesting to note that the asymptotic risk for $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ depends on both ϕ and ϕ_s because $\limsup_{k,n\to\infty} |\mathcal{I}_k^\pi|$ is finite, which makes the limiting risk of $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ different. Because $K_{N,M}$ defined in (4.11) is bounded above by 1 and $\limsup_{k,n\to\infty} K_{|\mathcal{I}_k^\pi|,M} < 1$ for any $M > 1$, $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ is a strictly better predictor then $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ in terms of the squared risk (i.e., $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ is inadmissible, even asymptotically).

Remark 4.4.9 (Comparison with distributed learning)**.** Theorem 4.4.1 considers the simple average of base predictors fitted on non-overlapped samples, which is also closely related to distributed learning (Mücke et al., 2022) that utilizes multiple computing devices to reduce overall training time. Mücke et al. (2022) only provide finite-sample upper bounds for the prediction risk of distributed ridgeless predictor, while Theorem 4.4.1 gives exact risk characterization. The distributed ridge predictors are also studied in Dobriban and Sheng (2020), though their goal is to obtain the optimal weight and the optimal regularization parameter.

Illustration of Theorem 4.4.6. In Figures 4.6 and 4.7, we provide numerical illustrations for Theorem 4.4.6 (bagged ridgeless and ridge predictors with $\lambda = 0.1$) under model (M-AR1-LI), with the number of bags M varying from 1 to ∞. The limiting data aspect ratio is fixed at 0.1 when $n > p$ and at 1.1 when $n < p$. We observe that the empirical risks very closely match the deterministic approximations as stated in Theorem 4.4.6 for both bagged ridge and ridgeless predictors. Similar to Figure 4.3, for any fixed M, the optimal ϕ_s may be strictly larger than ϕ, an implication of non-monotonic risk behavior.

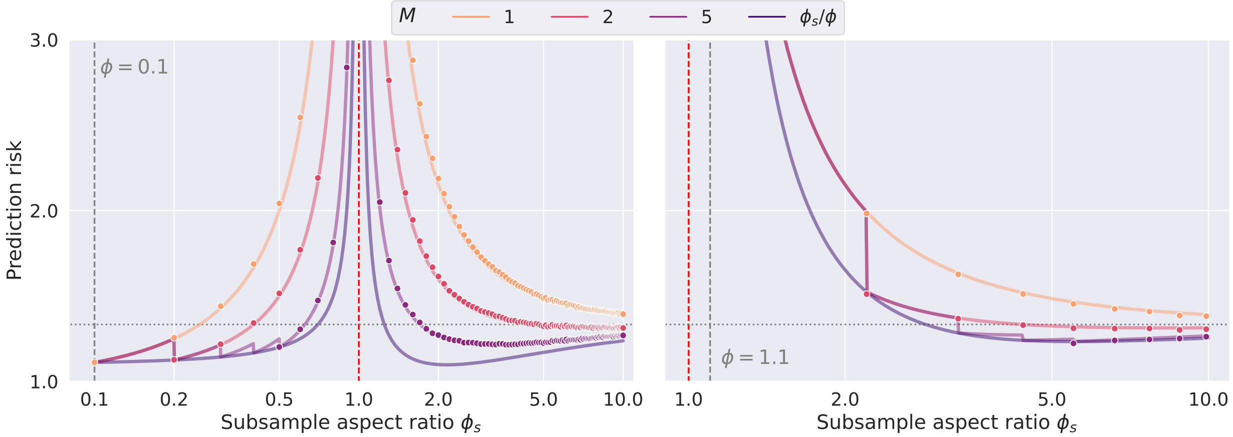


Figure 4.6: Asymptotic prediction risk curves in (4.32) for splagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying split sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $p = 500$ and $n = \lfloor p\phi \rfloor$.

Proof outline of Theorem 4.4.6. The proof of Theorem 4.4.6 follows the similar reduction strategy as in the proof of Theorem 4.4.1, where we first analyze the subsample conditional risks for $M = 1$ and $M = 2$, and appeal to Theorem 4.3.9 to obtain the result for data conditional and subsample conditional risks for any M.

1. The deterministic risk approximation to the subsample conditional risk for $M = 1$ splagging is exactly the same as that of subagging.

2. Under the linear model, the subsample conditional risk for $M = 2$ decomposes in the same form as (4.29), except in this case the datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ are independent of each other (conditional on I_1, I_2), which makes the analysis in this case slightly easier compared to that in subagging. By conditioning on each of the datasets successively, and utilizing the closed-form expression of the ridge estimator, we obtain the desired deterministic approximations.

3. Finally, as with the case of Theorem 4.4.1, we obtain results for the ridgeless predictor as the limiting risk approximations to the risk of the ridge predictor in the limit as $\lambda \to 0^+$ using uniformity arguments.

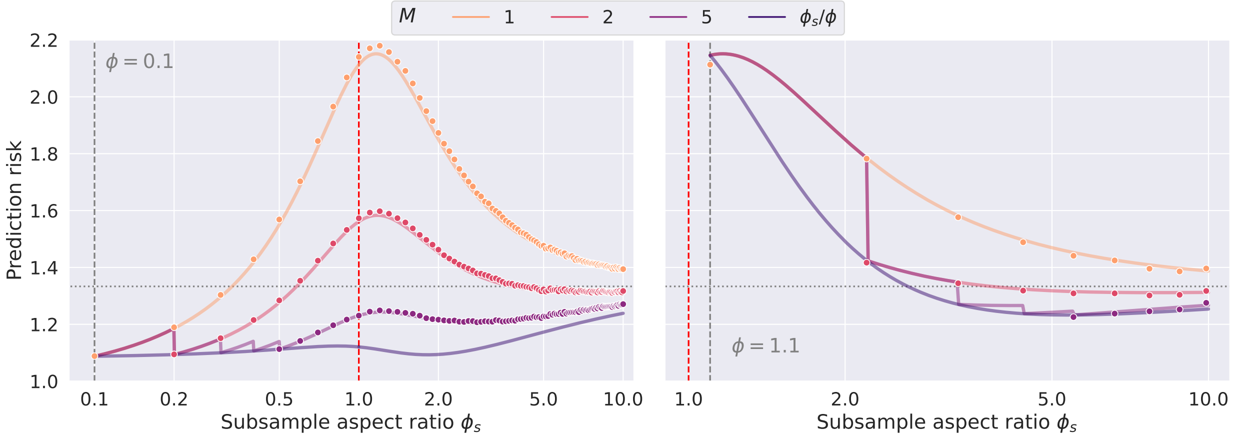


Figure 4.7: Asymptotic prediction risk curves in (4.32) for splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying split sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $p = 500$ and $n = \lfloor p\phi \rfloor$.

4.4.3.2 Monotonicity of bias and variance in number of bags

Similar to subagging, the asymptotic bias and variance of the conditional risk for splagging are also monotonically decreasing in the number of bags M, as shown in Proposition 4.4.10.

Proposition 4.4.10 (Improvement due to splagging)**.** *Fix any pair (ϕ, ϕ_s) such that $\phi_s \geq \phi$. Then for all $M \in \{1, \dots, \lfloor \phi_s/\phi \rfloor\}$,*

$$\mathscr{B}^{\texttt{spl}}_{\lambda,\lfloor \phi_s/\phi \rfloor}(\phi, \phi_s) \leq \mathscr{B}^{\texttt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s), \tag{4.33}$$

$$\mathscr{V}^{\texttt{spl}}_{\lambda,\lfloor \phi_s/\phi \rfloor}(\phi, \phi_s) \leq \mathscr{V}^{\texttt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s). \tag{4.34}$$

The inequalities in (4.33) are strict whenever $\rho^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$), while the inequalities in (4.34) are strict when $\sigma^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$). Thus, the asymptotic risk is monotonically decreasing in M: $\mathscr{R}^{\texttt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$.

Because the deterministic risk approximation for splagging is defined as a constant in M for $M \geq \lfloor \phi_s/\phi \rfloor$, Proposition 4.4.10 implies that the for every fixed pair (ϕ, ϕ_s), the optimal splagged predictor uses $M = \lfloor \phi_s/\phi \rfloor$ many bags.

4.5 Risk profile monotonization

The results presented in the previous sections provide risk characterizations for different versions of bagged predictors, per (4.4) and (4.5), for all possible subsample aspect ratios ϕ_s. In practice, the choice of ϕ_s is important to yield good prediction performance. Following the cross-validation strategy discussed in Patil et al. (2022a), one can apply cross-validation to choose the optimal ϕ_s in order to obtain the best possible prediction performance by subagging or splagging the base predictor across different subsample sizes. In Section 4.5.1, we first describe the risk monotonization results for general predictors, going back to the general setting in Section 4.3. In Section 4.5.2, we then specialize the general risk monotonization results to the bagged ridge and ridgeless predictors. In Section 4.5.3, we provide a comparison of the best subagged with the best splagged predictor among all possible choices of both ϕ_s and M, when the base predictor is either ridge or ridgeless.

4.5.1 Bagged general predictors

Several commonly used prediction procedures such as min-ℓ_2-norm least squares and ridge regression exhibit a non-monotonic risk behavior as a function of the data aspect ratio ϕ. This is referred to in the literature as double/multiple descent. The deterministic risk approximation as a function of the aspect ratio ϕ first increases with ϕ, reaches a peak, and then decreases with ϕ. Reinterpreting this phenomenon with a fixed dimension and changing sample size n reads as follows: the risk first decreases as the sample size increases up to some threshold, and then increases as the sample size increases. This is a counter-intuitive behavior from a statistical point of view as this indicates that more data might hurt. But theoretically, more information can only yield better performance. The underlying problem here is not the theory but the prediction procedure being used in that they are sub-optimal when applied as is on the full data.

There are at least two ways in which one can think of improving a given predictor:

1. Obtain a new predictor whose risk is the greatest monotone minorant of the risk of the given prediction procedure. This can be achieved by computing the predictor on a smaller sample size if necessary. Such a procedure was called the zero-step procedure (with $M = 1$) in Patil et al. (2022a); see Algorithm 4 below for details. It does the bare minimum to get monotone risk.

2. The zero-step procedure (with $M = 1$) is not a genuine improvement of the base predictor in that it is just the same predictor computed on a smaller dataset. From the positive effects of subagging or splagging mentioned in previous sections, we can improve on the zero-step procedure by aggregating over several subsets of the data. This was already eluded to and illustrated in Patil et al. (2022a). In this section, we discuss this point further.

We note from Theorem 4.4.1 and Figures 4.3 and 4.4 that for each ϕ, there are essentially infinitely many risk values possible (one for each pair of subsample aspect ratio ϕ_s and number of bags M). The zero-step procedure (with $M = 1$) improves on the base predictor by optimizing over ϕ_s, but fixing $M = 1$. Going one step forward, based on our results above, we can consider optimizing over ϕ_s and $M \geq 1$ (or just over ϕ_s, but fixing $M \geq 1$). In the following, we present an actionable algorithm to attain the optimum over ϕ_s for any fixed $M \geq 1$. (Note that we have already proved monotonicity over $M \geq 1$ and one can always choose M to be as large as feasible in practice.) Then, we present Theorem 4.5.1 where we prove that the general cross-validation attains the optimum over ϕ_s (asymptotically). Under the setting in Section 4.2, Theorem 4.5.1 provides theoretical guarantees of the cross-validation procedure for general base predictors, which extends the results of Patil et al. (2022a) to subagging and splagging.

Theorem 4.5.1 (Risk monotonization by cross-validation)**.** *Suppose that as $n, p \to \infty$, $p/n \to \phi \in (0, \infty)$. Let $\mathcal{K}_n$ be the set of subsample sizes defined in Algorithm 4 and $\mathcal{I}_k$ be the set of subsets of $\mathcal{S}_{\mathrm{tr}}$ of size $k \in \mathcal{K}_n$ according to the sampling scheme. Suppose for any $k \in \mathcal{K}_n$, as $n, k, p \to \infty$, and $p/k \to \phi_s \in [\phi, \infty)$, there exists a deterministic function $\mathscr{R} : (0, \infty]^2 \to [0, \infty]$ such that*

(i) For any $I \in \mathcal{I}_k$ and $\{I_{k,1}, I_{k,2}\}$ a simple random sample from $\mathcal{I}_k$,

$$R(\widetilde{f}_1; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} \mathscr{R}(\phi_s, \phi_s), \qquad R(\widetilde{f}_2; \mathcal{D}_n, \{I_{k,1}, I_{k,2}\}) \xrightarrow{\text{a.s.}} \mathscr{R}(\phi, \phi_s).$$

(ii) For any $\phi \in (0, \infty)$, $\phi_s \mapsto \mathscr{R}(\phi, \phi_s)$ is proper and lower semi-continuous over $[\phi, \infty]$, and is continuous on the set $\arg\min_{\{\psi:\psi \geq \phi\}} \mathscr{R}(\phi, \psi)$.

Let $\widehat{f}_M^{\mathrm{cv}}$ be the cross-validated predictor returned by Algorithm 4 with base predictor $\widehat{f}$. If the estimated risk $\widehat{R}(\widetilde{f}_{M,k})$ defined in (4.35) or (4.36) is uniformly (in $k \in \mathcal{K}_n$) close to the subsample conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)$ with probability converging to 1, then the following conclusions hold. For subagging with or without replacement, or splagging without replacement, it holds for all $M \in \mathbb{N}$ that,

$$\left(R(\widehat{f}_M^{\mathrm{cv}}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s \geq \phi} \mathscr{R}_M(\phi, \phi_s) \right)_+ \xrightarrow{\text{p}} 0,$$

Algorithm 4 Cross-validation for subagging or splagging

```
Input: A dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le n\}$, a positive integer $n_{\mathrm{te}} < n$ (number of test samples), a base prediction procedure $\widehat{f}$, a real number $\nu \in (0,1)$ (bag size unit parameter), a natural number $M$ (number of bags), a centering procedure $\texttt{CEN} \in \{\texttt{AVG}, \texttt{MOM}\}$, a real number $\eta$ when $\texttt{CEN} = \texttt{MOM}$.
Data splitting: Randomly split $\mathcal{D}_n$ into training set $\mathcal{D}_{\mathrm{tr}}$ and test set $\mathcal{D}_{\mathrm{te}}$ as:
    $\mathcal{D}_{\mathrm{tr}} = \{(\boldsymbol{x}_i, y_i) : i \in \mathcal{S}_{\mathrm{tr}}\}, \qquad \mathcal{D}_{\mathrm{te}} = \{(\boldsymbol{x}_j, y_j) : j \in \mathcal{S}_{\mathrm{te}}\},$
    where $\mathcal{S}_{\mathrm{te}} \subset [n]$ with $|\mathcal{S}_{\mathrm{te}}| = n_{\mathrm{te}}$ and $\mathcal{S}_{\mathrm{tr}} = [n] \setminus \mathcal{S}_{\mathrm{te}}$.
Bag sample sizes grid construction: Let $k_0 = \lfloor n^{\nu} \rfloor$ and $\mathcal{K}_n = \{k_0, 2k_0, \dots, \lfloor n/k_0 \rfloor k_0\}$.
Subagging or splagging predictors: For each $k \in \mathcal{K}_n$, define $\widetilde{f}_{M,k}$ trained on $\mathcal{D}_{\mathrm{tr}}$ as:
    • For subagging, let $\widetilde{f}_{M,k}(\cdot) = \widetilde{f}_M(\cdot; \{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^M)$ denote the subagged predictor as in (4.4) with $M$ bags. Here, $I_{k,1}, \dots, I_{k,M}$ represent a simple random sample with or without replacement from the set of all subsets of $\mathcal{S}_{\mathrm{tr}}$ of size $k$.
    • For splagging, $\widetilde{f}_{M,k}(\cdot)$ is the same as above but now $I_{k,1}, \dots, I_{k,M}$ represent a simple random sample without replacement from a random split of $\mathcal{S}_{\mathrm{tr}}$ into $\lfloor n/k \rfloor$ parts with each part containing $k$ elements. As explained in Section 4.4.1, for $M > \lfloor n/k \rfloor$, no such splitting exists. In this case, we return $\widetilde{f}_{\lfloor n/k \rfloor, k}$. Hence in general, we have $\widetilde{f}_{M,k} = \widetilde{f}_{\min\{M, \lfloor n/k \rfloor\}, k}$.
Risk estimation: For each $k \in \mathcal{K}_n$, estimate the conditional prediction risk on $\mathcal{D}_{\mathrm{te}}$ of $\widetilde{f}_{M,k}$ as:
    $\widehat{R}(\widetilde{f}_{M,k}) := \begin{cases} |\mathcal{S}_{\mathrm{te}}|^{-1} \sum_{j \in \mathcal{S}_{\mathrm{te}}} (y_j - \widetilde{f}_{M,k}(\boldsymbol{x}_j))^2, & \text{if } \texttt{CEN=AVG} \quad (4.35) \\ \mathrm{median}(\widehat{R}_1(\widetilde{f}_{M,k}), \dots, \widehat{R}_B(\widetilde{f}_{M,k})), & \text{if } \texttt{CEN=MOM}, \quad (4.36) \end{cases}$
    where $B = \lceil 8 \log(1/\eta) \rceil$, and $\widehat{R}_j(\widetilde{f}_{M,k})$, $1 \le j \le B$ is defined similarly to (4.35) for $B$ random splits of the test dataset $\mathcal{D}_{\mathrm{te}}$; see Patil et al. (2022a) for more details.
Cross-validation: Set $\widehat{k} \in \mathcal{K}_n$ to be the bagging sample size that minimizes the estimated prediction risk using
    $\widehat{k} \in \mathop{\mathrm{arg\,min}}_{k \in \mathcal{K}_n} \widehat{R}(\widetilde{f}_{M,k}). \quad (4.37)$
Output: Return the predictor $\widehat{f}^{\mathrm{cv}}_M(\cdot; \mathcal{D}_n) = \widetilde{f}_{M,\widehat{k}}(\cdot) = \widetilde{f}_M(\cdot; \{\mathcal{D}_{I_{\widehat{k},\ell}}\}_{\ell=1}^M)$.
```

where the function $\mathscr{R}_M(\phi, \phi_s)$ is defined as

$$\mathscr{R}_M(\phi, \phi_s) := (2\mathscr{R}(\phi, \phi_s) - \mathscr{R}(\phi_s, \phi_s)) + \frac{2}{M}(\mathscr{R}(\phi_s, \phi_s) - \mathscr{R}(\phi, \phi_s)).$$

Furthermore, if for any $\phi_s \in (0, \infty)$, $\phi \mapsto \mathscr{R}(\phi, \phi_s)$ is non-decreasing over $(0, \phi_s]$, then the function $\phi \mapsto \min_{\phi_s \ge \phi} \mathscr{R}_M(\phi, \phi_s)$ is monotonically increasing for every M.

Remark 4.5.2 (Asymptotic risks are different for subagging and splagging.)**.** Although Theorem 4.5.1 is presented in a unified manner for subagging and splagging, the actual limiting risks can be (and in most cases are) different. This difference arises in the different expressions for the asymptotic risks assumed in assumption (i) of Theorem 4.5.1.

Remark 4.5.3 (Exact risk characterization of the cross-validated predictor with stronger assumptions)**.** Note that Theorem 4.5.1 does not exactly characterize the risk of cross-validated bagged predictor; it only states that the subsample conditional risk of $\widetilde{f}^{\mathrm{cv}}_M$ is asymptotically no larger than $\min_{\phi_s} \mathscr{R}_M(\phi, \phi_s)$.

Nevertheless, this is an improvement over the results of Patil et al. (2022a), who proved that the subsample conditional risk of $\widetilde{f}_M^{\mathrm{cv}}$ is asymptotically no larger than $\min_{\phi_s} \mathscr{R}_1(\phi, \phi_s)$. For the exact risk characterization of $\widetilde{f}_M^{\mathrm{cv}}$, one can make the stronger assumption that as $n, p \to \infty$ and $p/n \to \phi$,

$$\sup_{k \le n} |R(\widetilde{f}_1; \mathcal{D}_n, \{I_1 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k\}) - \mathscr{R}(p/k, p/k)| \xrightarrow{\mathrm{p}} 0, \qquad \sup_{k \le n} |R(\widetilde{f}_2; \mathcal{D}_n, \{I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k\}) - \mathscr{R}(\phi, p/k)| \xrightarrow{\mathrm{p}} 0,$$

which can be used to conclude

$$R(\widehat{f}_M^{\mathrm{cv}}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) \xrightarrow{\mathrm{p}} \min_{\phi_s \ge \phi} \mathscr{R}_M(\phi, \phi_s).$$

The result for bagging without replacement can be extended analogously.

Remark 4.5.4 (Assumption of uniform consistency of the estimated risk)**.** The assumption of uniform (in $k \in \mathcal{K}_n$) closeness of the estimated risk $\widehat{R}(\widetilde{f}_{M,k})$ to the subsample conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)$ is meant to represent either

$$\max_{k \in \mathcal{K}_n} |\widehat{R}(\widetilde{f}_{M,k}) - R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)| = o_p(1), \quad \text{or} \quad \max_{k \in \mathcal{K}_n} \left| \frac{\widehat{R}(\widetilde{f}_{M,k})}{R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)} - 1 \right| = o_p(1).$$

In Section 2 of Patil et al. (2022a), the authors have provided several assumptions on the data distribution and the predictors such that this uniform closeness assumption holds true. In Section 4.5.2, we will apply Theorem 4.5.1 for bagged linear predictors which are themselves linear predictors. In this specific case, Theorem 2.22 in the aforementioned work shows that uniform closeness holds true under assumptions on the data distribution alone (no matter what linear predictor is, even those that have diverging risks); see Patil et al. (2022a, Remarks 2.19 and 2.20). We do not further discuss this uniform closeness condition, but only remark that Assumptions 4.1-4.5 imply the assumptions of Theorem 2.22 with $\texttt{CEN} = \texttt{MOM}$ (the median-of-means estimator). With $\texttt{CEN} = \texttt{AVG}$, sub-Gaussian features imply the assumptions of Theorem 2.22.

4.5.2 Bagged ridge and ridgeless predictors

Theorem 4.5.1 provides a very general result that describes the risk behavior of cross-validated bagged predictors in general. Following our results in previous sections that verify condition (i) of Theorem 4.5.1 for both ridge and ridgeless predictors, we now specialize Theorem 4.5.1 to these predictors under Assumptions 4.1-4.5.

Theorem 4.5.5 (Risk monotonicity in aspect ratio)**.** *Suppose that the cross-validated predictor $\widehat{f}_M^{\mathrm{cv}}$ is returned by Algorithm 4 with base predictor $\widehat{f}_\lambda$ and M bags, and the conditions in Theorem 4.4.1 (or Theorem 4.4.6) hold*[7] *with $\mathcal{R}_{\lambda,M}(\phi, \phi_s)$ being the limiting risk $\mathscr{R}_{\lambda,M}^{\texttt{sub}}(\phi, \phi_s)$ (or $\mathscr{R}_{\lambda,M}^{\texttt{spl}}(\phi, \phi_s)$). Then it holds for all $M \in \mathbb{N}$,*

$$\left(R(\widehat{f}_M^{\mathrm{cv}}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s \ge \phi} \mathcal{R}_{\lambda,M}(\phi, \phi_s) \right)_+ \xrightarrow{\mathrm{p}} 0. \tag{4.38}$$

Furthermore, $\phi \mapsto \min_{\phi_s \ge \phi} \mathcal{R}_{\lambda,M}(\phi, \phi_s)$ is a monotonically increasing function of ϕ for every M.

In Theorem 4.5.5, the monotonicity of $\phi \mapsto \min_{\phi_s \ge \phi} \mathcal{R}_{\lambda,M}$ implies that for every M, for the optimal bagged predictor, more data (i.e, increasing n) cannot hurt. In the plot of Figure 4.8, we observe slight non-monotonicity of the empirical risk profile for $M = 1$. This is because of the small sample size which does not allow for the optimal cross-validated predictor to be the null predictor. One way to not let this happen (in this specific case) is to always include a perfect "null" predictor in the set of predictors tuned with cross-validation in Algorithm 4.

[7] The statement as stated holds for $\texttt{CEN} = \texttt{MOM}$ in Algorithm 4. For $\texttt{CEN} = \texttt{AVG}$, we need to assume sub-Gaussian features as discussed in Remark 4.5.4.

For splagging without replacement, the simulation results are shown in Figure 4.8(b). As expected, as the limiting aspect ratio ϕ increases, the empirical excess risks are nearly monotone increasing and match with theoretical curves. Another pattern we observe in Figure 4.8 (splagging without replacement) is that the asymptotic risk may not be monotonically decreasing in M when ϕ is small. This is because the subsample aspect ratio ϕ_s is restricted by the number of bags M in that it cannot be below $M\phi$, and the differences in the range of ϕ_s when using different numbers of bags result in the non-monotonicity when ϕ is small. While in the overparameterized region when ϕ is large enough, the cross-validated risk for bagging without replacement is guaranteed to be monotonically decreasing in M. Furthermore, the choice of $M = \phi_s/\phi$ guarantees that the risk is always optimal compared to any other value of M.

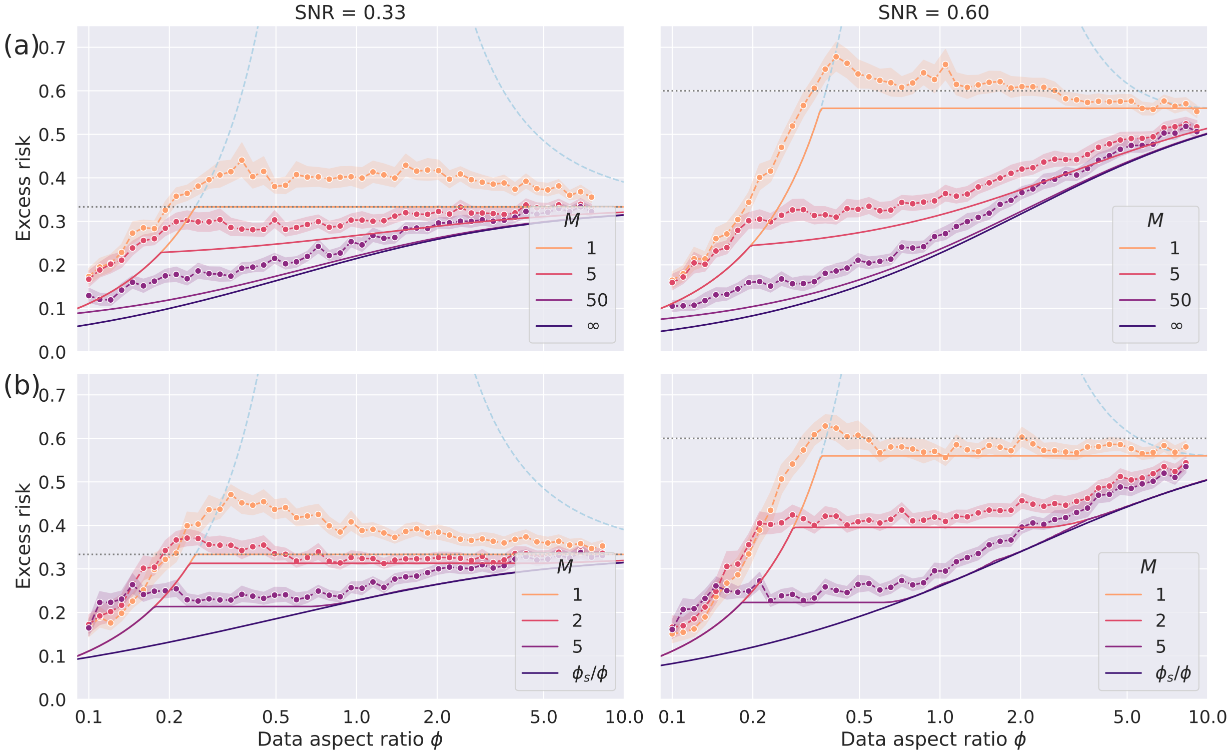


Figure 4.8: Asymptotic excess risk curves for cross-validated bagged ridgeless predictors ($\lambda = 0$) for (a) subagging and (b) splagging, under model (M-AR1-LI) when $\sigma^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $\texttt{SNR} = 0.33$ ($\rho_{\text{ar1}} = 0.25$) and 0.6 ($\rho_{\text{ar1}} = 0.5$), respectively. The excess null risks and the risks for the unbagged ridgeless predictors are marked as dotted lines and the dashed lines, respectively. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

4.5.3 Optimal subagging versus optimal splagging

The cross-validated predictors discussed previously yield asymptotic optimal risks over subsample aspect ratio ϕ_s for every M. As a step further, we can obtain the optimal subagging and optimal splagging by jointly optimizing over both ϕ_s and M. From the explicit formulas for the limiting risks for each pair of aspect ratios (ϕ, ϕ_s) and each M, the optimal bagged risks in the two cases can be compared.

Proposition 4.5.6 (Comparison of the optimal risk of subagging and splagging)**.** *Under Assumptions 4.3-4.5, let $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$ be defined as in Theorem 4.4.1 and Theorem 4.4.6, respectively.*

Then for any $\lambda \in [0,\infty)$ and $\phi \in (0,\infty)$, the following holds:

$$\inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]} \mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) \le \inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]} \mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s). \tag{4.39}$$

In words, optimal subagging is at least as good as optimal splagging (without replacement) in terms of squared loss for ridge predictors.

For any dataset with fixed aspect ratio ϕ, Proposition 4.5.6 indicates that the optimal risk for bagged predictor across all possible choices of M and subsample aspect ratio ϕ_s is always given by subagging. The optimal subagging and optimal splagging risks in Proposition 4.5.6 can be written as

$$\mathcal{R}^{\mathtt{sub}}_{\mathtt{opt}}(\phi) = \mathscr{R}^{\mathtt{sub}}_{\lambda,\infty}(\phi,\phi^{\mathtt{sub}}_s(\phi)), \quad \text{and} \quad \mathcal{R}^{\mathtt{spl}}_{\mathtt{opt}}(\phi) = \mathscr{R}^{\mathtt{spl}}_{\lambda,\phi^{\mathtt{spl}}_s(\phi)/\phi}(\phi,\phi^{\mathtt{spl}}_s(\phi)), \tag{4.40}$$

where the functions $\phi \mapsto \phi^{\mathtt{sub}}_s(\phi)$ and $\phi \mapsto \phi^{\mathtt{spl}}_s(\phi)$ are defined via

$$\phi^{\mathtt{sub}}_s(\phi) \ := \ \mathop{\arg\min}_{\phi_s\ge\phi} \mathscr{R}^{\mathtt{sub}}_{\lambda,\infty}(\phi,\phi_s), \quad \text{and} \quad \phi^{\mathtt{spl}}_s(\phi) \ := \ \mathop{\arg\min}_{\phi_s\ge\phi} \mathscr{R}^{\mathtt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s). \tag{4.41}$$

The fact that the optimal risks shown in Proposition 4.5.6 are the same as shown in (4.40) follows from the fact that the risks are monotonically decreasing in M for subagging and that the risk at $M = \phi_s/\phi$ is the best for splagging without replacement for any pair (ϕ,ϕ_s). The quantities $\phi^{\mathtt{sub}}_s(\cdot)$ and $\phi^{\mathtt{spl}}_s(\cdot)$ represent the best possible subsample aspect ratios for subagging and splagging (without replacement) for every data aspect ratio ϕ given. (Minimizers of lower semi-continuous functions over compact domains exist, which is true for the functions in (4.41) from Theorem 4.5.5.)

We calculate and present the theoretical optimal asymptotic risks (4.40) for bagged ridgeless predictors in Figure 4.9. The optimal risk $\min_{\phi_s\ge\phi}\mathscr{R}^{\mathtt{sub}}_{\lambda,1}(\phi,\phi_s) = \min_{\phi_s\ge\phi}\mathscr{R}^{\mathtt{spl}}_{\lambda,1}(\phi,\phi_s)$ of the bagged ridgeless predictor with $M = 1$ is also presented as the dashed line, which is the same as the monotone risk of the zero-step ridgeless predictor of Patil et al. (2022a) with $M = 1$. As shown in Figure 4.9(a), the optimal risk for the subagged ridgeless predictor is always smaller than the splagged ridgeless predictor without replacement. Both of them improve the risk for the ridgeless predictor with optimal subsample aspect ratio ϕ_s using only one bag ($M = 1$).

Oracle properties of optimal subsample aspect ratios. From the previous section, we see that optimal subagged ridge or ridgeless regression always outperforms the splagged one in terms of limiting risk. Due to the monotonicity in the number of bags M from Proposition 4.4.5, the optimal risk for subagging must be obtained at $M = \infty$ for any given subsample aspect ratio ϕ_s. One question that arises is: what is the optimal subsample aspect ratio ϕ_s? We provide a partial answer to this question in Proposition 4.5.7 specialized to ridgeless regression.

Proposition 4.5.7 (Optimal risk for bagged ridgeless predictor)**.** *Suppose the conditions in Theorems 4.4.1 and 4.4.6 hold, and $\sigma^2, \rho^2 \ge 0$ are the noise variance and signal strength from Assumptions 4.2 and 4.3. Let $\mathtt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0,\infty)$, the properties of the optimal asymptotic risks $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi^{\mathtt{sub}}_s(\phi))$ and $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi^{\mathtt{spl}}_s(\phi))$ in terms of $\mathtt{SNR}$ and ϕ are characterized as follows:*

(1) $\mathtt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \ne 0)$*: For all $\phi \ge 0$, the global minimum σ^2 of both $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi^{\mathtt{sub}}_s(\phi))$ and $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi^{\mathtt{spl}}_s(\phi))$ are obtained with $\phi^{\mathtt{sub}}_s(\phi) = \phi^{\mathtt{spl}}_s(\phi) = \infty$.*

(2) $\mathtt{SNR} > 0$*: For all $\phi \ge 0$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s)$ is obtained at $\phi^{\mathtt{sub}}_s(\phi) \in (1,\infty)$. For $\phi \ge 1$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ is obtained at $\phi^{\mathtt{spl}}_s(\phi) \in (1,\infty)$; for $\phi \in (0,1)$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ is obtained at $\phi^{\mathtt{spl}}_s(\phi) \in \{\phi\} \cup (1,\infty)$.*

(3) $\mathtt{SNR} = \infty$ $(\rho^2 \ne 0, \sigma^2 = 0)$*: If $\phi \in (0,1]$, the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi^{\mathtt{sub}}_s(\phi)) = \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi^{\mathtt{spl}}_s(\phi)) = 0$ is obtained with any $\phi^{\mathtt{sub}}_s(\phi), \phi^{\mathtt{spl}}_s(\phi) \in [\phi,1]$. If $\phi \in (1,\infty)$, then the global minimums $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi^{\mathtt{sub}}_s(\phi))$ and $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi^{\mathtt{spl}}_s(\phi))$ are obtained at $\phi^{\mathtt{sub}}_s(\phi), \phi^{\mathtt{spl}}_s(\phi) \in [\phi,\infty)$.*

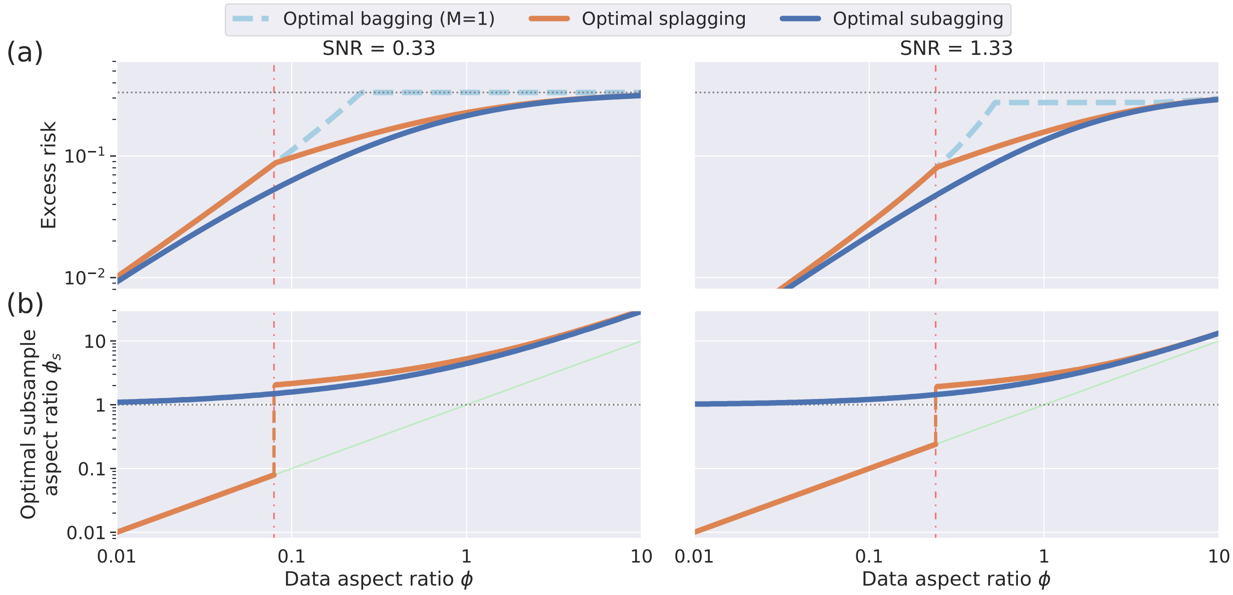


Figure 4.9: Comparison between optimal subagging and optimal splagging of ridgeless predictors ($\lambda = 0$) for varying limiting aspect ratios ϕ of p/n under model (M-AR1-LI) when $\sigma^2 = 1$. The left and right panels correspond to $\texttt{SNR} = 0.33$ ($\rho_{\rm ar1} = 0.25$) and $\texttt{SNR} = 0.6$ ($\rho_{\rm ar1} = 0.5$), respectively. The point of phase transition for splagging is marked as the red dash-dot line in every subplot. (a) Optimal asymptotic excess risk curves (4.39). The excess null risks are marked as gray dotted lines and the blue dashed lines represent the optimal risks of bagged ridgeless predictor with $M = 1$, which are the same as the risks from the zero-step procedure of Patil et al. (2022a). (b) The corresponding optimal subsample aspect ratio ϕ_s as a function of data aspect ratio ϕ. For subagging, the optimal subsample aspect ratio is always larger than one (above the gray dotted line). The line $\phi_s = \phi$ is colored in green.

Proposition 4.5.7 implies that the optimal subsample aspect ratio $\phi_s^{\texttt{sub}}(\phi)$ for subagging is always in $[1, \infty]$, i.e., the overparameterized regime. In other words, subagging interpolators with larger aspect ratios (larger than the full data aspect ratio ϕ) helps to reduce the prediction risk, even when $\phi < 1$. For splagging, however, the minimum risk can be obtained either using the full data or splagging interpolators, depending on the data aspect ratio ϕ and the signal-to-noise ratio.

It is interesting to note that the optimal subsampling aspect ratio for splagging is either ϕ or it is in the overparameterized regime $(1, \infty)$. This means that either splagging does not help, or when it helps, one has to splag interpolators. Whenever $\texttt{SNR}$ is positive, the optimal subsample aspect ratio is finite for any ϕ. Hence we are able to visualize $\phi_s^{\texttt{sub}}(\phi)$ and $\phi_s^{\texttt{spl}}(\phi)$ in Figure 4.9(b). As shown in Figure 4.9, there is a point of non-differentiability of $\phi_s^{\texttt{spl}}(\phi)$ for optimal splagging without replacement. Before this point of non-differentiability, $\phi_s^{\texttt{spl}}(\phi) = \phi$, which is the same as the optimal bagged ridgeless with $M = 1$. This is also the same as the ridgeless predictor trained on the full data. After the point of non-differentiability, the optimal risk for splagging without replacement is obtained in the overparameterized regime, i.e., $\phi_s^{\texttt{spl}}(\phi) > 1$. In contrast to splagging, $\phi_s^{\texttt{sub}}(\phi) \geq 1$ for all $\phi > 0$, meaning that it is always better to subag interpolators (i.e., the overparameterized regime).

These observations indicate that, when the number of bags is sufficiently large enough, splagging without replacement only helps when the limiting aspect ratio ϕ of the full dataset is above some threshold, but subagging is always beneficial in reducing the prediction risk, even in the underparameterized regime.

Remark 4.5.8 (Guidelines for practical data analysis)**.** Proposition 4.5.7 implies that when using $M = \infty$, one should consider bagging interpolators to get better predictive performance, at least when the linear model holds true. However, $M = \infty$ is practically infeasible particularly when $n, k \to \infty$. Note from Figures 4.3 and 4.4 that for M large enough, the same phenomenon holds true, i.e., it is better to bag interpolators with a large M. How large such an M should be depends on various unknowns related to the linear model and also on how much gap $\delta > 0$ from $M = \infty$ one is willing to allow. Given the form of the limiting risk as a function of M, we can figure out the necessary value of M as a function of the gap δ, based on the cross-validation procedure (Algorithm 4). Note that this is completely data-driven and model-agnostic. The procedure is as follows: (1) Run Algorithm 4 with $M = 1$ and $M = 2$ to obtain the estimators $\widehat{\mathscr{R}}_1(\phi_s, \phi)$ and $\widehat{\mathscr{R}}_2(\phi_s, \phi)$ of the limiting subsample conditional risks $M = 1, 2$, respectively, for a grid of values $\phi_s \geq \phi$. Following Proposition 4.3.3, this yields an estimator of the subsample conditional risk for every $M \geq 1$, in particular, for $M = \infty$. (2) Find $\widehat{\phi_s^{\mathtt{sub}}}(\phi)$, the minimizer of $\phi_s \mapsto 2\widehat{\mathscr{R}}_2(\phi_s, \phi) - \widehat{\mathscr{R}}_1(\phi_s, \phi)$. Note that this map is an estimator of the limiting risk for $M = \infty$. (3) Fix a tolerance level $\delta > 0$, and choose

$$M = \frac{2}{\delta}\left\{\widehat{\mathscr{R}}_1(\widehat{\phi_s^{\mathtt{sub}}}(\phi), \phi) - \widehat{\mathscr{R}}_2(\widehat{\phi_s^{\mathtt{sub}}}(\phi), \phi)\right\}.$$

Operating at $\widehat{\phi_s^{\mathtt{sub}}}(\phi)$ with such a value of M will yield an asymptotic risk that is δ close (in the additive sense) to the optimal risk.

4.6 Illustrations and insights: isotropic features

All the results presented before are derived under Assumptions 4.1-4.5. We now consider a much simpler case of isotropic features (i.e., $\mathbf{\Sigma} = \boldsymbol{I}_p$ in Assumption 4.1). In this case, the spectral distribution simplifies and allows us to compute the fixed point solutions analytically. We will primarily focus on the case of ridgeless predictor for the sake of illustration. It is possible to obtain similar results for ridge predictors, albeit slightly more involved. In Appendix D.7.3, we provide formulas for the fixed-point solutions for $\lambda > 0$ from which one can derive the risk as well as the individual bias and variance numerically for ridge predictors (with arbitrary $\lambda > 0$). In general, these quantities can always be computed numerically for nonisotropic models.

For isotropic features, the bias and variance functions in Theorems 4.4.1 and 4.4.6 admit relatively simple forms, as shown in Corollary 4.6.1. The asymptotic bias and variance can be further computed for all $M \in \mathbb{N}$ from (4.28).

Corollary 4.6.1 (Bias-variance components for isotropic design)**.** *Assume the conditions in Theorem 4.4.1 or Theorem 4.4.6 hold with $\mathbf{\Sigma} = \boldsymbol{I}_p$. Then we have*

$$B_0(\phi, \phi_s) = \rho^2 \frac{(\phi_s - 1)^2}{\phi_s^2 - \phi} \mathbb{1}_{(1,\infty]}(\phi_s), \qquad C(\phi_s) = \rho^2 \frac{(\phi_s - 1)^2}{\phi_s^2} \mathbb{1}_{(1,\infty]}(\phi_s),$$

$$V_0(\phi, \phi_s) = \begin{cases} \sigma^2 \dfrac{\phi}{1 - \phi}, & \phi_s \in (0, 1) \\ \infty, & \phi_s = 1 \\ \sigma^2 \dfrac{\phi}{\phi_s^2 - \phi}, & \phi_s \in (1, \infty]. \end{cases}$$

Subagging with replacement. From Corollary 4.6.1, we are able to evaluate the closed-form asymptotic risk under model (M-ISO-LI):

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p), \quad \boldsymbol{\beta}_0 \sim \mathcal{N}(0, p^{-1}\rho^2 \boldsymbol{I}_p), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{M-ISO-LI}$$

Extra experimental results under model (M-ISO-LI) are also included in Appendix D.10. We note that the Gaussianity of the noise ϵ_i in model (M-ISO-LI) is convenient for numerical evaluation, while it is not needed for Corollary 4.6.1. Instead, we only need the first and second moments to match as above. For $M \in \mathbb{N}$, the bias term is always increasing, while the variance term will blow up when the subsample aspect ratio ϕ_s approaches one. However, the variance for $M = \infty$ is different. It is decreasing in ϕ_s and continuous at $\phi_s = 1$. As a result, one might be interested in the optimal subsample aspect ratio $\phi_s^{\mathtt{sub}}(\phi)$, that best trades off the bias and variance, and minimizes the risk for a given value of ϕ and $M = \infty$.

Proposition 4.6.2 (Optimal risk for subagged ridgeless predictors with isotropic features)**.** *Suppose the conditions in Corollary 4.6.1 hold, and $\sigma^2, \rho^2 \geq 0$ are the noise variance and signal strength from*

Assumptions 4.2 and 4.3. Let $\mathtt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0, \infty)$, the properties of the asymptotic risk $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)$ as a function of ϕ_s are characterized as follows:

(1) $\mathtt{SNR} = 0$ ($\rho^2 = 0, \sigma^2 \neq 0$): The global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi)) = \sigma^2$ is obtained at $\phi^{\mathtt{sub}}_s(\phi) = \infty$.

(2) $\mathtt{SNR} > 0$: The global minimum

$$\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi)) = \frac{\sigma^2}{2}\left[1 + \frac{\phi - 1}{\phi}\mathtt{SNR} + \sqrt{\left(1 - \frac{\phi - 1}{\phi}\mathtt{SNR}\right)^2 + 4\mathtt{SNR}}\right] \tag{4.42}$$

is obtained at $\phi^{\mathtt{sub}}_s(\phi) = A + \sqrt{A^2 - \phi} \in (1, \infty)$ where $A = (\phi + 1 + \phi/\mathtt{SNR})/2$.

(3) $\mathtt{SNR} = \infty$ ($\rho^2 \neq 0, \sigma^2 = 0$): If $\phi \in (0, 1]$, then the global minimum is $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi)) = 0$ is attained at any $\phi_s \in [\phi, 1]$. If $\phi \in (1, \infty)$, then the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi)) = \sigma^2 + \rho^2(\phi - 1)/\phi$ is attained at $\phi^{\mathtt{sub}}_s(\phi) = \phi$.

As a specialization of Proposition 4.5.7, Proposition 4.6.2 provides the analytic expression of the optimal risk one could achieve by optimizing over all choices of the number of bags M and the subsample aspect ratio ϕ_s. Furthermore, it also reveals the relationship between the optimal risk and the $\mathtt{SNR}$, which is also visualized in Figure 4.10. Specifically, the optimal subagged risk is monotonically decreasing in $\mathtt{SNR}$ when σ^2 is fixed, which is an intuitive behavior as one would expect a larger $\mathtt{SNR}$ yields a smaller prediction risk. In contrast, such a property is not satisfied by the ridge or ridgeless predictor computed on the full data (Hastie et al., 2022, Figure 2). It can be shown that the gap between the optimal risk given in Proposition 4.6.2 and the underparameterized excess risk $\sigma^2\phi/(1 - \phi)$ obtained with the full dataset gets larger, when $\mathtt{SNR}$ gets smaller. Most importantly, it benefits more when the $\mathtt{SNR}$ gets smaller, with higher overparameterized aspect ratio $\phi^{\mathtt{sub}}_s(\phi)$.

Theorem 4.6.3 (Optimal subagged ridgeless risk versus optimal ridge risk)**.** *Under the conditions in Corollary 4.6.1, we have that for all $\phi \in (0, \infty)$,*

$$\min_{\phi_s \geq \phi} \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) \;\; = \;\; \min_{\lambda \geq 0} \mathscr{R}^{\mathtt{sub}}_{\lambda,1}(\phi, \phi).$$

In words, the optimal limiting risk of the subagged ridgeless predictors equals the optimal ridge predictors trained on the full data.

Theorem 4.6.3 is an unexpected connection between subagging and ridge regression. This result can also be interpreted as saying that subagging a ridge predictor with $\lambda = 0$ and optimizing over the subsample size is "same" as using the ridge predictor with $\lambda \geq 0$ and optimizing over λ. Consequently, this suggests that subsampling and optimizing over subsample size is a form of regularization. A similar connection between subsampling features and ridge regression was made by LeJeune et al. (2020, Theorem 3.6).

Remark 4.6.4 (Difference between optimal subagged ridgeless and optimal ridge predictors)**.** Though the two optimal limiting risks match under the isotopic model as suggested by Theorem 4.6.3, the risk monotonicity properties of them in the data aspect ratio ϕ are different. Further, the optimal risk of the subagged ridgeless predictor is expected to remain monotonically decreasing in ϕ from Theorem 4.5.5, while whether the optimal ridge predictor has the same property is unknown under general models.

Splagging without replacement. Unlike subagging, it is possible, though very cumbersome to obtain the optimal sub-sampling ratio $\phi^{\mathtt{spl}}_s(\phi)$ in this case. It involves solving a cubic equation (for a fixed M) or a quartic equation (for the optimal M). To this end, we compute $\phi^{\star}_s$ numerically and provide a qualitative behavior for ϕ_s as summarized below. As $\mathtt{SNR}$ increases the point of phase transition occurs at a larger value of ϕ. This indicates that when there are much more features than samples in the full dataset and the $\mathtt{SNR}$ is relatively large, then splagging does not help to reduce the prediction risk. However, when the $\mathtt{SNR}$ is small, splagging interpolators is beneficial, even when n is much larger than p in the full data.

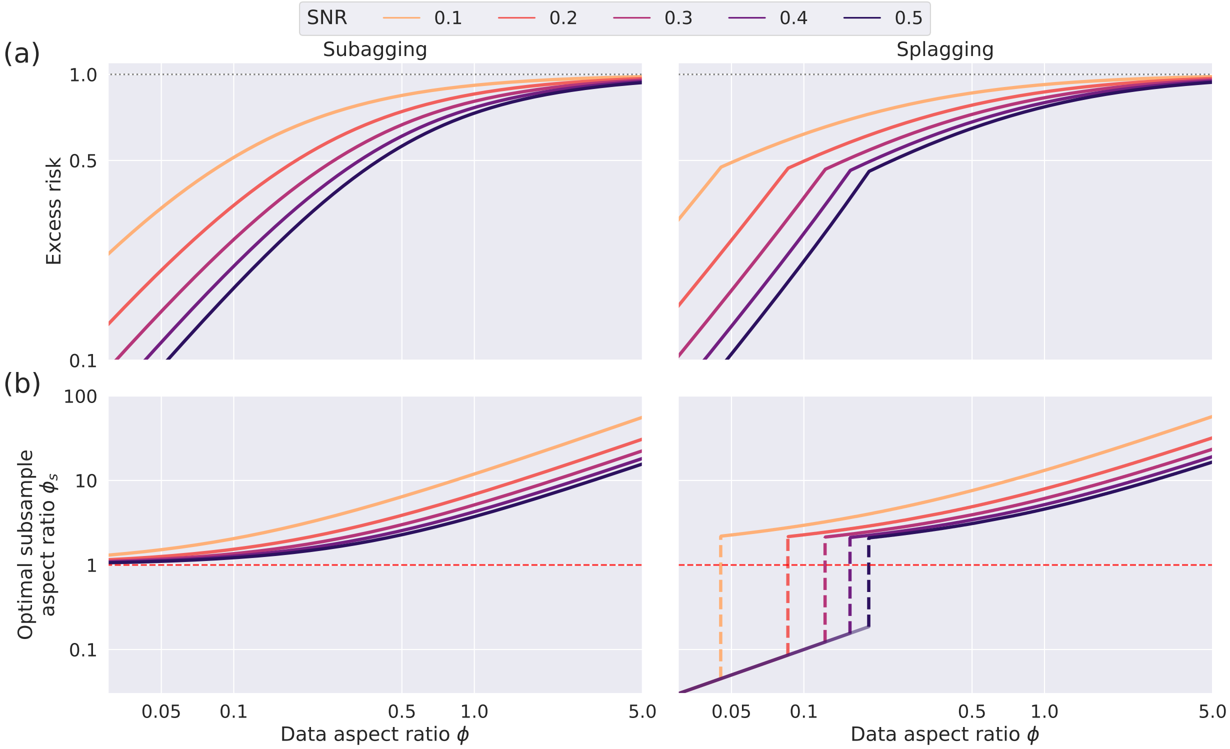


Figure 4.10: Properties of optimal bagged ridgeless predictors ($\lambda = 0$) under model (M-ISO-LI) when $\rho^2 = 1$, for varying signal noise ratio ($\mathtt{SNR} = \rho^2/\sigma^2$). (a) Optimal asymptotic excess risk curves of subagging (left panel) and splagging (right panel) over the number of bags M and subsample aspect ratio ϕ_s. The optimal numbers of bags are $M = \infty$ and $M = \phi_s/\phi$ for subagging and splagging, respectively. The gray dotted lines represent the excess null risk. (b) The corresponding optimal subsample aspect ratio ϕ_s as a function of data aspect ratio ϕ. For subagging, the optimal subsample aspect ratio is always larger than one (above the red dash line).

Subagging versus splagging. From the previous sections, we have observed the interesting phenomena of the prediction risks for subagging and splagging. Here we briefly summarize these findings concerning the similarities and differences between the two types of bagging strategies for ridgeless predictors.

- As revealed in Figure 4.10, for any data aspect ratio ϕ and any $\mathtt{SNR}$, subagging can help to reduce the risk with a suitable subsample aspect ratio in the overparameterized regime, if we have enough bags. In contrast, splagging may not help when $\phi < 1$ and $\mathtt{SNR}$ is large, even if we optimize over all possible numbers of bags and subsample aspect ratios jointly.

- For the cases when subagging or splagging is beneficial, the maximal gain compared to the predictor computed on the full data increases as the $\mathtt{SNR}$ decreases. When the full data aspect ratio ϕ is near to 1, both subagging and splagging substantially reduce the prediction risk; see Figures 4.3, 4.4, 4.6 and 4.7.

- Most surprisingly, even if the original dataset is heavily underparameterized, overparameterized subagging always helps, as shown in Figure 4.9(b). For example, recall in Figure 4.3 when $n = 5000$ and $p = 500$ (which is a favorable case in classical statistics), subagged ridgeless predictors trained on overparameterized subsampled datasets (e.g. with $n = 50$ and $p = 500$) with $M = 50$ bags have smaller prediction risk than least squares fitted on the original data.

4.7 Discussion

In this work, we have provided a generic reduction strategy for characterizing the squared risk of general bagged predictors (for two bagging strategies of subagging and splagging). As a function of the numbers of bags M, for the squared error loss, we show that the asymptotic risk of the M-bagged predictor can be expressed as $M^{-1}\mathfrak{R}_1 + (1 - M^{-1})\mathfrak{R}_\infty$, where $\mathfrak{R}_1$ and $\mathfrak{R}_\infty$ represent the asymptotic risks of the M-bagged predictor with $M = 1$ and $M = \infty$, respectively. More generally, for a smooth loss function, we show that the risk of the M-bagged predictor is sandwiched between similar convex combinations. Furthermore, we have provided a generic cross-validation framework to tune the subsample size that aims at obtaining the best subagged predictor, which also helps in monotonizing the risk profile of any given prediction procedure.

Following this general strategy, along with some new tools from random matrix theory, we have derived explicit risk characterization for bagged ridge and ridgeless predictors. The risk expressions reveal bias and variance monotonicity in the number of bags. In comparing different versions of bagging for ridge and ridgeless predictors, we show that subagging (with optimal subsample size) improves upon the divide-and-conquer or the data-splitting approach of averaging the predictors computed on different non-overlapping splits of data (with optimal split size). In the overparameterized regime, the latter data-splitting has been recently observed to improve upon the ridgeless predictor computed on the entire data (Mücke et al., 2022) under sub-Gaussian features.

Surprisingly, our results reveal that, under a well-specified linear model, subagging on properly chosen ridgeless interpolators constantly improves upon the ridgeless predictor trained on the complete data, even when the entire data has more observations than the number of features. Our generic and model-agnostic cross-validation procedure provably yields the best ridgeless interpolators for subagging. Further specializing to the case of isotropic features, we prove that the optimal subagged predictor has the asymptotic risk that matches the unbagged ridge predictor with optimally-tuned regularization parameter.

Several natural extensions of the current work can be considered going forward. We briefly discuss two of them below.

First, though the general strategy we proposed for analyzing bagged predictors can be helpful for other prediction procedures, we have only derived the exact bagged risk expressions when the underlying prediction procedure is ridge and ridgeless regression. In the context of the ridge and ridgeless predictors, we had to develop new random matrix theory tools related to deterministic equivalents. One might have to develop similar new tools to analyze other predictors based on our strategy. A natural prediction procedure to analyze next for bagging is lasso or lassoless regression. An empirical investigation of the bagged lassoless predictor is already conducted by Patil et al. (2022a). The traditional analysis of this predictor trained on the full data is performed via appropriate message passing (AMP) techniques (Li and Wei, 2021). It would be interesting to see if our general strategy can be combined with AMP, the convex Gaussian min-max theorem, or the leave-one-out perturbation analysis to yield a far more general strategy for analyzing bagging.

Second, we have analyzed the bagged ridge and ridgeless predictors under a well-specified linear model. It is interesting to extend the analysis to a general data-distributional setting for two main reasons: (1) to make the results more relevant for practical data analysis, and (2) to investigate whether bagging interpolators can still improve upon the ridgeless predictor trained on the full data. Regarding (1) above, techniques developed by Bartlett et al. (2021) are useful in relaxing the linear model assumptions. Regarding (2) above, we performed a simple simulation study that suggests that even in the misspecified nonlinear model, bagging properly selected interpolators can improve the unbagged ridgeless predictor. See Figure 4.11 for more details.

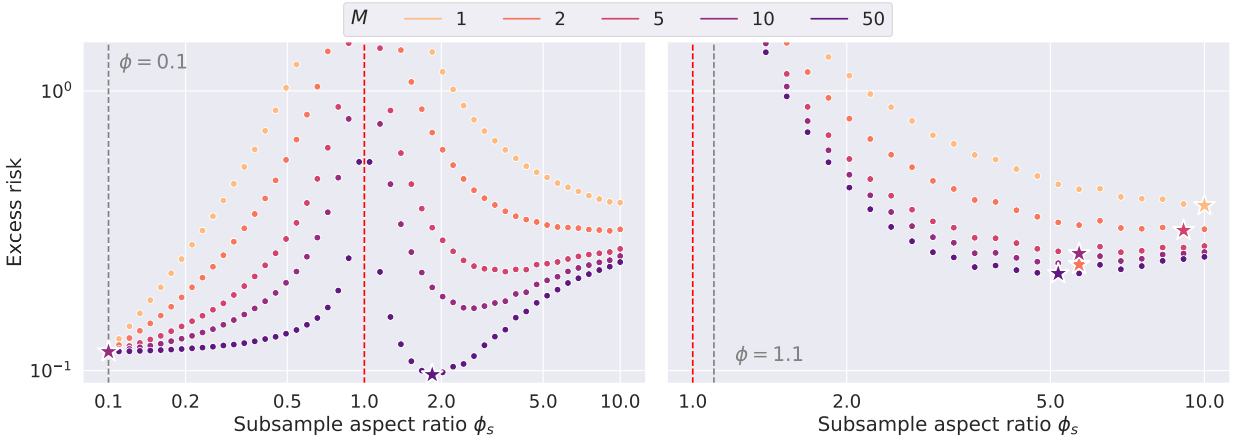


Figure 4.11: Finite-sample prediction risks for subagged ridgeless predictors ($\lambda = 0$) under a nonlinear model, averaged over 100 dataset repetitions, for varying bag size $k = [p\phi_s]$ and number of bags M with replacement, with $n = [p/\phi]$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. We generated data from a nonlinear model where the response y_i for $i \in [n]$ is generated from a nonlinear function of $\boldsymbol{x}_i$ with additive noise: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \frac{1}{p}(\|\boldsymbol{x}_i\|_2^2 - \operatorname{tr}(\boldsymbol{\Sigma}_{\text{ar1}})) + \epsilon_i$ and $\boldsymbol{\beta}_0, \boldsymbol{X}, \boldsymbol{\epsilon}$ are generated as in (M-AR1-LI) with $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$. We observe the similar pattern as in Figure 4.3 that the risk of the subagged ridgeless predictor with $M = 50$ and $\phi_s \approx 1.5$ is smaller than the risk of the ridgeless predictor fitted on the full data. As a consequence, the main results about subagging are likely to in a much wider range of cases.

Chapter 5

Revisiting model complexity

A note to the reader: This chapter is work in progress, and the results presented are partial in nature. As such, please pardon any omissions and lack of clarifications in the meantime that the work is completed.

5.1 Introduction

Modern machine learning involves fitting a large number of parameters relative to the number of observations. Such overparameterized models are typically trained to (nearly) interpolate noisy in-sample data, and yet generalize reasonable well on out-of-sample data in many settings (Zhang et al., 2017). A series of recent work has investigated this surprising phenomenon for different models, including linear regression (Belkin et al., 2019a; Hastie et al., 2022; Muthukumar et al., 2020; Bartlett et al., 2020), random features regression (Mei and Montanari, 2022), sparse regression (Li and Wei, 2021), kernel regression (Liang and Rakhlin, 2020), linear classification (Deng et al., 2019; Montanari et al., 2019b), boosting (Liang and Sur, 2020b), among several others; see Bartlett et al. (2021); Dar et al. (2021) for more examples.

A peculiar feature of overparameterized models is the so-called "double descent" (or even "multiple descent") behavior in the generalization error curve when plotted against the raw number of model parameters or some analogous notion of model complexity. This leads us to ask the following motivating questions in this work:

1. Is there a better and more principled measure of model complexity in general for overparameterized models?

2. More specifically, how do we compare complexity of different (near) interpolating models?

We address these questions through the lens of degrees of freedom, by borrowing and extending classical ideas from optimism theory. In particular, we propose two measures of model complexity, namely *emergent and intrinsic random-X degrees of freedom*. We show the utility of our proposed complexity measures through examples of linear smoothers and interpolators, and illustrate how our proposed measures may help "reconcile" the surprising "multiple descent" generalization behaviors in modern machine learning with the "single descent" bias-variance tradeoff in classical statistical learning. In what follows, we fist summarize our proposals in Section 5.2, and then provide illustrative examples in Section 5.3.

5.2 New proposal for random-X degrees of freedom

Consider the standard regression setup with i.i.d. observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, such that $y_i = f(x_i) + \varepsilon_i$, where $f : \mathbb{R}^p \to \mathbb{R}$ is the regression function, and ε_i has mean 0 and variance σ^2. Denote by $X \in \mathbb{R}^{n \times p}$ the corresponding feature matrix and by $y \in \mathbb{R}^n$ the associated response vector. Let $\mathcal{A}$ be any fitting algorithm that maps $(X, y) \overset{\mathcal{A}}{\mapsto} \widehat{f}$, where $\widehat{f} : \mathbb{R}^p \to \mathbb{R}$ is the resulting fitted predictor. Associated with $\widehat{f}$ are three error metrics: (a) the training error, $\mathrm{ErrT}(\widehat{f}) = n^{-1} \sum_{i=1}^n (y_i - \widehat{f}(x_i))^2$, (b) the fixed-X prediction error, $\mathrm{ErrF}(\widehat{f}) = n^{-1} \sum_{i=1}^{n} \mathbb{E}[(\widetilde{y}_i - \widehat{f}(x_i))^2 | X, y]$, where $\widetilde{y} \in \mathbb{R}^n$ is an independent copy of y at the training points X, and (c) the random-X prediction error, $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2 | X, y]$, where (x_0, y_0) is a test observation sampled independently from the same distribution as the training data.

The training error underestimates both the fixed-X and random-X prediction error in general. In classical statistics, such downward bias is referred to as training *optimism* (Hastie et al., 2009). Define the fixed-X optimism, $\mathrm{OptF}(\widehat{f}) = \mathbb{E}[\mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f}) | X]$, and the random-X optimism, $\mathrm{OptR}(\widehat{f}) = \mathbb{E}[\mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f}) | X]$ (Rosset and Tibshirani, 2019). The fixed-X optimism has been studied extensively and leads to the definition of *fixed-X degrees of freedom* as $\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i) | X)$ (Efron, 1983, 1986; Hastie and Tibshirani, 1990; Efron, 2004), which under certain regularity conditions, is the same as as $\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^{n} \mathbb{E}[\partial \widehat{f}(x_i) / \partial y_i | X]$ (Ye, 1998; Stein, 1981). In some cases, $\mathrm{DofF}(\widehat{f})$ can be computed explicitly: e.g., for linear smoothers $\widehat{f}(X) = L(X)y$, it is given by $\mathrm{tr}[L(X)]$ (Craven and Wahba, 1978, 1979); for lasso, it is given by the expected number of non-zero coefficients in the fitted estimator (Zou et al., 2007; Tibshirani and Taylor, 2012); see Kaufman and Rosset (2014); Janson et al. (2015); Tibshirani (2015) for various other generalizations. In classical statistics, DofF is a widely agreed-upon qualitative measure of complexity and is algorithm-specific, however it is only defined for the fixed-X setup. Despite 50+ years of work on DofF, there is no notion of random-X degrees of freedom that we know of. The goal of this work is to propose a definition for random-X degrees of freedom, denoted by DofR, suitable for the random-X setup underlying most predictive problems.

Towards defining DofR, we first cast the classical definition of the fixed-X degrees of freedom from a different perspective. For a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ can be shown to be equal to the value of k that satisfy the following relation: $\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_{n \times k}, v))$, where $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm, and $U_k \in \mathbb{R}^{n \times k}$ is a certain design matrix consisting n observations and $k \le n$ features, and $v \in \mathbb{R}^n$ is a noise vector with mean 0_n and covariance I_n (see Theorem 5.2.1 for more details). We then extend the same analogy and use the least squares as the reference algorithm and "match" random-X optimisms. We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k (we can show that such k always exists and is unique assuming $k \le n$; see the remarks after Theorem 5.2.1) for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)). \qquad \text{(DofR, emergent)}$$

This measure, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f. We call it *emergent random-X degrees of freedom*. We also define *intrinsic random-X degrees of freedom*, denoted by DofR^i, as the k (which again exists and is unique assuming $k \le n$) for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)). \qquad \text{(DofR, intrinsic)}$$

Apart from analogy with fixed-X degrees of freedom, another reason for choosing the least squares reference algorithm to match optimisms is the following invariance property of OptR that we can show for least squares:

Theorem 5.2.1. *Let $U_k = Z_k \Sigma_k^{1/2}$, where Z_k contains i.i.d. entries of mean 0, variance 1, and bounded moment of order $4 + \mu$ for some $\mu > 0$ and $\Sigma_{k \times k}$ is a positive definite matrix whose minimum and maximum eigenvalues are uniformly bounded away from 0 and ∞. Let v contain i.i.d. entries of mean 0, variance σ^2, and bounded moment of order $4 + \nu$ for some $\nu > 0$. Denote the normalized random-X optimisms of $\widehat{f}$ by $\phi := \mathrm{OptR}(\mathcal{A}(X, y)) / \sigma^2$ and $\psi := \mathrm{OptR}(\mathcal{A}(X, v)) / \sigma^2$ Then, as $n, k \to \infty$ and $k/n \to \xi \in (0, 1)$, we have*

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1 - (1 - \xi)^2}{1 - \xi}, \quad \mathrm{DofR}(\widehat{f}) \to 1 + \frac{\phi}{2} - \sqrt{1 + \frac{\phi^2}{4}}, \quad \mathrm{DofR}^i(\widehat{f}) \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

Remarks: There is remarkable universality in above limits: (1) They do not depend on the exact form of the distributions of U_k and v. (2) They are also independent of Σ_k. This further justifies the choice of the least squares reference algorithm for matching random-X optimisms. We can show an immediate interesting property of the random-X degrees of freedom: There is a unique number that satisfies the desired relations between $[0, n]$. We find this to be a very interpretable range for random-X degrees of freedom. The least complex predictor has DofR of 0, and the most complex predictor has DofR of n, as if the saturated model.

5.3 Explicit and numerical illustrative examples

In general, the random-X degrees of freedom depend of the exact form of the algorithm, but as with DofF, for linear smoothers, DofR^i takes a special interpretable form. It also shows how DofR^i is related to DofF.

Proposition 5.3.1. *Recall the setting of Theorem 5.2.1. Suppose $\widehat{f}$ is a linear smoother such that $\widehat{f}(X) = L(X)y$ and $\widehat{f}(x_0) = \ell(x_0)^\top y$ for some smoothing matrix $L \in \mathbb{R}^{n\times n}$ and smoothing weight function $\ell : \mathbb{R}^p \to \mathbb{R}^n$. Then, we have*

$$\psi = 2\,\mathrm{tr}[L(X)]/n + \mathbb{E}[\ell(x_0)^\top \ell(x_0)] - \mathrm{tr}[L(X)^\top L(X)]/n, \quad \text{and} \quad \mathrm{DofR}^i(\widehat{f}) \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

Remarks: Some special cases of interest are: (1) Interpolating models for which $L(X) = I_n$. In this case, ψ simplifies to $1 + \mathbb{E}[\ell(x_0)^\top \ell(x_0)]$. As a result, DofR^i differs between different interpolating models as opposed to DofF which is always $\mathrm{tr}[L(X)] = n$ for any interpolating model. (2) In the special case of min ℓ_2-norm interpolator, we can prove the following interesting property: in the underparameterized regime when $p \leq n$, we have DofR^i/n strictly increasing from $[0, 1]$ as expected, while in the overparameterized regime when $p > n$, DofR^i/n is strictly decreasing from $(1, 0)$, so DofR^i is maximized at $p = n$. This result holds for any feature covariance Σ and shows overparameterization indeed reduces the intrinsic complexity.

Beyond linear smoothers, properties of DofR and DofR^i depend on the specific fitting procedure. Below we compare min ℓ_2-norm interpolator with min ℓ_1-norm interpolator, abbreviated mn2ls and mn1ls, whose risks are recently shown to exhibit double (Hastie et al., 2022) and multiple descents (Li and Wei, 2021), respectively. Note that latter is a non-linear procedure. We observe from Figure 5.1 that our proposed notion of intrinsic degrees-of-freedom reconciles the "bias-variance" tradeoff and turns modern "double descents" into classical "single descents".

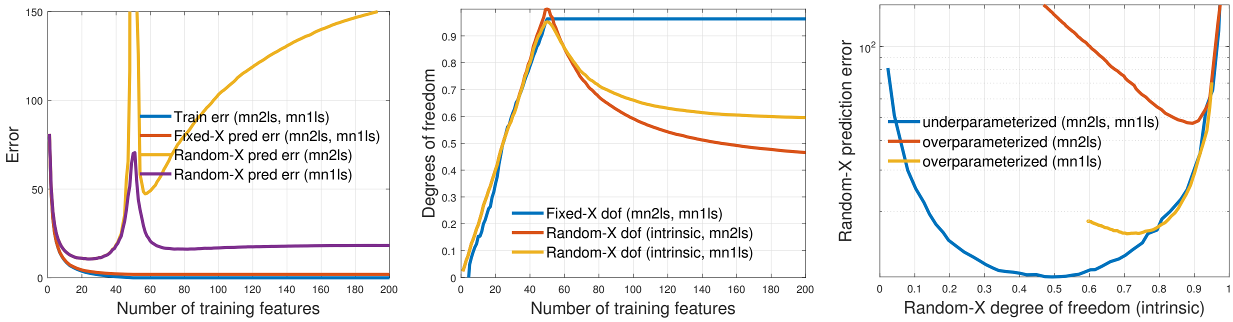


Figure 5.1: We consider a fixed data generating model with $n = 200$ and response non-linear in $p = 200$ feature, and consider training estimators with varying number of features. This model is similar to that used in Belkin et al. (2019a).

Appendix A

Supplement for Chapter 1

This supplement contains proofs and additional details for Chapter 1. The content of the supplement is organized as follows.

- In Appendix A.1, we provide proofs of the constituent Lemmas 1.5.1 to 1.5.4 related to Theorem 1.4.1, along with the remaining steps to complete the proof of Theorem 1.4.1.
- In Appendix A.2, we provide proof of the constituent Lemma 1.5.6 related to Theorem 1.4.2, along with the remaining steps to complete the proof of Theorem 1.4.2.
- In Appendix A.3, we list and prove auxiliary lemmas that we need in other proofs.
- In Appendix A.4, we list useful concentration results that are used in the proofs throughout.

A.1 Proofs related to Theorem 1.4.1

A.1.1 Proof of Lemma 1.5.1

Recall from (1.2) that the expected out-of-sample prediction error of the ridge estimator $\widehat{\beta}_\lambda$ is defined as

$$\mathrm{Err}(\widehat{\beta}_\lambda) = \mathbb{E}_{x_0, y_0}\left[(y_0 - x_0^T \widehat{\beta}_\lambda)^2 \mid X, y\right].$$

Under a well-specified linear response $y_0 = x_0^T \beta_0 + \varepsilon_0$, the prediction error can be decomposed as

$$\begin{aligned}\mathrm{Err}(\widehat{\beta}_\lambda) &= \mathbb{E}\left[(\beta_0 - \widehat{\beta}_\lambda)^T x_0 x_0^T (\beta_0 - \widehat{\beta}_\lambda) \mid X, y\right] + \mathbb{E}\left[(\beta_0 - \widehat{\beta}_\lambda)^T x_0 \varepsilon_0 \mid X, y\right] + \mathbb{E}\left[\varepsilon_0^2 \mid X, y\right] \\ &= (\beta_0 - \widehat{\beta}_\lambda)^T \Sigma (\beta_0 - \widehat{\beta}_\lambda) + \sigma^2. \end{aligned} \tag{A.1}$$

Here we used the fact that $\mathbb{E}\big[x_0 \varepsilon_0\big] = 0$ as ε_0 is independent of x_0. Using the expression of $\widehat{\beta}_\lambda$ from (1.1), the deviation $\beta_0 - \widehat{\beta}_\lambda$ can be expressed as

$$\begin{aligned}\beta_0 - \widehat{\beta}_\lambda &= \beta_0 - (X^T X/n + \lambda I_p)^+ X^T y/n \\ &= \beta_0 - (X^T X/n + \lambda I_p)^+ X^T (X\beta_0 + y - X\beta_0)/n \\ &= \big(I_p - (X^T X/n + \lambda I_p)^+ X^T X/n\big)\beta_0 - (X^T X/n + \lambda I_p)^+ X^T \varepsilon/n. \end{aligned}$$

Note that the first component depends on the signal parameter β_0 and the second depends on the error vector ε. Plugging this into (A.1), and denoting $X^T X/n$ by $\widehat{\Sigma}$ and $\mathrm{Err}(\widehat{\beta}(\lambda))$ by $\mathrm{err}(\lambda)$, we have the following decomposition of the prediction error for any $\lambda \in \mathbb{R}$:

$$\mathrm{err}(\lambda) = \mathrm{err}_b(\lambda) + \mathrm{err}_c(\lambda) + \mathrm{err}_v(\lambda), \tag{A.2}$$

where $\mathrm{err}_b(\lambda)$, $\mathrm{err}_v(\lambda)$, and $\mathrm{err}_c(\lambda)$ are the bias, variance, and cross components in the decomposition given by

$$\begin{aligned}\mathrm{err}_b(\lambda) &= \beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0,\\ \mathrm{err}_c(\lambda) &= -2\beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+ X^T\varepsilon/n,\\ \mathrm{err}_v(\lambda) &= \varepsilon^T \big(X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big)\varepsilon/n + \sigma^2.\end{aligned}$$

For any $\lambda \in (\lambda_{\min}, \infty)$, we establish below that

$$\mathrm{err}_c(\lambda) \xrightarrow{\text{a.s.}} 0 \tag{A.3}$$

under proportional asymptotic limit. The desired decomposition in Lemma 1.5.1 then follows by plugging convergence in (A.3) into (A.2).

To establish the convergence in (A.3), let us write $\mathrm{err}_c(\lambda) = a_n^T\varepsilon/n$ where $a_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$a_n = -2X(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0.$$

We note that for $\lambda \in (\lambda_{\min}, \infty)$,

$$\begin{aligned}\|a_n\|^2/n &= 4\beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0\\ &\le C\Big\|\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Big\|\\ &\le C,\end{aligned}$$

where the first inequality uses bound on the signal energy from Assumption 1.4 and the second inequality holds almost surely for large n by using the facts that $\|\widehat{\Sigma}\| \le C(\sqrt{\gamma}+1)^2\|\Sigma\|$, $\|(\widehat{\Sigma} + \lambda I_p)^+\| \le (\lambda - \lambda_{\min})^{-1}$ almost surely for n large enough from Assumption 1.2 and $\|\Sigma\| \le r_{\max}$ from Assumption 1.3. In addition, ε has i.i.d. entries satisfying Assumption 1.1. The desired result then follows from application of Lemma A.4.1.

A.1.2 Proof of Lemma 1.5.2

We start by writing the GCV risk estimate $\mathrm{gcv}(\lambda)$ for the ridge estimator from (1.5) as

$$\mathrm{gcv}(\lambda) = \frac{y^T(I_n - L_\lambda)^2 y/n}{\big(1 - \mathrm{tr}[L_\lambda]/n\big)^2} \tag{A.4}$$

where L_λ is the ridge smoothing matrix. Note that (A.4) is of the form $\frac{0}{0}$ when $L_\lambda = I_n$ (which happens when $\lambda = 0$ and X has rank n). In this case, we define the GCV risk estimate as the corresponding limit as $\lambda \to 0$. We handle this case separately below.

The denominator of (A.4) can be expressed as

$$\begin{aligned}1 - \mathrm{tr}[L_\lambda]/n &= 1 - \mathrm{tr}\left[X(X^TX/n + \lambda I_p)^+X^T/n\right]/n\\ &= 1 - \mathrm{tr}\left[(X^TX/n + \lambda I_p)^+X^TX/n\right]/n.\end{aligned}$$

The numerator of (A.4) can be expressed as

$$\begin{aligned}y^T(I_n - L_\lambda)^2 y/n &= (X\beta_0 + \varepsilon)^T(I_n - L_\lambda)^2(X\beta_0 + \varepsilon)/n\\ &= \beta_0^T X^T(I_n - L_\lambda)^2 X\beta_0/n + 2\beta_0^T X^T(I_n - L_\lambda)^2\varepsilon/n + \varepsilon^T(I_n - L_\lambda)^2\varepsilon/n.\end{aligned}$$

Consider the first term of the numerator expression. The factor $X^T(I_n - L_\lambda)^2X$ can be expressed as

$$\begin{aligned}X^T(I_n - L_\lambda)^2X &= X^T\big(I_n - X(X^TX/n + \lambda I_p)^+X^T/n\big)^2X\\ &= \big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+X^T\big)\big(X - X(X^TX/n + \lambda I_p)^+X^TX/n\big)\\ &= \big(I_p - X^TX/n(X^TX/n + \lambda I)^+\big)X^TX\big(I_p - (X^TX/n + \lambda I_p)^+X^TX/n\big).\end{aligned}$$

Consider the second term of the numerator expression. The factor $X^T(I_n - L_\lambda)^2$ can be expressed as

$$\begin{aligned}
X^T(I_n - L_\lambda)^2 &= X^T\big(I_n - X(X^TX/n + \lambda I_p)^+X^T/n\big)^2 \\
&= \big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+X^T\big)\big(I_n - X(X^TX/n + \lambda I_p)^+X^T/n\big) \\
&= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)X^T\big(I_n - X(X^TX/n + \lambda I_p)^+X^T/n\big) \\
&= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)\big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+X^T\big) \\
&= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)\big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)X^T
\end{aligned}$$

Consider the third term of the numerator expansion. The factor $(I_n - L_\lambda)^2$ can be expressed as

$$(I_n - L_\lambda)^2 = \big(I_n - X(X^TX/n + \lambda I_p)^+X^T/n\big)^2$$

Case when $\lambda \neq 0$. The GCV denominator $1 - \operatorname{tr}\big[(X^TX/n + \lambda I_p)^+X^TX/n\big]/n \neq 0$ when $\lambda \neq 0$. Thus plugging the denominator and numerator expansions into (A.4) and denoting X^TX/n by $\widehat{\Sigma}$, the GCV risk estimate can be decomposed as

$$\operatorname{gcv}(\lambda) = \frac{\operatorname{gcv}_b(\lambda) + \operatorname{gcv}_c(\lambda) + \operatorname{gcv}_v(\lambda)}{\operatorname{gcv}_d(\lambda)}, \tag{A.5}$$

where $\operatorname{gcv}_b(\lambda)$, $\operatorname{gcv}_v(\lambda)$, and $\operatorname{gcv}_c(\lambda)$ are the bias-like, variance-like, and cross components in the decomposition given by

$$\operatorname{gcv}_b(\lambda) = \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0,$$

$$\operatorname{gcv}_c(\lambda) = 2\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)^2X^T\varepsilon/n,$$

$$\operatorname{gcv}_v(\lambda) = \varepsilon^T\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)^2\varepsilon/n,$$

and $\operatorname{gcv}_d(\lambda)$ is the normalization factor given by

$$\operatorname{gcv}_d(\lambda) = \big(1 - \operatorname{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+]/n\big)^2.$$

Similar to the proof of Lemma 1.5.1, we now establish that

$$\operatorname{gcv}_c(\lambda) \xrightarrow{\text{a.s.}} 0 \tag{A.6}$$

under proportional asymptotic limit. Let us write $\operatorname{gcv}_c(\lambda) = b_n^T\varepsilon/n$ where $b_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$b_n = 2X\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big)^2\beta_0.$$

As argued in the proof of Lemma 1.5.1, for $\lambda \in (\lambda_{\min}, \infty)$,

$$\begin{aligned}
\|b_n\|^2/n &= 4\beta_0^T\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big)^2\widehat{\Sigma}\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big)^2\beta_0 \\
&\le C\Big\|\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big)^2\widehat{\Sigma}\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big)^2\Big\| \\
&\le C
\end{aligned}$$

almost surely for large n, and since ε has i.i.d. entries satisfying Assumption 1.1, the convergence in (A.6) follow from application of Lemma A.4.1.

Limiting case when $\lambda = 0$. To handle the case when $\mathrm{gcv}_d(\lambda)$ can be zero, we note that when $\lambda \neq 0$ using Lemma A.3.2 the components in the decomposition (A.5) can be alternately expressed as

$$\mathrm{gcv}_b(\lambda) = \beta_0^T \lambda^2 (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I_p)^+ \beta_0,$$

$$\mathrm{gcv}_b(\lambda) = 2\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I_p)^+ (\widehat{\Sigma} + \lambda I_p)^+ X^T \varepsilon / n,$$

$$\mathrm{gcv}_v(\lambda) = \lambda^2 \varepsilon^T (XX^T/n + \lambda I_n)^+ (XX^T/n + \lambda I_n)^+ \varepsilon,$$

$$\mathrm{gcv}_d(\lambda) = \lambda^2 \big(\mathrm{tr}[(XX^T/n + \lambda I_n)^+]/n \big)^2.$$

We can then cancel the factor of λ^2 and take the limit $\lambda \to 0$ to get the limiting GCV decomposition as

$$\mathrm{gcv}(0) = \frac{\mathrm{gcv}_b(0) + \mathrm{gcv}_b(0) + \mathrm{gcv}_v(0)}{\mathrm{gcv}_d(0)}, \tag{A.7}$$

where the limiting bias-like, variance-like and cross components in the decomposition are given by

$$\mathrm{gcv}_b(0) = \beta_0^T \widehat{\Sigma}^+ \widehat{\Sigma} \widehat{\Sigma}^+ \beta_0 = \beta_0^T \widehat{\Sigma}^+ \beta_0,$$

$$\mathrm{gcv}_c(0) = 2\beta_0^T \widehat{\Sigma}^{+2} X^T \varepsilon / n,$$

$$\mathrm{gcv}_v(0) = \varepsilon^T (XX^T/n)^{+2} \varepsilon / n,$$

and the limiting normalization can be written as

$$\mathrm{gcv}_d(0) = \big(\mathrm{tr}[\widehat{\Sigma}^+]/n \big)^2$$

by noting that $\mathrm{tr}[(XX^T/n)^+] = \mathrm{tr}[(X^T X/n)^+]$. As before, let us establish that

$$\mathrm{gcv}_c(0) \xrightarrow{\text{a.s.}} 0 \tag{A.8}$$

under proportional asymptotics. We write $\mathrm{gcv}_c(0) = b_n^T \varepsilon / n$ where $b_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$b_n = 2X\widehat{\Sigma}^{+2}\beta_0.$$

We note that $\|b_n\|^2/n$ is almost surely bounded for large n and ε contains i.i.d. entries satisfying Assumption 1.1. Using Lemma A.4.1, we conclude the convergence.

The desired decomposition in Lemma 1.5.2 then follows by using the convergences in (A.6) and (A.8) into (A.5) and (A.7), respectively.

A.1.3 Proof of Lemma 1.5.3

We start with $\mathrm{gcv}_b(\lambda)$ and first establish that

$$\beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \widehat{\Sigma} \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \beta_0 - \frac{\beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \Sigma \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \beta_0}{\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2} \xrightarrow{\text{a.s.}} 0. \tag{A.9}$$

To that end, let $B := \beta_0 \beta_0^T$ and break the left-hand side into sum of quadratic forms evaluated at the n observations as follows:

$$\begin{aligned}
&\beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \widehat{\Sigma} \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \beta_0 \\
&= \mathrm{tr}\left[B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \widehat{\Sigma} \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \right] \\
&= \mathrm{tr}\left[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \widehat{\Sigma} \right] \\
&= \mathrm{tr}\left[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) \sum_{i=1}^{n} x_i x_i^T / n \right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr} \Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) x_i x_i^T \Big] \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) x_i.
\end{aligned}
$$

The summands $x_i^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) x_i$ are quadratic forms where the point of evaluation x_i and the matrix $\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big)$ are dependent. To break the dependence, we use the standard leave-one-out trick and the Sherman-Morrison-Woodbury formula with Moore-Penrose pseudo-inverse (Meyer, 1973). Let us temporarily call $w_i := B(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}) x_i$ and proceed as follows:

$$
\begin{aligned}
& x_i^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) x_i \\
&= w_i^T (I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}) x_i \\
&= w_i^T \big(I_p - (\widehat{\Sigma}_{-i} + x_i x_i^T/n)(\widehat{\Sigma}_{-i} + \lambda I_p + x_i x_i^T/n)^{+}\big) x_i \\
&= w_i^T \left(I_p - (\widehat{\Sigma}_{-i} + x_i x_i^T/n) \left((\widehat{\Sigma}_{-i} + \lambda I_p)^{+} - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right) \right) x_i \\
&= w_i^T x_i - w_i^T \left(\widehat{\Sigma}_{-i} + x_i x_i^T/n\right) \left((\widehat{\Sigma}_{-i} + \lambda I_p)^{+} - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right) x_i \\
&= w_i^T x_i - w_i^T \left(\widehat{\Sigma}_{-i} + x_i x_i^T/n\right) \left((\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right) \\
&= w_i^T x_i - w_i^T \left(\widehat{\Sigma}_{-i} + x_i x_i^T/n\right) \\
&\qquad \cdot \left(\frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i + (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T (\widehat{\Sigma} + \lambda I_p)^{+} x_i/n - (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right) \\
&= w_i^T x_i - \frac{w_i^T (\widehat{\Sigma}_{-i} + x_i x_i^T/n)(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \\
&= \frac{w_i^T x_i + w_i^T x_i x_i^T (\widehat{\Sigma} + \lambda I_p)^{+} x_i/n - w_i^T \widehat{\Sigma}_{-i} (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i - w_i^T x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \\
&= \frac{w_i^T x_i - w_i^T \widehat{\Sigma}_{-i} (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \\
&= \frac{w_i^T (I_p - \widehat{\Sigma}_{-i} (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}) x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \\
&= \frac{x_i^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n}.
\end{aligned}
$$

By carrying our similar leave-one-out strategy on the other side, we can further simplify

$$
\frac{x_i^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} = \frac{x_i^T \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n\big)^2}.
$$

We now split the error to the target in (A.9) as follows:

$$
\operatorname{tr} \Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \widehat{\Sigma} \Big] - \frac{\operatorname{tr} \Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+} \Sigma\big]/n\Big)^2}
$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^T\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{\big(1 + x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+ x_i/n\big)^2} - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n\Big)^2}$$

$$= e_1 + e_2, \text{ where}$$

$$e_1 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i^T\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{\Big(1 + x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+ x_i/n\Big)^2} - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n\Big)^2} \right),$$

$$e_2 := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n\Big)^2} - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n\Big)^2} \right).$$

In Appendix A.1.6, we show that both terms e_1 and e_2 almost surely approach 0 under proportional asymptotics.

Let us provide some intuition as follows. On one hand, in the error term e_1, conditional on X_{-i}, expected value of $x_i^T\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i$ is $\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]$ and the expected value of $x_i^T(\widehat{\Sigma}_{-i}+\lambda I)^+ x_i/n$ is $\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I)^+\Sigma\big]/n$. Because of concentration of these quantities around their respective expectations rapid enough, the error term e_1 is almost surely 0. On the other hand, for e_2, $\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]$ and $\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]$, and $\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n$ and $\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n$, the matrices involved differ by rank-1 component. The difference is almost surely 0 in the proportional asymptotic limit. We note that this strategy is similar to the ones used by, for example, Rubio and Mestre (2011); Ledoit and Péché (2011) to obtain expressions for certain functionals involving Σ and $\widehat{\Sigma}$ in terms of Σ. The main difference is that the eventual target in our case is defined solely in terms of $\widehat{\Sigma}$ rather than Σ.

We have so far established that

$$\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\Big] - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2} \xrightarrow{\text{a.s.}} 0,$$

which after expressing B in terms of β_0 and moving the denominator across yields

$$\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2 \beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0 - \beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0 \xrightarrow{\text{a.s.}} 0. \tag{A.10}$$

Case when $\lambda \neq 0$. We now use the $\lambda \neq 0$ case of Lemma A.3.1 to get

$$\frac{\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\widehat{\Sigma}\big]/n\Big)^2} - \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0 \xrightarrow{\text{a.s.}} 0$$

under proportional asymptotics as desired.

Limiting case when $\lambda = 0$. To handle the $\lambda = 0$ case, we first express $I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+ = \lambda(\widehat{\Sigma}+\lambda I_p)^+$ when $\lambda \neq 0$ using Lemma A.3.2. We can then move factor of λ^2 from $\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0$ to $\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2$ such that

$$\begin{aligned}
&\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2 \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0 \\
&= \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2 \lambda^2 \beta_0^T(\widehat{\Sigma}+\lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\beta_0
\end{aligned}$$

$$
\begin{aligned}
&= \Big(\lambda + \lambda \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^{+}\Sigma\big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\beta_0 \\
&= \Big(\lambda + \operatorname{tr}\big[\lambda(\widehat{\Sigma} + \lambda I)^{+}\Sigma\big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\beta_0 \\
&= \Big(\lambda + \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big)\Sigma\Big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\beta_0.
\end{aligned}
$$

Using the above expression in (A.10) and sending $\lambda \to 0$ thus yields

$$
\Big(\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma\big]/n\Big)^2 \beta_0^T \widehat{\Sigma}^{+}\widehat{\Sigma}\widehat{\Sigma}^{+}\beta_0 - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\beta_0 \xrightarrow{\text{a.s.}} 0,
$$

or in other words,

$$
\Big(\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma\big]/n\Big)^2 \beta_0^T \widehat{\Sigma}^{+}\beta_0 - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\beta_0 \xrightarrow{\text{a.s.}} 0.
$$

Using Lemma A.3.1 for this case, we then have

$$
\frac{\beta_0^T \widehat{\Sigma}^{+}\beta_0}{\big(\operatorname{tr}[\widehat{\Sigma}^{+}]/n\big)^2} - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^{+})\beta_0 \xrightarrow{\text{a.s.}} 0
$$

under proportional asymptotics, completing both the cases in Lemma 1.5.3.

A.1.4 Proof of Lemma 1.5.4

Case when $\lambda \neq 0$. Under proportional asymptotic limit, our goal is to show that

$$
\varepsilon^T\big(X(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)\varepsilon/n + \sigma^2 - \frac{\varepsilon^T\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)^2\varepsilon/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\big]/n\big)^2} \xrightarrow{\text{a.s.}} 0.
$$

We first note that $\varepsilon^T\varepsilon/n$ almost surely approaches σ^2 from the strong law of large numbers. Thus we can slightly rephrase our goals to show as

$$
\varepsilon^T\left[\big(X(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big) + I_n - \frac{\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)^2}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\big]/n\big)^2}\right]\varepsilon/n \xrightarrow{\text{a.s.}} 0.
$$

Our main strategy is to show that under proportional asymptotic limit

$$
\operatorname{tr}\big[X(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big]/n + 1 - \frac{\operatorname{tr}\Big[\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)^2\Big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\big]/n\big)^2} \xrightarrow{\text{a.s.}} 0. \tag{A.11}
$$

The desired convergence then follows by using Lemma A.4.2.

We proceed by decomposing the first component of (A.11) as follows:

$$
\begin{aligned}
\operatorname{tr}\big[X(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)X^T/n\big]/n &= \operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}\big]/n \\
&= \operatorname{tr}\big[\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}\big]/n - \operatorname{tr}\big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}\big]/n.
\end{aligned}
$$

For the numerator of the second component of (A.11), we note that

$$
\begin{aligned}
&\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)^2 \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big)\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big) - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^{+}X^T/n\big) - X(X^TX/n + \lambda I_p)^{+}X^T/n\big(I_n - X(X^TX/n + \lambda I_p)^{+}X^T/n\big)
\end{aligned}
$$

$$
\begin{aligned}
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(X^TX/n + \lambda I_p)^+ \big(X^T/n - X^TX/n(X^TX/n + \lambda I_p)^+ X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(X^TX/n + \lambda I_p)^+ \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big) X^T/n.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\frac{\operatorname{tr}\big[I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big]^2/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\big)^2} \\
&= \frac{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n - \operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\big)^2} \\
&= \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n} - \frac{\operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ (I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+)\big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\big)^2}.
\end{aligned}
$$

To establish the desired equivalence, we now use the following two individual equivalences:

$$
\operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n - \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n} + 1 \xrightarrow{\text{a.s.}} 0,
$$

which follows from Lemma A.3.1, and

$$
\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\Big]/n - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Big]/n}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)^2} \xrightarrow{\text{a.s.}} 0,
$$

which follows analogously from the equivalence established in the proof of Lemma 1.5.3 with $B = I_p$.

Limiting case when $\lambda = 0$. To handle the case when $\lambda = 0$, we observe that when $\lambda \neq 0$, we can write

$$
\operatorname{tr}\big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big]/n = 1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n + \lambda^2 \operatorname{tr}\big[(XX^T/n + \lambda I_n)^{+2}\big]/n,
$$

along with

$$
1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n = \lambda \operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n,
$$

which follow from Lemma A.3.2. This allows us to cancel the factor of λ^2 to write

$$
\operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ \Sigma(\widehat{\Sigma} + \lambda I_p)^+\big]/n - \frac{\operatorname{tr}\big[(XX^T/n + \lambda I_n)^{+2}\big]/n}{\big(\operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n\big)^2} + 1 \xrightarrow{\text{a.s.}} 0,
$$

which in the limiting case by sending $\lambda \to 0$ provides the equivalence

$$
\operatorname{tr}[\widehat{\Sigma}^+ \Sigma]/n - \frac{\operatorname{tr}[\widehat{\Sigma}^{+2}]/n}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2} + 1 \xrightarrow{\text{a.s.}} 0
$$

under proportional asymptotic limit. Note that we have written the final expression in terms $\widehat{\Sigma}$ instead of XX^T/n simply for consistency with the $\lambda \neq 0$ case. Combining the two cases, we have the desired limiting equivalences in Lemma 1.5.4.

A.1.5 Completing the proof of Theorem 1.4.1

Lemmas 1.5.1 to 1.5.4 establish the almost sure pointwise convergence of $\operatorname{gcv}(\lambda)$ to $\operatorname{err}(\lambda)$ under proportional asymptotics for $\lambda \in (\lambda_{\min}, \infty)$. To complete the proof of Theorem 1.4.1, we now show that the convergence holds uniformly over compact subintervals of $(\lambda_{\min}, \infty)$ and subsequently show the convergence of tuned risks over such intervals.

The strategy is show that, on any compact subinterval $I \subseteq (\lambda_{\min}, \infty)$, $\mathrm{gcv}(\lambda)$ and $\mathrm{err}(\lambda)$, and their derivatives, as functions of λ are bounded over I. This provides equicontinuity of family as functions of λ over I. The Arzela-Ascoli theorem then provides the desired uniform convergence. The convergence of tuned risks subsequently follows from a standard argument.

We start by writing the GCV estimate (A.4) for the ridge estimator as

$$\mathrm{gcv}(\lambda) = \frac{y^T (I_n - L_\lambda)^2 y/n}{\big(\operatorname{tr}[I_n - L_\lambda]/n\big)^2}.$$

It is convenient to first assume $\lambda \neq 0$ and express $I_n - L_\lambda$ as $\lambda(XX^T/n + \lambda I_n)^+$ using Lemma A.3.2 and then cancel the factor of λ^2 from both the numerator and denominator, which also covers the limiting $\lambda \to 0$ case. This lets us write the GCV estimate as

$$\mathrm{gcv}(\lambda) = \frac{u_n(\lambda)}{v_n(\lambda)}, \tag{A.12}$$

where $u_n(\lambda) = y^T (XX^T/n + \lambda I_n)^{+2} y/n$, and the denominator $v_n(\lambda) = \Big(\operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n\Big)^2$. We first bound the numerator and denominator appropriately. Let $s_{\min}$ and $s_{\max}$ denote the minimum non-zero and maximum eigenvalues of XX^T/n, respectively. We can upper bound the numerator as

$$|u_n(\lambda)| \le \frac{\|y\|^2}{n} \frac{1}{(s_{\min} + \lambda)^2}, \tag{A.13}$$

and we can lower bound the denominator as

$$|v_n(\lambda)| \ge \frac{1}{(s_{\max} + \lambda)^2}. \tag{A.14}$$

Using the two bounds in (A.13) and (A.14) into (A.12), we have the following upper bound on the GCV estimate:

$$|\mathrm{gcv}(\lambda)| \le \frac{\|y\|^2}{n} \left(\frac{s_{\max} + \lambda}{s_{\min} + \lambda} \right)^2.$$

From the strong law of large numbers we note that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n. From Bai and Silverstein (1998), we have that $s_{\max} \le C(1 + \sqrt{\gamma})^2 r_{\max}$ for any $C > 1$ and $s_{\min} \ge c(1 - \sqrt{\gamma})^2 r_{\min}$ for any $c < 1$ almost surely for sufficiently large n, where $r_{\min}$ and $r_{\max}$ denote the bounds on the minimum and maximum eigenvalues of Σ from Assumption 1.3. Thus, over any compact subinterval I of $(\lambda_{\min}, \infty)$, $\mathrm{gcv}(\lambda)$ is bounded almost surely for sufficiently large n.

We next bound the derivative of $\mathrm{gcv}(\lambda)$ as a function of λ. We start with the quotient rule of the derivatives to write:

$$\mathrm{gcv}'(\lambda) = \frac{u_n'(\lambda) v_n(\lambda) - u_n(\lambda) v_n'(\lambda)}{v_n(\lambda)^2}. \tag{A.15}$$

We now upper bound the derivatives of $u_n(\lambda)$ and $v_n(\lambda)$, and additionally obtain an upper bound on $v_n(\lambda)$. From short calculations, we can upper bound the derivative of the numerator as

$$|u_n'(\lambda)| \le \frac{2\|y\|^2}{n} \left| \frac{1}{(s_{\min} + \lambda)^3} \right|, \tag{A.16}$$

and the derivative of the denominator as

$$|v_n'(\lambda)| \le \left| \frac{2}{(s_{\min} + \lambda)^3} \right|. \tag{A.17}$$

In addition, we can upper bound the denominator as

$$|v_n(\lambda)| \le \frac{1}{(s_{\min} + \lambda)^2}. \tag{A.18}$$

Combining the bounds in (A.16) to (A.18), along with the bounds in (A.13) and (A.14), into (A.15), we get the following upper bound on the derivative:

$$|\mathrm{gcv}'(\lambda)| \le \frac{4\|y\|^2}{n} \left| \frac{(s_{\max} + \lambda)^4}{(s_{\min} + \lambda)^5} \right|. \tag{A.19}$$

As before, we note that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n, and $s_{\max}$ is upper bounded and $s_{\min}$ lower bounded above $(\sqrt{\gamma} - 1)^2 r_{\min}$ for sufficiently large n. Thus, over any compact subinterval I of $(\lambda_{\min}, \infty)$, $|\mathrm{gcv}'(\lambda)|$ is almost surely upper bounded for sufficiently large n.

By similar arguments, we can bound the $\mathrm{err}(\lambda)$ and its derivative as a function of λ. Together, we have that the function $\mathrm{err}(\lambda) - \mathrm{gcv}(\lambda)$ forms an equicontinous family of functions of λ over any compact subinterval of $(\lambda_{\min}, \infty)$. Applying the Arzela-Ascoli theorem, we conclude uniform convergence for a subsequence, and since the difference converges pointwise to 0, the uniform convergence holds for the entire sequence.

Finally, we use the uniform convergence to establish the convergence of the tuned risks by a standard argument. We start with the observation that $\mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) \le \mathrm{gcv}(\lambda)$ for any $\lambda \in I$ using the optimality of $\widehat{\lambda}_I^{\mathrm{gcv}}$. Using the specific $\lambda = \lambda_I^\star$, we thus have that $\mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) \le \mathrm{gcv}(\lambda_I^\star)$. We next note that

$$\begin{aligned}
\mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{err}(\lambda_I^\star) &= \mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) + \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\lambda_I^\star) + \mathrm{gcv}(\lambda_I^\star) - \mathrm{err}(\lambda_I^\star) \\
&\le \mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) + \mathrm{gcv}(\lambda_I^\star) - \mathrm{err}(\lambda_I^\star) \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

where the inequality follows from the optimality of $\widehat{\lambda}_I^{\mathrm{gcv}}$ for $\mathrm{gcv}(\lambda)$ and the two almost sure convergences follow from the uniform convergence. This concludes the proof of Theorem 1.4.1.

A.1.6 Error terms in the proof of Lemma 1.5.3

It is convenient to further split $e_1 = e_{11} + e_{12}$ where the suberror terms e_{11} and e_{12} are defined as follows:

$$e_{11} := \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i^T \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i / n\big)^2} - \frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i / n\big)^2} \right),$$

$$e_{12} := \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i / n\big)^2} - \frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\right] / n\big)^2} \right).$$

We similarly split $e_2 = e_{21} + e_{22}$ where the suberror terms e_{21} and e_{22} are defined as follows:

$$e_{21} := \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\right] / n\big)^2} - \frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right] / n\big)^2} \right)$$

$$e_{22} := \frac{1}{n} \sum_{i=1}^n \left(\frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right] / n\big)^2} - \frac{\mathrm{tr}\left[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right] / n\big)^2} \right).$$

Below we show that for $\lambda \in (\lambda_{\min}, \infty)$ all the suberror terms almost surely approach 0 as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Note that we use a generic letter C to denote a constant (that does not depend on n or p) whose value can change from line to line and the inequality sign is used in an asymptotic sense which holds almost surely for sufficiently large n.

Error term e_{11}

We bound the error term e_{11} as follows:

$$|e_{11}| = \left| \frac{1}{n} \sum_{i=1}^n \frac{x_i^T \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i - \mathrm{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)\Sigma\right]}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i / n\big)^2} \right|$$

$$
\begin{aligned}
&\leq C\left|\frac{1}{n}\sum_{i=1}^{n} x_i^T\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)x_i-\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)\Sigma\Big]\right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the first inequality follows by noting that from Lemma A.4.2 the quadratic form $x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n$ converges almost surely to $\operatorname{tr}[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma]/n$ (as operator norm of $(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}$ is almost surely bounded for large n) and the fact that $\left|1/\big(1+\operatorname{tr}[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma]/n\big)\right|$ is bounded by viweing $\operatorname{tr}[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma]/n$ as a Stieljes transform of a measure with bounded total mass (see, for example, Paul and Silverstein (2009); Couillet and Hachem (2014)). The convergence in the final step follows from application of Lemma A.4.4 since $\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)$ has trace norm almost surely bounded for large n (as trace norm of B is bounded and the operator norm of $\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)$ is almost surely bounded for large n).

Error term e_{12}

We bound the error term e_{12} as follows:

$$
\begin{aligned}
|e_{12}| &= \left|\frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)\Sigma\Big]\left(\frac{1}{\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2}-\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right)\right| \\
&\leq C\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2}-\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right| \\
&= C\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2-\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2}{\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right| \\
&\leq C\left|\frac{1}{n}\sum_{i=1}^{n}\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2-\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2\right| \\
&\leq C\max_{i=1,\ldots,n}\left|\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n\big)^2-\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2\right| \\
&\leq C\max_{i=1,\ldots,n}\left|x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n-\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\right|\left|2+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\right| \\
&\leq C\max_{i=1,\ldots,n}\left|x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}x_i/n-\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the first inequality bound follows from noting that the matrix $\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)\Sigma$ almost surely has bounded trace norm for large n (since trace norm of $\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)$ is bounded almost surely for large n as argued for the error term e_{11} above and the operator norm of Σ is bounded) and the final convergence follows from using Lemma A.4.3 by noting that the operator norm of $(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}$ is almost surely bounded for large n.

Error term e_{21}

We bound the error term e_{21} as follows:

$$
\begin{aligned}
|e_{21}| &= \left|\frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\big)\Sigma\Big]\left(\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}-\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right)\right| \\
&\leq C\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}-\frac{1}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right| \\
&= \frac{C}{n}\left|\sum_{i=1}^{n}\frac{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma\big]/n\big)^2-\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}{\big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^{+}\Sigma\big]/n\big)^2\big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma\big]/n\big)^2}\right|
\end{aligned}
$$

$$
\begin{aligned}
&\le \frac{C}{n} \sum_{i=1}^{n} \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n\Big)^{2} - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]/n\Big)^{2} \right| \\
&\le \frac{C}{n} \sum_{i=1}^{n} \left| \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n - \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]/n \right| \left| 2 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]/n \right| \\
&\le \frac{C}{n} \sum_{i=1}^{n} \left| \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n - \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]/n \right| \\
&\le \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the final convergence follows by noting that

$$
(\widehat{\Sigma} + \lambda I_p)^{+} - (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} = - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T / n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n},
$$

which after multiplying by Σ, taking the trace, and normalizing by n gives

$$
\begin{aligned}
\left| \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n - \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]/n \right| &= \frac{1}{n} \left| \frac{\operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T / n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\big]}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n} \right| \\
&= \frac{1}{n} \left| \frac{x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n} \right| \\
&\le \frac{C}{n},
\end{aligned}
$$

where the last bound follows by noting that operator norm of $(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma$ is almost surely bounded for large n.

Error term e_{22}

We bound the error term e_{22} as follows:

$$
\begin{aligned}
|e_{22}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big)\Sigma\Big] - \operatorname{tr}\Big[(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}) B (I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+})\Sigma\Big]}{\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\big]/n\Big)^{2}} \right| \\
&\le \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big)\Sigma\Big] - \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Sigma\Big] \right| \\
&\le \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big)\Sigma\Big] - \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Sigma\Big] \right| \\
&\quad + \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Sigma\Big] - \operatorname{tr}\Big[\Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big) B \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Sigma\Big] \right| \\
&\le \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\Big[\Sigma\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) B \Big\{\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) - \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Big\}\Big] \right| \\
&\quad + \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\Big[\Big\{\Big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Big) - \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Big\} B \Big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\Big)\Sigma\Big] \right| \\
&\le \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the last inequality bound follows by noting that

$$
\begin{aligned}
&\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \\
&= (\widehat{\Sigma}_{-i} + x_i x_i^T/n)(\widehat{\Sigma}_{-i} + x_i x_i^T/n + \lambda I_p)^+ - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \\
&= (\widehat{\Sigma}_{-i} + x_i x_i^T/n) \left((\widehat{\Sigma}_{-i} + \lambda I_p)^+ - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + \frac{1}{n} x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i} \right) - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \\
&= \frac{x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= \frac{\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n},
\end{aligned}
$$

which after multiplying by $\Sigma(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+)B$ and taking the trace can be bounded as follows:

$$
\begin{aligned}
&\left| \operatorname{tr}\left[\Sigma\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \Big\{ \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Big\} \right] \right| \\
&= \left| \frac{\operatorname{tr}\left[\Sigma\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \right]}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \right| \\
&= \frac{1}{n} \left| \frac{x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \right| \\
&\le \frac{C}{n},
\end{aligned}
$$

where the last bound follows by noting that the matrix $(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+)\Sigma$ has almost surely bounded trace norm for large n (since trace norm of B is bounded and the operator norm of the remaining matrix component is almost surely bounded for large n). The second term can be bounded analogously.

A.2 Proofs related to Theorem 1.4.2

A.2.1 Proof of Lemma 1.5.6

We start by writing the leave-one-out risk estimate $\operatorname{loo}(\lambda)$ from (1.4) as

$$
\operatorname{loo}(\lambda) = y^T (I_n - L_\lambda)^2 D_\lambda^{-2} y/n,
$$

where L_λ is the ridge smoothing matrix and $D_\lambda \in \mathbb{R}^{n \times n}$ is a diagonal matrix with entries $1 - [L_\lambda]_{ii}$ for $i = 1, \dots, n$. Under proportional asymptotic limit, we show below that for any $\lambda \in (\lambda_{\min}, \infty)$,

$$
\operatorname{loo}(\lambda) - y^T (I_n - L_\lambda)^2 \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0, \tag{A.20}
$$

which after substituting back for L_λ proves the desired convergence.

Observe that for any $i = 1, \dots, n$,

$$
\begin{aligned}
[D_\lambda^{-1}]_{ii} = \frac{1}{1 - [L_\lambda]_{ii}} &= \frac{1}{1 - \big[X(X^T X/n + \lambda I_p)^+ X^T/n\big]_{ii}} \\
&= \frac{1}{1 - x_i^T/\sqrt{n}(X^T X/n + \lambda I_p)^+ x_i/\sqrt{n}}.
\end{aligned}
$$

Denoting X^TX/n by $\widehat{\Sigma}$ and using the Woodbury matrix identity as explained in the proof of Lemma A.3.1, we have that

$$\frac{1}{1 - x_i^T(\widehat{\Sigma} + \lambda I_p)^+ x_i/n} = 1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n.$$

The diagonal entries of the matrix D_λ^{-1} are thus $1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n$ for $i = 1, \dots, n$.

We proceed to bound the difference in the two quantities of (A.20) as follows:

$$\begin{aligned}
&\left|\operatorname{loo}(\lambda) - y^T(I_n - L_\lambda)^2\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2 y/n\right| \\
&= \left|y^T(I_n - L_\lambda)^2 D_\lambda^{-2} y/n - y^T(I_n - L_\lambda)^2\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2 y/n\right| \\
&\le y^T(I_n - L_\lambda)^2 y/n \max_{i=1,\dots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2\right| \\
&\le C \max_{i=1,\dots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2\right|,
\end{aligned}$$

where the bound in the last inequality holds almost surely for sufficiently large n by noting that $y^T(I_n - L_\lambda)^2 y/n$ is almost surely bounded for sufficiently large n as explained in the proof of Theorem 1.4.1 Note that we do not require that the response y is well-specified. Finally, similar to the proof of Lemma 1.5.3, we decompose the error as

$$\max_{i=1,\dots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2\right| \le \xi_1 + \xi_2,$$

where the error terms ξ_1 and ξ_2 are defined as follows:

$$\xi_1 := \max_{i=1,\dots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n\Big)^2\right|, \tag{A.21}$$

$$\xi_2 := \max_{i=1,\dots,n}\left|\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n\Big)^2 - \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2\right|. \tag{A.22}$$

Both of the error terms approach 0 under proportional asymptotic limit using the final parts of the arguments used for e_{12} and e_{21} in the proof of Lemma 1.5.3.

A.2.2 Completing the proof of Theorem 1.4.2

Case when $\lambda \neq 0$. Recall from (A.4) that the GCV risk estimate $\operatorname{gcv}(\lambda)$ in this case can be expressed as

$$\operatorname{gcv}(\lambda) = \frac{y^T(I_n - L_\lambda)^2 y/n}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\Big)^2}.$$

On the other hand, from Lemma 1.5.6, under proportional asymptotics we have that

$$\operatorname{loo}(\lambda) - y^T(I_n - L_\lambda)^2\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0.$$

The result then follows by noting that

$$\begin{aligned}
&\left|y^T(I_n - L_\lambda)^2\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2 y/n - \operatorname{gcv}(\lambda)\right| \\
&= \left|y^T(I_n - L_\lambda)^2\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n\Big)^2 y/n - \frac{y^T(I_n - L_\lambda)^2 y/n}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\Big)^2}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq y^T (I_n - L_\lambda)^2 y/n \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big)^2 - \frac{1}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)^2} \right| \\
&\leq C \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big)^2 - \frac{1}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)^2} \right| \\
&\leq C \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big) - \frac{1}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)} \right| \\
&\qquad \cdot \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big) + \frac{1}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)} \right| \\
&\leq C \left| \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big) - \frac{1}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)} \right| \\
&\xrightarrow{\text{a.s.}} 0
\end{aligned}
$$

under proportional asymptotics using the first part of Lemma A.3.1. Note that the bound in the second inequality again follows from the fact that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n, and the operator norm of $I_n - L_\lambda$ is bounded almost surely for large n for $\lambda \in (\lambda_{\min}, \infty)$.

Limiting case when $\lambda = 0$ Similar to the proofs of Lemma 1.5.3 and Lemma 1.5.4, to handle the case when $\lambda = 0$, we observe that for $\lambda \neq 0$, we can extract a factor of λ^2 from $(I_n - L_\lambda)^2$ and absorb into $\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big)^2$ and take $\lambda \to 0$ to write the limiting LOOCV risk estimate under proportional asymptotics as

$$
\operatorname{loo}(0) - y^T (XX^T/n)^{+2} \Big(\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\big]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0,
$$

while the limiting GCV estimate is given by

$$
\operatorname{gcv}(0) = \frac{y^T (XX^T/n)^{+2} y/n}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2}.
$$

As above, we can then bound the difference to get

$$
\begin{aligned}
&\left| y^T (XX^T/n)^{+2} \Big(\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\Big]/n\Big)^2 y/n - \frac{y^T (XX^T/n)^{+2} y/n}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2} \right| \\
&\leq C \left| \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\Big]/n - \frac{1}{\operatorname{tr}[\widehat{\Sigma}^+]/n} \right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the convergence follows from the second part of Lemma A.3.1.

Putting things together, this establishes the almost sure pointwise convergence of $\operatorname{loo}(\lambda)$ to $\operatorname{gcv}(\lambda)$. To show uniform convergence and the convergence of tuned risks, we similarly bound the estimate $\operatorname{loo}(\lambda)$ and its derivative as a function of λ to establish equicontinuity as done in the proof of Theorem 1.4.1. We omit the details due to similarity.

A.3 Auxiliary lemmas

In this section, we state and prove auxiliary lemmas that we often make use of in other proofs. Note that Lemma 1.5.5 is a special case of Lemma 1.5.3 and its proof follows analogous steps as the proof of Lemma 1.5.3 in Appendix A.1.3 and is omitted.

Lemma A.3.1 (Basic GCV denominator lemma)**.** *Under Assumption 1.2 and Assumption 1.3, for* $\lambda \in (\lambda_{\min}, \infty) \setminus \{0\}$,

$$1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n - \frac{1}{1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\right]/n} \xrightarrow{\text{a.s.}} 0 \tag{A.23}$$

as $n, p \to \infty$ *with* $p/n \to \gamma \in (0, \infty)$*. In the case when* $\lambda = 0$*,*

$$\operatorname{tr}\left[(I_p - \widehat{\Sigma}^{+}\widehat{\Sigma})\Sigma\right]/n - \frac{1}{\operatorname{tr}\left[\widehat{\Sigma}^{+}\right]/n} \xrightarrow{\text{a.s.}} 0 \tag{A.24}$$

as $n, p \to \infty$ *with* $p/n \to \gamma \in (0, \infty)$*.*

Proof. We start with the the GCV denominator (the denominator of the second term of (A.23)) and establish that under proportional asymptotics

$$1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\right]/n - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n} \xrightarrow{\text{a.s.}} 0.$$

To that end, we use the standard leave-one-out trick to break the trace functional $1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\right]/n$ into random quadratic forms where the point of evaluation is independent of the inner matrix as follows:

$$\begin{aligned}
1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\widehat{\Sigma}\right]/n &= 1 - \frac{1}{n}\operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\sum_{i=1}^{n} x_i x_i^T/n\right] \\
&= 1 - \frac{1}{n}\sum_{i=1}^{n}\operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+} x_i x_i^T/n\right] \\
&= 1 - \frac{1}{n}\sum_{i=1}^{n} x_i^T(\widehat{\Sigma} + \lambda I_p)^{+} x_i/n \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - x_i^T(\widehat{\Sigma} + \lambda I_p)^{+} x_i/n\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n}.
\end{aligned}$$

Here the last equality follows from the following simplification using the Sherman-Morrison-Woodbury formula with Moore-Penrose inverse (Meyer, 1973):

$$\begin{aligned}
&1 - x_i^T(\widehat{\Sigma} + \lambda I_p)^{+} x_i/n \\
&= 1 - x_i^T\left(\widehat{\Sigma}_{-i} + \lambda I_p + x_i x_i^T/n\right)^{+} x_i/n \\
&= 1 - x_i^T\left((\widehat{\Sigma}_{-i} + \lambda I_p)^{+} - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n}\right) x_i/n \\
&= 1 - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n + x_i^T\frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} x_i/n \\
&= 1 - \frac{x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n x_i^T(\widehat{\Sigma} + \lambda I_p)^{+} x_i/n + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n}
\end{aligned}$$

$$
\begin{aligned}
&= 1 - \frac{x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= \frac{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}.
\end{aligned}
$$

We now break the error in (A.23) as

$$
\begin{aligned}
1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n &- \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n} \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n} \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n}\right) \\
&= \delta_1 + \delta_2,
\end{aligned}
$$

where the error terms δ_1 and δ_2 are defined as follows:

$$
\delta_1 := \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\right]/n}\right),
$$

$$
\delta_2 := \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\right]/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n}\right),
$$

In Appendix A.3.1, we show that both the error terms δ_1 and δ_2 almost surely approach 0 under proportional asymptotics for $\lambda \in (\lambda_{\min}, \infty)$ under Assumption 1.2 and Assumption 1.3.

We now finish the final step by considering the two cases of $\lambda \neq 0$ and $\lambda = 0$.

Case when $\lambda \neq 0$. We so far have that

$$
1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n} \xrightarrow{\text{a.s.}} 0,
$$

which we can rewrite as

$$
\left(1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\right)\left(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\right) - 1 \xrightarrow{\text{a.s.}} 0.
$$

When $\lambda \neq 0$, the GCV denominator $1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n \neq 0$, and we can safely take the inverse to get

$$
1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n - \frac{1}{1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n} \xrightarrow{\text{a.s.}} 0
$$

under proportional asymptotic limit as desired.

Limiting case when $\lambda = 0$. In this case, $1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n$ can be zero (in particular, it is zero when $p \geq n$ and X has rank n). As before, we start with $\lambda \neq 0$ and using Lemma A.3.2, express

$$
1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n = \lambda \operatorname{tr}\left[(XX^T/n + \lambda I_n)^+\right]/n,
$$

along with

$$
\lambda \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right] = \operatorname{tr}\left[\left(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\right)\Sigma\right]/n.
$$

This allows us to move λ across to write

$$\Big(\operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n\Big)\Big(\lambda + \operatorname{tr}\big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\big]/n\Big) - 1 \xrightarrow{\text{a.s.}} 0.$$

Sending $\lambda \to 0$, writing $\operatorname{tr}[(XX^T/n^+)]/n = \operatorname{tr}[\widehat{\Sigma}^+]/n$, and inverting safely, we have

$$\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\big]/n - \frac{1}{\operatorname{tr}[\widehat{\Sigma}^+]/n} \xrightarrow{\text{a.s.}} 0$$

under proportional asymptotic limit as desired. □

Lemma A.3.2 (Gram and sample covariance matrix simplifications). *Suppose $X^TX/n+\lambda I_p$ and $XX^T/n+\lambda I_n$ are invertible. Then it holds that*

$$I_n - X(X^TX/n + \lambda I_p)^+X^T/n = \lambda(XX^T/n + \lambda I_n)^+,$$

$$I_p - \left(X^TX/n + \lambda I_p\right)^+ X^TX/n = \lambda(X^TX/n + \lambda I_p)^+.$$

Proof. Recall the Woodbury matrix identity

$$A^{-1} - A^{-1}U(VA^{-1}U + C^{-1})^{-1}VA^{-1} = (UCV + A)^{-1}.$$

Letting $A = I_n$, $U = X/\sqrt{n}$, $C = 1/\lambda I_p$, $V = X^T/\sqrt{n}$, we get

$$\begin{aligned} I_n - X(X^TX/n + \lambda I_p)^{-1}X^T/n &= (X/\sqrt{n}\, 1/\lambda I_p\, X^T/\sqrt{n} + I_n)^{-1} \\ &= \lambda(XX^T/n + \lambda I_n)^{-1}. \end{aligned}$$

On the other hand, letting $A = I_p$, $U = I_p$, $V = X^TX/n$, $C = 1/\lambda I_p$, we get

$$\begin{aligned} I_p - \left(X^TX/n + \lambda I_p\right)^{-1} X^TX/n &= \left(1/\lambda I_p\, X^TX/n + I_p\right)^{-1} \\ &= \lambda\left(X^TX/n + \lambda I_p\right)^{-1}. \end{aligned}$$

□

A.3.1 Error terms in the proof of Lemma A.3.1

Below we show that for $\lambda \in (\lambda_{\min}, \infty)$ both the error terms δ_1 and δ_2 almost surely approach 0 as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. The arguments mirror parts of the error analysis for terms e_{12} and e_{21} in Appendix A.1.6.

Error term δ_1

$$\begin{aligned} |\delta_1| &= \left|\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+x_i/n} - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n}\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n \frac{\operatorname{tr}\big[\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+x_i/n}{\big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+x_i/n\big)\big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n\big)}\right| \\ &\le C\left|\frac{1}{n}\sum_{i=1}^n \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+x_i/n\right| \\ &\le C \max_{i=1,\dots,n}\left|\operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\big]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+x_i/n\right| \\ &\xrightarrow{\text{a.s.}} 0, \end{aligned}$$

where the final convergence follows from using Lemma A.4.4 as argued for the suberror term e_{12} in Appendix A.1.6.

Error term δ_2

$$
\begin{aligned}
|\delta_2| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n} \right| \\
&= \frac{1}{n} \left| \sum_{i=1}^{n} \frac{\operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n - \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n}{\left(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\right)\left(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\right)} \right| \\
&\leq \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}\right]/n - \operatorname{tr}\left[\Sigma(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\right]/n \right| \\
&\leq \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the last inequality follows analogous simplification as done for the suberror term e_{21} in Appendix A.1.6.

A.4 Useful results

The following lemma is a standard concentration of linear combination of i.i.d. entries.

Lemma A.4.1 (Concentration of linear form with independent components)**.** *Let ε be a random vector in R^n that satisfy conditions of error vector in Assumption 1.1. Let b_n be a sequence of random vectors in $\mathbb{R}^n$ independent of ε such that $\sup_n \|b_n\|^2/n < \infty$ almost surely. Then as $n \to \infty$,*

$$b_n^T \varepsilon / n \xrightarrow{\text{a.s.}} 0.$$

The following lemma is adapted from Dobriban and Wager (2018, Lemma 7.6).

Lemma A.4.2 (Concentration of quadratic form with independent components)**.** *Let $\varepsilon \in \mathbb{R}^n$ be a random vector that satisfy conditions of error vector in Assumption 1.1. Let D_n be a sequence of random matrices in $\mathbb{R}^{n\times n}$ that are independent of ε and have operator norm uniformly bounded in n. Then as $n \to \infty$,*

$$\varepsilon^T D_n \varepsilon / n - \sigma^2 \operatorname{tr}[D_n]/n \xrightarrow{\text{a.s.}} 0.$$

The following lemma is adapted from an argument in Hastie et al. (2022, Theorem 7) using union bound along with a lemma from Bai and Silverstein (2010, Lemma B.26).

Lemma A.4.3 (Concentration of maximum of quadratic forms with independent components)**.** *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^p$ that satisfy Assumption 1.2 and Assumption 1.3. Let $G_1, \ldots, G_n$ be random matrices in $\mathbb{R}^{p\times p}$ such that G_i is independent of x_i (but may depend on all of X_{-i}) and have operator norm uniformly bounded in n. Then as $n \to \infty$,*

$$\max_{i=1,\ldots,n} \left| x_i^T G_i x_i / n - \operatorname{tr}[G_i \Sigma]/n \right| \xrightarrow{\text{a.s.}} 0.$$

The following lemma is adapted from Rubio and Mestre (2011, Lemma 4).

Lemma A.4.4 (Concentration of sum of quadratic forms with independent components)**.** *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^p$ that satisfy Assumption 1.2 and Assumption 1.3. Let $H_1, \ldots, H_n$ be random matrices in $\mathbb{R}^{p\times p}$ such that H_i is independent of x_i (but may depend on all of X_{-i}) that have trace norm uniformly bounded in n. Then as $n \to \infty$,*

$$\left| \sum_{i=1}^{n} x_i^T H_i x_i / n - \operatorname{tr}[H_i \Sigma]/n \right| \xrightarrow{\text{a.s.}} 0.$$

Appendix B

Supplement for Chapter 2

This supplement contains additional details, proofs, and numerical experiments for Chapter 2. The content of the supplement is organized as follows.

- In Appendices B.1 to B.3, we first provide proofs related to Theorems 2.4.1 to 2.4.3, respectively, along with supporting lemmas used in the process, as they constitute building blocks for other theoretical results.
- In Appendix B.4, we then present proof of Theorem 2.3.1.
- In Appendix B.5, we present proofs related to Theorem 2.5.1, along with further theoretical results related to quantile estimation.
- In Appendix B.7, we provide additional numerical results and experimental details
- In Appendix B.6, we collect statements of supplementary results from the literature that are used in various proofs throughout the supplement.

A note about constants throughout the supplement: We use the letter C (either standalone or with a subscript such as C_1) to denote a generic constant whose value can change from line to line. Additionally, some of the inequalities only hold almost surely for sufficiently large n. We will sometimes use the term eventually almost surely to indicate such statements.

B.1 Proofs related to Theorem 2.4.1

As suggested in the proof overview in Section 2.4, we will first show the second part of the theorem statement: $\widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, and use it to show the first part: $\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

- To prove $\widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we introduce an intermediate quantity $\widetilde{T}_\lambda$ as in (2.19) and break the difference

  $$T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}} = (T_\lambda - \widetilde{T}_\lambda) + (\widetilde{T}_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}). \tag{B.1}$$

  We will show that both terms in the decomposition (B.1) almost surely vanish. Appendix B.1.1 shows the convergence for the first term, while Appendix B.1.2 shows the convergence for the second term.
- To prove $\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we similarly break the difference

  $$T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}} = (T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}) + (\widehat{T}_\lambda^{\mathrm{loo}} - \widehat{T}_\lambda^{\mathrm{gcv}}). \tag{B.2}$$

  We have already dealt with the first term in the decomposition (B.2) in (B.1). We show the second term almost surely goes to zero in Appendix B.1.3.

We will show the three aforementioned converges first under a slight stronger assumption that the error function t is uniformly continuous. Using a truncation argument, we will then relax them to continuous error functions t in Appendix B.1.4. Let $\omega_t : [0, \infty] \to [0, \infty]$ denote a modulus of continuity of t. Without of loss of generality, we can assume ω_t to be non-decreasing and continuous. Since the error function is assumed to be uniformly continuous, such a modulus exits (see, e.g., Chapter 2 of DeVore and Lorentz, 1993). In addition, let $\overline{\omega}_t$ denote the least concave majorant of ω_t. From DeVore and Lorentz (1993, Lemma 6.1), $\overline{\omega}_t$ is also a modulus of continuity and satisfies $\overline{\omega}_t(r) \le 2\omega_t(r)$ for $r \ge 0$. We will make use of these properties below.

B.1.1 Functional to LOO functional

Towards showing $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$, we begin by manipulating the desired difference using properties of conditional expectation as follows:

$$\begin{aligned}
T_\lambda - \widetilde{T}_\lambda &= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}, x_i, y_i\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big].
\end{aligned}$$

The second equality above uses independence of (y_0, x_0) and (X_{-i}, y_{-i}), while the third equality uses independence of (y_0, x_0), $\widehat{\beta}_{-i,\lambda}$, and (x_i, y_i). We will next show below that under proportional asymptotics absolute value of the right-hand side of the last display almost surely goes to zero; in other words, we will show

$$\left| \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big] \right| \xrightarrow{\text{a.s.}} 0. \tag{B.3}$$

Using the modulus of continuity of t and its least concave majorant, we first bound the summands in (B.3) for $i = 1, \dots, n$ as

$$\begin{aligned}
\big|t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda})\big| &\le \omega_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big) \\
&\le \overline{\omega}_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big).
\end{aligned}$$

We can then bound the summation in (B.3) as

$$\begin{aligned}
\left| \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big] \right| &\le \frac{1}{n}\sum_{i=1}^n \left|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big]\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big|t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] \\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\overline{\omega}_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big) \mid X, y\Big] \\
&\le \frac{1}{n}\sum_{i=1}^n \overline{\omega}_t\Big(\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big]\Big)
\end{aligned}$$

$$
\begin{aligned}
&\le \overline{\omega}_t \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big| \mid X, y \Big] \right) \\
&\le 2\omega_t \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big| \mid X, y \Big] \right).
\end{aligned}
$$

In the above chain of inequalities, the second, forth, and fifth inequalities follow from repeated use of Jensen's inequality (on the absolute value function and the concave majorant function). To finish the proof, we will finally show below that

$$
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big| \mid X, y \Big] \xrightarrow{\text{a.s.}} 0, \tag{B.4}
$$

which along with the continuity of the modulus that vanishes at 0 shows (B.3), leading to the desired conclusion that $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$.

Towards showing (B.4), first note that under Assumption 2.1, we can bound the summands for each $i = 1, \dots, n$ as

$$
\begin{aligned}
\mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big| \mid X, y \Big] &\le \left(\mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big|^2 \mid X, y \Big] \right)^{1/2} \\
&= \left(\mathbb{E}\Big[\big| z_0^\top \Sigma^{1/2} (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big|^2 \mid X, y \Big] \right)^{1/2} \\
&= \left(\mathbb{E}\Big[(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})^\top \Sigma^{1/2} z_0 z_0^\top \Sigma^{1/2} (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \mid X, y \Big] \right)^{1/2} \\
&= \left((\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \Sigma (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \right)^{1/2} \\
&\le \left(r_{\max} (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \right)^{1/2} \\
&= \sqrt{r_{\max}} \big\| (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big\|_2 .
\end{aligned}
$$

The inequality in the first line uses Jensen's inequality (on the square root function), and the inequality in the forth line follows since the maximum eigenvalue of Σ is upper bounded by $r_{\max}$. Hence, overall we can bound the left-hand side of (B.4) by

$$
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \big| \mid X, y \Big] \le \sqrt{r_{\max}} \left(\frac{1}{n} \sum_{i=1}^{n} \big\| \widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda} \big\|_2 \right). \tag{B.5}
$$

We show in Lemma B.1.2 that the term in the parenthesis on the right-hand side of (B.5) almost surely goes to zero under Assumptions 2.1 and 2.2, proving (B.4) and completing the proof.

B.1.2 LOO functional to LOOCV estimator

To show $\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}} \xrightarrow{\text{a.s.}} 0$, we start by breaking the difference into two pieces:

$$
\begin{aligned}
\big| \widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}} \big| &= \left| \widetilde{T}_\lambda - \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) + \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}} \right| \\
&\le \left| \widetilde{T}_\lambda - \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \right| + \left| \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}} \right|.
\end{aligned} \tag{B.6}
$$

In the sequel, we will show that each of two pieces in (B.6) vanishes almost surely under proportional asymptotics.

For the second piece in (B.6), using the modulus of t and its concave majorant, we can bound the difference as

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda^{\mathrm{loo}} \right| &= \left| \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^{\rightarrow} \widehat{\beta}_{-i,\lambda}) - \frac{1}{n} \sum_{i=1}^{n} t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right) \right| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \left| t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right) \right| \\
&\le \frac{1}{n} \sum_{i=1}^{n} \overline{\omega}_t \left(\left| y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right| \right) \\
&\le \overline{\omega}_t \left(\frac{1}{n} \sum_{i=1}^{n} \left| y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right| \right) \\
&\le 2\omega \left(\frac{1}{n} \sum_{i=1}^{n} \left| y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \right| \right), \qquad \text{(B.7)}
\end{aligned}$$

where line four uses Jensen's inequality (on the concave majorant). Note that the above is valid when $1 - [L_\lambda]_{ii} \neq 0$ for any of $i = 1, \dots, n$. For the case of min-norm estimator where $[L_0]_{ii} = 0$, we similarly bound

$$\left| \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,0}) - \widetilde{T}_\lambda^{\mathrm{loo}} \right| \le 2\omega \left(\frac{1}{n} \sum_{i=1}^{n} \left| y_i - x_i^\top \widehat{\beta}_{-i,0} - \frac{[(XX^\top/n)^\dagger]_i}{[(XX^\top/n)^\dagger]_{ii}} \right| \right). \qquad \text{(B.8)}$$

The argument of ω in either cases of (B.7) and (B.8) goes to 0 almost surely, and thus the continuity of ω provides the desired convergence of the second piece in (B.6) It is worth mentioning that the only reason we need to worry about (B.7) and (B.8) is the way we have defined ridge estimator in (2.1) where the leave-one-out estimator $\widehat{\beta}_{-i,\lambda}$ gets a dividing factor of $(n-1)$ instead of n, otherwise these terms would be exactly 0. It is a short straightforward calculation to show however that this does not make a difference as $n \to \infty$.

We now focus on the first piece in the decomposition (B.6). Note that we can express

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda &= \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \bigm| X_{-i}, y_{-i} \big] \\
&= \frac{1}{n} \sum_{i=1}^{n} \Big\{ t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \bigm| X_{-i}, y_{-i} \big] \Big\}. \qquad \text{(B.9)}
\end{aligned}$$

For $i = 1, \dots, n$, let $\mathcal{F}_i$ denote the increasing σ-field generated by $(x_1, y_1), \dots, (x_i, y_i)$. Observe that

$$\Big\{ t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \bigm| X_{-i}, y_{-i} \big] \Big\}_{i=1}^{n}$$

forms a martingale difference array with respect to the filtration $\{\mathcal{F}_i\}_{i=1}^n$. To see this, note that

$$\begin{aligned}
&\mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \bigm| X_{-i}, y_{-i} \big] \Bigm| \mathcal{F}_{i-1} \Big] \\
&= \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \Bigm| \mathcal{F}_{i-1} \Big] - \mathbb{E}\Big[\mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \bigm| X_{-i}, y_{-i} \big] \Bigm| \mathcal{F}_{i-1} \Big] \\
&= \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \Bigm| \mathcal{F}_{i-1} \Big] - \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \Bigm| \mathcal{F}_{i-1} \Big] \\
&= 0,
\end{aligned}$$

where for the second equality we used the tower property of conditional expectation as $\mathcal{F}_{i-1}$ is a subset of the σ-field generated by (X_{-i}, y_{-i}). This observation allows us to use the Burkholder inequality (see Lemma B.6.1 for an exact statement) to bound q-th moment of the difference for $q \ge 2$.

Applying the Burkholder inequality to our martingale sequence, we can bound

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\le C\mathbb{E}\left[\left\{\sum_{i=1}^{n} \mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \,\middle|\, \mathcal{F}_{i-1}\right]\right\}^{q/2}\right] \\
&\quad + C\mathbb{E}\left[\sum_{i=1}^{n} \left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \qquad \text{(B.10)}
\end{aligned}$$

for some constant $C > 0$. We next bound each of the terms in turn. Denote by X_{i+i}^n and y_{i+i}^n dataset consisting of observations $(x_{i+1}, y_{i+1}), \cdots, (x_n, y_n)$.

For the first term, from the law of total expectation observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \,\middle|\, \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \,\middle|\, \mathcal{F}_{i-1}, X_{i+1}^n, y_{i+1}^n\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \,\middle|\, X_{-i}, y_{-i}\right\}\right] \\
&\le 4\mathbb{E}\left[\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^2 \,\Big|\, X_{-i}, y_{-i}\Big]\right],
\end{aligned}$$

where in the last step we used the inequality $\mathbb{E}[|a+b|^2] \le 2\big(\mathbb{E}[|a|^2] + \mathbb{E}[|b|^2]\big)$.

For the second term, similarly note that

$$\begin{aligned}
&\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \,\middle|\, X_{-i}, y_{-i}\right] \\
&\le 2^q\mathbb{E}\left[\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\right|^q \,\middle|\, X_{-i}, y_{-i}\right]\right],
\end{aligned}$$

where the last step follows from using the inequality $\mathbb{E}[|a+b|^q] \le 2^{q-1}\big(\mathbb{E}[|a|^q] + \mathbb{E}[|b|^q]\big)$ for $q > 1$.

In addition, from Jensen's inequality, we have for $q \ge 2$

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^2 \mid X_{-i}, y_{-i}\Big] \le \mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \mid X_{-i}, y_{-i}\Big].$$

Hence, to bound both the terms, it is sufficient to control q-th moment of the functional. From Lemma B.1.1, for $q \le 2 + \min\{\mu/2, \nu/2\}$,

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \mid X_{-i}, y_{-i}\Big] \le \big(C_1 + C_2\big\|\widehat{\beta}_{-i,\lambda}\big\|_2\big)^{2q}$$

for some positive constants C_1 and C_2. Combined Lemma B.1.3 that implies $\|\widehat{\beta}_{-i,\lambda}\|_2 \le C$ almost surely for n large enough under Assumptions 2.1 and 2.2, we have

$$\mathbb{E}\left[\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \mid X_{-i}, y_{-i}\Big]\right] \le C$$

for some constant $C > 0$ and $2 \le q \le 2 + \min\{\mu/2, \nu/2\}$.

Therefore, from (B.10) we can bound q-th moment of normalized sum (B.9) to get

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\le \frac{(nC)^{q/2} + nC}{n^q} \\
&\le C\frac{1}{n^{q/2}} + C\frac{1}{n^{q-1}}.
\end{aligned}$$

Finally, choosing $2 < q \le 2 + \min\{\mu/2, \nu/2\}$ and applying Lemma B.6.7 provides the desired convergence for the first piece in (B.6). This concludes the proof.

B.1.3 LOOCV estimator to GCV estimator

To prove $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{T}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$, we start by bounding the absolute difference of interest by the average of absolute differences for $i = 1, \dots, n$:

$$\begin{aligned}
\big|\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{T}_\lambda^{\mathrm{loo}}\big| &= \left|\frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n}\right) - \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n}\right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right|. \qquad \text{(B.11)}
\end{aligned}$$

We will show below that the right-hand side of the expression (B.11) almost surely goes to zero. As with the proof of $\widetilde{T}_\lambda - \widehat{T}_\lambda \xrightarrow{\text{a.s.}} 0$, we will first assume $L_{ii} \ne 0$ so (B.11) is well defined. We will indicate the changes that we need to make when $L_{ii} = 0$ towards the end of the proof.

Using the modulus of continuity of t and it least concave majorant, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n}\right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| &\le \frac{1}{n}\sum_{i=1}^{n} \omega_t\left(\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le \frac{1}{n}\sum_{i=1}^{n} \overline{\omega}_t\left(\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le \overline{\omega}_t\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le 2\omega_t\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le 2\omega_t\left(\frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i^\top \widehat{\beta}_\lambda\right|\left|\frac{1}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right|\right).
\end{aligned}$$

In the above chain on inequalities, we used Jensen's inequality on the concave majorant $\overline{\omega}_t$ for the third line, and monotonicity of ω_t on the fifth line.

Thus, from continuity of ω_t at 0, we will be done by showing

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i^\top \widehat{\beta}_\lambda\right|\left|\frac{1}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right| \xrightarrow{\text{a.s.}} 0. \qquad \text{(B.12)}$$

To build towards proving (B.12), let us denote by $r \in \mathbb{R}^n$ the vector of residuals $y_i - x_i^\top \widehat{\beta}^\lambda$ and by $d \in \mathbb{R}^n$ the vector of differences $(1 - \operatorname{tr}[L_\lambda]/n)^{-1} - (1 - [L_\lambda]_{ii})^{-1}$. Observe that

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i^\top \widehat{\beta}_\lambda\right|\left|\frac{1}{1 - \operatorname{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right| = \frac{1}{n} r^\top d$$

$$
\begin{aligned}
&\leq \frac{1}{n}\|r\|_1\|d\|_\infty \\
&\leq \frac{1}{\sqrt{n}}\|r\|_2\|d\|_\infty,
\end{aligned}
$$

where we used Hölder's inequality in the second line and the the bound $\|a\|_1 \leq \sqrt{n}\|a\|_2$ for any $a \in \mathbb{R}^n$ in the last line. Since $r = (I - L_\lambda)y$, and the operator norm of $I - L_\lambda$ is bounded for $\lambda \in (\lambda_{\min}, 0)$ and $\|y\|_2/\sqrt{n}$ is almost surely bounded for sufficiently large n from the strong law of large numbers under Assumption 2.2, we have that $\|r\|_2/\sqrt{n}$ is eventually almost surely bounded. We now show in the sequel that $\|d\|_\infty \xrightarrow{\text{a.s.}} 0$ leading to the desired conclusion.

First for each $i = 1, \ldots, n$, by adding and subtracting $1 + \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n$, and $\operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n$, we decompose the difference

$$
\begin{aligned}
&\left|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \frac{1}{1-[L_\lambda]_{ii}}\right| \\
&= \Bigg|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \left(1 + \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\right) + \operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n - \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n \\
&\qquad + \left(1 + \operatorname{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\right) - \frac{1}{1-[L_\lambda]_{ii}}\Bigg| \\
&\leq \left|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \left(1 - \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\right)\right| \\
&\qquad + \left|\operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \operatorname{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\right| \\
&\qquad\quad + \left|\left(1 - \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\right) - \frac{1}{1-[L_\lambda]_{ii}}\right|.
\end{aligned}
$$

This lets us decompose

$$
\begin{aligned}
\|d\|_\infty &= \max_{1\leq i\leq n}\left|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \frac{1}{1-[L_\lambda]_{ii}}\right| \\
&\leq \left|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \left(1 - \operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n\right)\right| \\
&\qquad + \max_{1\leq i\leq n}\left|\operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n - \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n\right| \\
&\qquad\quad + \max_{1\leq i\leq n}\left|\left(1 - \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n\right) - \frac{1}{1-[L_\lambda]_{ii}}\right|.
\end{aligned}
$$

Finally, we verify that each of the term in the decomposition almost surely vanishes. Using the $\lambda \neq 0$ case of Lemma B.6.4, we have for the first term

$$
\left|\frac{1}{1-\operatorname{tr}[L_\lambda]/n} - \left(1 - \operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n\right)\right| \xrightarrow{\text{a.s.}} 0.
$$

For the second term, following the proof of Lemma B.6.4, for $i = 1, \ldots, n$ we can bound

$$
\left|\operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n - \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n\right| \leq C/n,
$$

almost surely for sufficiently large n. This uses the Sherman-Morrison-Woodbury formula with Moore-Penrose inverse to express the difference

$$
(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger = -\frac{(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i x_i^\top/n(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger}{1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i}. \tag{B.13}
$$

The second term thus almost surely goes to zero. For the third term, note that from using the Sherman-Morrison-Woodbury formula again, we can simplify

$$
1 - [L_\lambda]_{ii} = 1 - x_i^\top (X^\top X/n + \lambda I)^\dagger x_i/n
$$

$$
\begin{aligned}
&= 1 - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I + x_i x_i^\top/n)^\dagger x_i/n \\
&= \frac{1}{1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i/n}.
\end{aligned}
$$

Therefore, for $q \geq 2$, we can now proceed to bound the q-th moment of the second term as

$$
\begin{aligned}
&\mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|1 + \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - \frac{1}{1-[L_\lambda]_{ii}}\right|\right\}^q\right] \\
&= \mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|1 + \operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - \big(1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\big)\right|\right\}^q\right] \\
&= \mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|\operatorname{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\right|\right\}^q\right] \\
&\leq \max_{1\leq i\leq n}\mathbb{E}\left[\left\{\left|\operatorname{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\right|\right\}^q\right] \\
&\leq n\mathbb{E}\left[\left\{\operatorname{tr}\left[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger \Sigma\right]/n - x_j^\top (X_{-j}^\top X_{-j}/n + \lambda I)^\dagger x_j/n\right\}^q\right]
\end{aligned}
$$

for any $j = 1, \ldots, n$. Note that the last line follows from noting that $\operatorname{tr}\big[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger \Sigma\big]/n$, and $x_i^\top \left(X_{-i}^\top X_{-i}/n + \lambda I\right)^\dagger x_i$ are identically distributed for $i = 1, \ldots, n$. Since

$$
\operatorname{tr}\big[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger\big]/n \leq C/n
$$

almost surely for sufficiently large n, using Lemma B.6.3, the above quantity is of order $O(n/n^q)$. Choosing $q > 2$ and applying Lemma B.6.7 thus provides the desired almost sure convergence.

The above argument assumed that $L_{ii} \neq 0$. For the case of min-norm interpolator when $L_{ii} = 0$, we follow exactly similar steps as above using the modified errors defined in (2.13) and (2.14). (For more details on the λ cancellation for modified errors, see the proof of $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$ in Appendix B.1.4.) This reduces to showing

$$
\frac{1}{n}\sum_{i=1}^n \left|[(XX^\top/n)^\dagger y]_i\right| \left|\frac{1}{\operatorname{tr}[(XX^\top/n)^\dagger]/n} - \frac{1}{[(XX^\top/n)^\dagger]_{ii}}\right| \xrightarrow{\text{a.s.}} 0. \tag{B.14}
$$

The same way we argued the almost sure boundedness of $\|r\|_2$, we can bound the norm of modified error vector $(XX^\top/n)^\dagger y$ as shown in Appendix B.1.4. Finally, analogous to the argument used to bound d, we can now use the case of $\lambda = 0$ equivalence in Lemma B.6.4 for the difference vector in the modified errors of (B.14). This takes care of both the cases and concludes the proof.

B.1.4 Truncation arguments

We established the converges in Appendices B.1.1 to B.1.3 under the the assumption that the error function t is uniformly continuous. In this section, we relax this assumption to t being only continuous by a truncation argument. Let $\mathbb{I}\{\mathcal{A}\}$ denote the indicator function for set $\mathcal{A}$.

Let t be a continuous error function. Define $w : \mathbb{R} \to \mathbb{R}$ to be the truncation of t on the compact interval $[-n, n]$, in other words, $w(r) = t(r)\mathbb{I}\{|r| \leq n\}$. Let W_λ denote the linear functional (2.5) corresponding to the error function w, and let $\widetilde{W}_\lambda$ be the intermediate averaged LOO functional defined analogously to (2.19) using w. Let $\widehat{W}_\lambda^{\mathrm{gcv}}$ and $\widehat{W}_\lambda^{\mathrm{loo}}$ denote the plug-in GCV and LOOCV estimators associated with w. The arguments in Appendices B.1.1 to B.1.3 establish $W_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$, $\widetilde{W}_\lambda - \widehat{W}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$, and $\widehat{W}_\lambda^{\mathrm{loo}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$. We will now show that $T_\lambda - W_\lambda \xrightarrow{\text{a.s.}} 0$, $\widetilde{T}_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$, $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$, $\widehat{T}_\lambda^{\mathrm{loo}} - \widehat{W}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$ to finish the proof of Theorem 2.4.1. Since the proof of LOOCV mirrors that for GCV, we will only show the argument for GCV to avoid repetition.

Showing $T_\lambda - W_\lambda \xrightarrow{\text{a.s.}} 0$.

We can bound the absolute difference as follows:

$$\begin{aligned}
|T_\lambda - W_\lambda| &= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \mathbb{E}\big[w(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big]\Big| \\
&= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - w(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big]\Big| \\
&= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda)\mathbb{I}\{|y_0 - x_0^\top \widehat{\beta}| > n\} \mid X, y\big]\Big| \\
&\le \sqrt{\mathbb{E}\big[|t(y_0 - x_0^\top \widehat{\beta}_\lambda)|^2 \mid X, y\big]}\sqrt{\mathbb{P}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| > n \mid X, y\big]} \\
&\le C\sqrt{\mathbb{P}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| > n \mid X, y\big]} \\
&\le C\sqrt{\frac{\mathbb{E}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda|^2 \mid X, y\big]}{n^2}} \\
&\le \frac{C}{n} \to 0,
\end{aligned}$$

where the third line uses the Cauchy-Schwarz inequality, the fourth line uses Lemmas B.1.1 and B.1.3 with $q = 2$, the fifth line uses Chebychev's inequality, and the last line again uses Lemmas B.1.1 and B.1.3 with t as the identity function and $q = 2$.

Showing $\widetilde{T}_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$.

We can bound the absolute difference as follows:

$$\begin{aligned}
\Big|\widetilde{T}_\lambda - \widetilde{W}_\lambda\Big| &= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[w(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - w(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right| \\
&\le \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\mathbb{I}\big\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\big\} \;\Big|\; X_{-i}, y_{-i}\Big]\right| \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}\Big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \;\Big|\; X_{-i}, y_{-i}\Big]}\sqrt{\mathbb{P}\Big\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n \;\Big|\; X_{-i}, y_{-i}\Big\}} \\
&\le \frac{1}{n}\sum_{i=i}^{n}\sqrt{\mathbb{E}\big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \mid X_{-i}, y_{-i}\big]}\sqrt{\mathbb{P}\left\{\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n \;\Big|\; X, y\right\}} \\
&\le \left|\frac{1}{n}\sum_{i=i}^{n}\sqrt{\mathbb{E}\big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \mid X_{-i}, y_{-i}\big]}\right|\sqrt{\mathbb{P}\left\{\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\}} \\
&\le C\sqrt{\mathbb{P}\left\{\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\}}.
\end{aligned}$$

Above, line four uses the Cauchy-Schwarz inequality, line five uses the fact that the event $|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n$ for any $i = 1, \dots, n$ is contained inside the event $\max_{j=1}^{n}|y_j - x_j^\top \widehat{\beta}_{-j,\lambda}| > n$, and the last line follows from the q-th moment control as done in Appendix B.1.2 with $q = 2$. It therefore suffices to bound the probability of the event $\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n$ which we do below.

Starting with union bound, we have that

$$\mathbb{P}\left\{\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\} \le \sum_{i=1}^{n}\mathbb{P}\left\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\}$$

$$
\begin{aligned}
&\le \sum_{i=1}^{n} \frac{\mathbb{E}\big[|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}|^2\big]}{n^2} \\
&\le \sum_{i=1}^{n} \frac{C}{n^2} \\
&\le \frac{C}{n} \to 0.
\end{aligned}
$$

Showing $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$.

By following similar argument used to bound $\big|\widetilde{T}_\lambda - \widetilde{W}_\lambda\big|$, it suffices to show that

$$
\mathbb{P}\left\{ \max_{j=1}^{n} \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} > n \right\} \to 0.
$$

Using the union bound, it is thus enough to show that almost surely

$$
\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \right)^2 \le C.
$$

Note that this is valid when $\lambda \neq 0$. To cover the case of min-norm interpolator, we start by rewriting the residuals in an alternate form as follows:

$$
\begin{aligned}
y_i - x_i^\top \widehat{\beta}_\lambda &= y_i - x_i^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n \\
&= y_i - [X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\
&= [y - X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\
&= [(I - X^\top (X^\top X/n + \lambda I)^\dagger X/n) y]_i \\
&= \lambda [(X X^\top/n + \lambda I)^\dagger y]_i
\end{aligned} \tag{B.15}
$$

Similarly, we rewrite the denominator of GCV using

$$
\begin{aligned}
1 - \mathrm{tr}[L_\lambda]/n &= 1 - \mathrm{tr}[X (X X^\top/n + \lambda I)^\dagger X^\top]/n \\
&= \mathrm{tr}[I - X (X X^\top/n + \lambda I)^\dagger X^\top]/n \\
&= \lambda \, \mathrm{tr}[(X X^\top/n + \lambda I)^\dagger]/n.
\end{aligned} \tag{B.16}
$$

This lets us rewrite the invidual GCV reweighted errors as

$$
\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} = \frac{\lambda [(X X^\top/n + \lambda I)^\dagger y]_i}{\lambda \, \mathrm{tr}[(X X^\top/n + \lambda I)^\dagger]/n} = \frac{[(X X^\top/n + \lambda I)^\dagger y]_i}{\mathrm{tr}[(X X^\top/n + \lambda I)^\dagger]/n}.
$$

Thus, we can now bound

$$
\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \right)^2 &= \frac{\big\|(X X^\top/n + \lambda I)^\dagger y\big\|_2^2/n}{\big(\mathrm{tr}[(X X^\top/n + \lambda I)^\dagger]/n\big)^2} \\
&\le \frac{\big\|(X X^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}^2 \big\|y\big\|_2^2/n}{\big(\mathrm{tr}[(X X^\top/n + \lambda I)^\dagger]/n\big)^2}.
\end{aligned}
$$

Each term in the above ratio is almost surely bounded for sufficiently large n under Assumption 2.1 and Assumption 2.2 as explained in the proof of Lemma B.1.3. This finishes the argument.

B.1.5 Auxiliary lemmas

In this section, we gather supporting lemmas used in the proofs in Appendices B.1.1 to B.1.3, along with their proofs.

Lemma B.1.1 (Bounding conditional q-th moment of the i-th LOO residual)**.** *Suppose Assumptions 2.1 and 2.2 hold, and the error function t satisfies Assumption 2.3. Then, for $q \le \min\{\mu/2, \nu/2\}$ and $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \;\big| X_{-i}, y_{-i}\;\Big] \le \big(C_1 + C_2\, \|\widehat{\beta}_{-i,\lambda}\|_2\big)^{2q}$$

for some positive constants C_1 and C_2.

Proof. Note that under Assumption 2.3, $\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \le a\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} + b\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^q + c$ for some positive constants a, b, c. Because $\mathbb{E}\big[Z^{q_l}\big] \le \mathbb{E}\big[Z^{q_h}\big]^{q_l/q_h}$ for $q_l \le q_h$ from Jensen's inequality, it suffices to bound $\mathbb{E}\big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\big]$, which we do below.

From the triangle inequality for the conditional L_q norm, observe that

$$\begin{aligned}\mathbb{E}\Big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} &\le \mathbb{E}\Big[\big|y_i\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} + \mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q}\\ &\le \mathbb{E}\Big[\big|y_i\big|^{2q}\Big]^{1/2q} + \mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q}.\end{aligned}$$

The first term is bounded for $q \le 2 + \mu/2$ under Assumption 2.2. For the second term, start by writing

$$\mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big] = \mathbb{E}\Big[\big|z_i^\top \Sigma^{1/2} \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big].$$

Note that conditional on X_{-i} and y_{-i}, $\Sigma^{1/2}\widehat{\beta}_{-i,\lambda}$ is a fixed vector in $\mathbb{R}^p$. For $q \le 2 + \nu/2$, Lemma B.6.2 then provides

$$\mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} \le C\|\Sigma^{1/2}\widehat{\beta}_{-i,\lambda}\|_2 \le C\sqrt{r_{\max}}\|\widehat{\beta}_{-i,\lambda}\|_2,$$

where the last inequality follows since the maximum eigenvalue of Σ is bounded by $r_{\max}$. Therefore, for $q \le 2 + \min\{\mu/2, \nu/2\}$, we get

$$\mathbb{E}\Big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big] \le \big(C_1 + C_2\|\widehat{\beta}_{-i,\lambda}\|_2\big)^{2q}$$

for some positive constants C_1 and C_2 as desired. This completes the proof. □

Lemma B.1.2 (Bounding norm of the difference of leave-one-out ridge estimators)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\frac{1}{n}\sum_{i=1}^{n} \big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

Proof. For each $i = 1, \dots, n$, we start by breaking the difference

$$\begin{aligned}\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda} &= (X^\top X/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X_i^\top y_{-i}/(n-1)\\ &= (X^\top X/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X^\top y/n\\ &\quad + (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X_{-i}^\top y_{-i}/(n-1)\\ &= \big\{(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\} X^\top y/n\\ &\quad + (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}.\end{aligned}$$

Applying the triangle inequality, for each $i = 1, \dots, n$, we can then bound

$$\begin{aligned}\big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \le \big\|\big\{(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\}X^\top y/n\big\|_2 \\ + \big\|(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}\big\|_2.\end{aligned}$$

Averaging the bounds above thus provides

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 &\le \frac{1}{n}\sum_{i=1}^n \big\|(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}\big\| \\ &+ \frac{1}{n}\sum_{i=1}^n \big\|\big\{(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\}X^\top y/n\big\|. \qquad \text{(B.17)}\end{aligned}$$

We will see below that each of the two terms on the right-hand side of (B.17) almost surely goes to zero providing the desired convergence. Note that for each $i = 1, \dots, n$, we can bound

$$\begin{aligned}\big\|(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}\big\|_2 &\le \big\|(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\|_{\mathrm{op}}\big\|X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\|_2 \\ &\le C\big\|X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\|_2 \\ &= C\left\|\frac{x_i y_i}{n} - \sum_{j\neq i}\frac{x_j y_j}{(n-1)n}\right\|_2 \\ &\le \frac{C}{\sqrt{n}}\frac{\|x_i y_i\|_2}{\sqrt{n}} + \frac{C}{(n-1)\sqrt{n}}\sum_{j\neq i}\frac{\|x_j y_j\|_2}{\sqrt{n}},\end{aligned}$$

where the second line follows from the fact that $\big\|(X_{-i}^T X_{-i}/n + \lambda I)^\dagger\big\|_{\mathrm{op}}$ is almost surely bounded for n large enough (as explained in the proof of Lemma B.1.3), and last line uses triangle inequality. Now writing $x_i = \Sigma^{1/2} z_i$, note that for each $i = 1, \dots, n$,

$$\big\|x_i y_i\big\|_2/\sqrt{n} = \big\|\Sigma^{1/2} z_i y_i\big\|_2/\sqrt{n} \le \big\|\Sigma^{1/2}\big\|_{\mathrm{op}} y_i \big\|z_i\big\|_2/\sqrt{n} \le y_i\big\|z_i\big\|_2/\sqrt{n} \le C y_i$$

almost surely for sufficiently large n since $\|z_i\|_2/\sqrt{n}$ is eventually almost surely bounded from the strong law of large numbers. Hence, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \big\|(X_{-i}^\top X_{-i} + \lambda I)^\dagger\big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}\big\| &\le \frac{C}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n |y_i| + \frac{C}{(n-1)\sqrt{n}}\frac{1}{n}\sum_{i=i}^n\sum_{j\neq i}|y_j| \\ &\le \frac{C}{\sqrt{n}}\frac{(2n-1)}{(n-1)n}\sum_{i=1}^n |y_i| \\ &\le \frac{C}{\sqrt{n}} \to 0. \qquad \text{(B.18)}\end{aligned}$$

Here the second inequality follows by adding $|y_i|$ to the second term, and the last inequality follows because $\sum_{i=1}^n |y_i|/n$ is eventually almost surely bounded from the strong law of large numbers under Assumption 2.2. Using the leave-one-out sample covariance difference (B.13), we can similarly show that the second term goes to zero almost surely. Hence, we have that (B.17) almost surely goes to zero. This completes the proof.

□

Lemma B.1.3 (Bounding norm of the ridge estimator)**.** *Suppose Assumption 2.1 and Assumption 2.2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$, $\|\widehat{\beta}_\lambda\|_2 \le C$ for some positive constant C eventually almost surely.*

Proof. We can bound the norm of ridge estimator as

$$\begin{aligned}\big\|\widehat{\beta}_\lambda\big\|_2 &= \big\|(X^\top X/n + \lambda I)^\dagger X^\top y/n\big\|_2 \\ &\le \big\|(X^\top X/n + \lambda I)^\dagger X^\top/\sqrt{n}\big\|_{\mathrm{op}} \|y\|_2/\sqrt{n} \\ &\le \big\|(X^\top X/n + \lambda I)^\dagger\big\|_{\mathrm{op}} \big\|X^\top/\sqrt{n}\big\|_{\mathrm{op}} \|y\|_2/\sqrt{n}. \end{aligned} \tag{B.19}$$

Now for $\lambda \in (\lambda_{\min}, \infty)$, the first two terms in the product (B.19) are almost surely bounded for n large enough. This is because the maximum eigenvalue of $X^\top X/n$ is upper bounded by $C(1+\sqrt{\gamma})^2 r_{\max}$ for some $C > 1$ and the minimum non-zero eigenvalue is lower bounded by $c(1-\sqrt{\gamma})^2 r_{\min}$ for some $c < 1$ almost surely for sufficiently large n under Assumption 2.1 (Bai and Silverstein, 1998). From the strong law of large numbers, the final term is eventually almost surely bounded as the second moment of the response is bouned under Assumption 2.2. Hence, the product is eventually almost surely bounded, finishing the proof. □

B.2 Proofs related to Theorem 2.4.2

To show almost sure uniform convergence (in λ), we will appeal to Lemma B.6.5. A sufficient condition to establish strong stochastic equicontinuity in the current differentiable case is uniform boundness of the associated functions and their derivatives (with respect to λ) (e.g., Chpater 21 of Davidson, 1994). We will show that both T_λ and $\widehat{T}_\lambda^{\mathrm{gcv}}$ and their derivates are bounded over Λ, implying strong stochastic equicontinuity of the family of functions $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda\in\Lambda}$. Analogous analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}\}_{\lambda\in\Lambda}$, which we omit due to its similarity with the GCV analysis. Recall that Λ is a compact set in $(\lambda_{\min}, \infty)$. In the following, let $\Lambda \subset [\underline{\lambda}, \overline{\lambda}]$ where $\lambda_{\min} < \underline{\lambda} \le \overline{\lambda} < \infty$.

Bounding T_λ. We start with T_λ. Using Lemma B.1.1 with $q = 1$, under Assumptions 2.1 and 2.2, for error function t satisfying Assumption 2.3, we can bound T_λ in terms of the norm of the ridge estimator $\widehat{\beta}_\lambda$ as

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \le \big(C_1 + C_2\|\widehat{\beta}_\lambda\|_2\big)^2, \tag{B.20}$$

for some positive constants C_1 and C_2. Now following Lemma B.1.3, over Λ, we have that $\|\widehat{\beta}_\lambda\|_2$ is eventually almost surely bounded by $C\sqrt{r_{\max}}(\lambda_{\min} + \underline{\lambda})^{-1}$ for some positive constant C (independent of λ). This shows that T_λ is eventually almost surely bounded over $\lambda \in \Lambda$.

Bounding $\widehat{T}_\lambda^{\mathrm{gcv}}$. We next consider $\widehat{T}_\lambda^{\mathrm{gcv}}$. Using the alternate representation (B.15), for error function t satisfying Assumption 2.3, for some positive constants C, C_1, C_2, we can bound

$$\begin{aligned}\widehat{T}_\lambda^{\mathrm{gcv}} &= \frac{1}{n}\sum_{i=1}^n t\left(\frac{[(XX^\top/n + \lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n}\right) \\ &\le \frac{C_2}{n}\sum_{i=1}^n \frac{\big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2}{\big\{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big\}^2} + \frac{C_1}{n}\sum_{i=1}^n \frac{\big|[(XX^\top/n + \lambda I)^\dagger y]_i\big|}{\big|\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big|} + C \\ &\le \frac{C_2}{n}\sum_{i=1}^n \big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2 + \frac{C_1}{n}\sum_{i=1}^n \big|[(XX^\top/n + \lambda I)^\dagger y]_i\big| + C. \end{aligned} \tag{B.21}$$

The last inequality above follows by noting that the map $\lambda \mapsto \mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is non-increasing over $[\underline{\lambda}, \overline{\lambda}]$, so $\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is lower bounded by $\mathrm{tr}[(XX^\top/n + \underline{\lambda} I)^\dagger/n]$. Since $\lambda_{\min} < \underline{\lambda}$, we then have that $\{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\}^{-1}$ is upper bounded by $(\lambda_{\min} + \underline{\lambda})^{-1}$. Now, observe that for the first term in (B.21):

$$\frac{1}{n}\sum_{i=i}^n \big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2 = \frac{1}{n}\big\|(XX^\top/n + \lambda I)^\dagger y\big\|_2^2 \le \frac{1}{n}\big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}^2 \big\|y\big\|_2^2.$$

Similarly, note that for the second term in (B.21):

$$\frac{1}{n}\sum_{i=1}^{n}\big|[(XX^\top/n{+}\lambda I)^\dagger y]_i\big| = \frac{1}{n}\big\|(XX^\top/n{+}\lambda I)^\dagger y\big\|_1 \le \frac{1}{\sqrt{n}}\big\|(XX^\top/n{+}\lambda I)^\dagger y\big\|_2 \le \frac{1}{\sqrt{n}}\big\|(XX^\top/n{+}\lambda I)^\dagger\big\|_{\mathrm{op}}\big\|y\big\|_2.$$

Since $\|(XX^\top/n+\lambda I)^\dagger\|_{\mathrm{op}}$ is uniformly bounded over $\lambda \in \Lambda$ under Assumption 2.1 as argued above, and $\|y\|_2^2/n$ is almost surely bouned for n large enough from the law of large numbers under Assumption 2.2, it follows that $\widehat{T}_\lambda^{\mathrm{gcv}}$ is almost surely bounded over $\lambda \in \Lambda$.

Bounding derivative of T_λ. We now turn to bounding the derivaties of the map $\lambda \mapsto T_\lambda$. First note that since $\mathbb{E}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| \mid X, y\big] \le \mathbb{E}\big[|y_0 - x_0^\top\widehat{\beta}_\lambda|^2 \mid X, y\big]^{1/2}$, and since the latter is almost surely bounded as shown above, we can switch the order of differentiation and integration. The derivative of T_λ with respect to λ can then be bounded above by

$$T_\lambda' = \mathbb{E}\big[t'(y_0{-}x_0^\top\widehat{\beta}_\lambda)\, x_0^\top\widehat{\beta}_\lambda' \mid X, y\big] \le \mathbb{E}\big[\{t'(y_0{-}x_0^\top\widehat{\beta}_\lambda)\}^2 \mid X, y\big]^{1/2}{\cdot}\mathbb{E}\big[(\widehat{\beta}_\lambda')^\top x_0 x_0^\top \widehat{\beta}_\lambda' \mid X, y\big] \le C\sqrt{r_{\max}}\|\widehat{\beta}_\lambda'\|_2. \tag{B.22}$$

In the above chain, the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from the bounding of T_λ per (B.20) above (because under Assumption 2.3, t' is bounded above by a linear function), and the fact that $\|\Sigma\|_{\mathrm{op}} \le r_{\max}$. Applying Lemma B.2.1 on the last term of (B.22), we thus conclude that the derivative of T_λ is almost surely uniformly bounded over $\lambda \in \Lambda$, as desired.

Bounding derivative of $\widehat{T}_\lambda^{\mathrm{gcv}}$. Finally, we bound the derivative of the map $\lambda \mapsto \widehat{T}_\lambda^{\mathrm{gcv}}$. From the chain rule, the derivative of $\widehat{T}_\lambda^{\mathrm{gcv}}$ with respect to λ can be expressed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\\
&\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} &\text{(B.23)}\\
&\le C\sqrt{\sum_{i=1}^{n}\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} &\text{(B.24)}
\end{aligned}$$

The first inequality above again follows from the Cauchy-Schwarz inequalty. The second inequality follows since, from Assumption 2.3, t' is bouned above by a linear function, and the bounding of $\widehat{T}_\lambda^{\mathrm{gcv}}$ per (B.21) above shows that the first term of (B.23) is almost surely bounded. Applying Lemma B.2.2, we can now upper bound the final term of (B.24). This leads the derivative of $\widehat{T}_\lambda^{\mathrm{gcv}}$ to be almost surely bounded over $\lambda \in \Lambda$ and concludes the proof.

Lemma B.2.1 (Bounding norm of the derivative of ridge estimator)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$, $\|\widehat{\beta}_\lambda'\|_2 \le C$ eventually almost surely for some positive constant C.*

Proof. The proof follows from a straightforward calculation. Expressing the ridge estimation in the gram form, observe that

$$\frac{d\widehat{\beta}_\lambda}{d\lambda} = \frac{dX^\top(XX^\top/n+\lambda I)^\dagger y/n}{d\lambda} = X^\top(XX^\top/n+I)^\dagger(XX^\top/n+\lambda I)^\dagger y/n.$$

In the above, we use the fact that for $\lambda \in (\lambda_{\min}, \infty)$, the map $\lambda \mapsto (XX^\top/n+\lambda I)^\dagger$ is almost surely differentiable for n large enough, with the derivative given by $(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger$. The result then follows by noting that the opeator norms of $X/\sqrt{n}$ and $(XX^\top/n+\lambda I)^\dagger$ are uniformly bounded over Λ as argued above, and $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough, as explained in the proof of Lemma B.1.3. □

Lemma B.2.2 (Bounding norm of the derivative of modified GCV residuals)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$, we have that*

$$\frac{1}{\sqrt{n}} \left\| \frac{d}{d\lambda} \left(\frac{(XX^\top/n + \lambda I)^\dagger y}{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n} \right) \right\|_2 \le C$$

eventually almost surely for some positive contant C.

Proof. The proof uses straightforward matrix calculus (Petersen et al., 2008). Using the chain rule, we can write

$$\begin{aligned} \frac{d}{d\lambda} \left(\frac{(XX^\top/n + \lambda I)^\dagger y}{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n} \right) &= -\frac{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger (XX^\top/n + \lambda I)^\dagger]/n}{\{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n\}^2} (XX^\top/n + \lambda I)^\dagger y \\ &\quad + \frac{1}{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n} \frac{d}{d\lambda} \big((XX^\top/n + \lambda y)^\dagger y \big). \end{aligned}$$

Note that $\{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n\}^{-1}$ is almost surely bounded for n sufficiently large as argued above. In addition, since the operator norm of $(XX^\top/n + \lambda I)^\dagger$ is uniformly upper bounded for $\lambda \in \Lambda$, we also have that $\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger (XX^\top/n + \lambda I)^\dagger]/n$ is uniformly upper bounded over Λ. Next, observe that

$$\frac{d}{d\lambda} \big((XX^\top/n + \lambda I)^\dagger y \big) = (XX^\top/n + \lambda I)^\dagger (XX^\top/n + \lambda I)^\dagger y.$$

As above, since the opeator norm of $(XX^\top/n + \lambda I)^\dagger$ is uniformly bounded for $\lambda \in \Lambda$, and $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough, the result then follows from simple application of the triangle inequality (with respect to the ℓ_2 norm). This finishes the proof. □

B.3 Proofs related to Theorem 2.4.3

The proof is similar to that of proof of Theorem 2.4.2. We will again use Lemma B.6.5. In the current the nonsmooth case, it is sufficient to show that the family of random functions under consideration is almost surely Lipschitz continuous, along with the almost sure uniform bounds as shown in the proof of Theorem 2.4.2 (see, e.g., Chpater 21 of Davidson, 1994). We will show in the two helper lemmas below that this holds for $\{T_\lambda\}_{\lambda \in \Lambda}$ and $\{\widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda \in \Lambda}$, assuming that the loss function t is Lipschitz continuous. This will show that $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda \in \Lambda}$ is almost surely Lipschitz continuous from which the theorem follows. A similar analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}\}_{\lambda \in \Lambda}$.

Lemma B.3.1 (Lipschitz continuity of the out-of-sample functional)**.** *Suppose Assumption 2.1 and Assumption 2.2 hold, and the error function t is Lipschitz continuous. Let Λ be a compact set in $(\lambda_{\min}, \infty)$. Then, over Λ, the random map $\lambda \mapsto T_\lambda$ is almost surely Lipschitz continuous.*

Proof. Since Λ is compact, let $\Lambda \subseteq [\underline{\lambda}, \overline{\lambda}]$ where $\lambda_{\min} < \underline{\lambda} \le \overline{\lambda} < \infty$. For any $\lambda_1, \lambda_2 \in [\underline{\lambda}, \overline{\lambda}]$, using the Lipschitz continuity of the error function, we have

$$\big| t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2}) \big| \le L \big| x_0^\top (\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}) \big|$$

for some $L \ge 0$. Now consider

$$\begin{aligned} \big| T_{\lambda_1} - T_{\lambda_2} \big| &= \Big| \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2}) \mid X, y \big] \Big| \\ &\le \mathbb{E}\Big[\big| t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2}) \big| \mid X, y \Big] \\ &\le L \mathbb{E}\Big[\big| x_0^\top (\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}) \big| \mid X, y \Big] \\ &= L \mathbb{E}\Big[\sqrt{ \big| x_0^\top (\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}) \big|^2 } \mid X, y \Big] \end{aligned}$$

$$
\begin{aligned}
&\le L\sqrt{\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\big|^2 \mid X, y\Big]} \\
&\le L\sqrt{\mathbb{E}\Big[\big|(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})^\top x_0 x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\big|^2 \mid X, y\Big]} \\
&\le L\sqrt{(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})^\top \Sigma (\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})} \\
&\le L\sqrt{r_{\max}}\big\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\big\|_2.
\end{aligned}
$$

Above, the second and fourth lines follow from using Jensen's inequality (on the absolute and square root functions, respectively), the third line follows from the Lipschitz bound on the error function, and the last inequality follow since the operator norm of Σ is bounded above by $r_{\max}$.

To complete the proof, we show below that over $[\underline{\lambda}, \overline{\lambda}]$, $\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\| \le C|\lambda_1 - \lambda_2|$ for some constant C that is eventually almost surely bounded. To see this, we start by writing the difference using equivalent gram representation for ridge estimator:

$$
\begin{aligned}
\big\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\big\|_2 &= \big\|X(XX^\top/n + \lambda_1)^\dagger y/n - X(XX^\top/n + \lambda_2)^\dagger y/n\big\|_2 \\
&\le \big\|X/\sqrt{n}\big\|_{\mathrm{op}}\big\|(XX^\top/n + \lambda_1) - (XX^\top/n + \lambda_2)\big\|_{\mathrm{op}}\big\|y\big\|_2/\sqrt{n}. \qquad \text{(B.25)}
\end{aligned}
$$

As argued before, both the first and the last term in the product (B.25) are eventually almost surely bounded under Assumptions 2.1 and 2.2. For the middle term, note that on $[\underline{\lambda}, \overline{\lambda}]$, since $\lambda_{\min} < \underline{\lambda}$, the map $\lambda \mapsto (XX^\top/n + \lambda I)^\dagger$ is differentiable on $[\underline{\lambda}, \overline{\lambda}]$ with the derivative with respect to λ equal to $(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger$. Thus, using the mean value theorem, for some $\lambda \in (\underline{\lambda}, \overline{\lambda})$, we can bound

$$
\big|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\big| \le \big|(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger\big|\,|\lambda_1 - \lambda_2|\,.
$$

Hence, we can bound the second term as

$$
\begin{aligned}
\big\|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\big\|_{\mathrm{op}} &\le \big\|(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}\,|\lambda_1 - \lambda_2| \\
&\le \big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}\,\big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}\,|\lambda_1 - \lambda_2| \\
&\le C\,|\lambda_1 - \lambda_2|\,, \qquad \text{(B.26)}
\end{aligned}
$$

where the last inequality follows because $\lambda \ge \underline{\lambda} > \lambda_{\min}$ as explained in the proof of Lemma B.1.3. This concludes the proof. □

Lemma B.3.2 (Lipschitz continuity of the GCV functional)**.** *Suppose Assumption 2.1 and Assumption 2.2 hold, and the error function t is Lipschitz continuous. Let Λ be a compact set in $(\lambda_{\min}, \infty)$. Then, over Λ, the random map $\lambda \mapsto \widehat{T}_\lambda^{\mathrm{gcv}}$ is almost surely Lipschitz continuous.*

Proof. Let $\Lambda \subseteq [\underline{\lambda}, \overline{\lambda}]$, where $\lambda_{\min} < \underline{\lambda} \le \overline{\lambda} < \infty$. Using the alternate representation (B.15) for the numerator and (B.16) for the denominator of GCV reweighted errors, we can rewrite the plug-in functional $\widehat{T}_\lambda^{\mathrm{gcv}}$ as

$$
\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n t\left(\frac{\big[(XX^\top/n + \lambda I)^\dagger y\big]_i}{\operatorname{tr}\big[(XX^\top/n + \lambda I)^\dagger\big]/n}\right).
$$

For $\lambda_1, \lambda_2 \in \Lambda$ using the Lipschitz continuity of the error function, note that

$$
\begin{aligned}
&\widehat{T}_{\lambda_1}^{\mathrm{gcv}} - \widehat{T}_{\lambda_2}^{\mathrm{gcv}} \qquad \text{(B.27)} \\
&= \frac{1}{n}\sum_{i=1}^n t\left(\frac{\big[(XX^\top/n + \lambda_1 I)^\dagger y\big]_i}{\operatorname{tr}\big[(XX^\top/n + \lambda_1 I)^\dagger\big]/n}\right) - t\left(\frac{\big[(XX^\top/n + \lambda_2 I)^\dagger y\big]_i}{\operatorname{tr}\big[(XX^\top/n + \lambda_2 I)\big]/n}\right) \\
&\le \frac{1}{n}\sum_{i=1}^n L\left|\frac{\big[(XX^\top/n + \lambda_1 I)^\dagger y\big]_i}{\operatorname{tr}\big[(XX^\top/n + \lambda_1 I)^\dagger\big]/n} - \frac{\big[(XX^\top/n + \lambda_2 I)^\dagger y\big]_i}{\operatorname{tr}\big[(XX^\top/n + \lambda_2 I)^\dagger\big]/n}\right|
\end{aligned}
$$

$$
\begin{aligned}
&\le L \left| \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_1 I)^\dagger\big]/n} - \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_2 I)^\dagger\big]/n} \right| \frac{1}{n}\sum_{i=1}^n \left| \Big[(XX^\top/n+\lambda_1 I)^\dagger y\Big]_i - \Big[(XX^\top/n+\lambda_2 I)^\dagger y\Big]_i \right| \\
&\le L \left| \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_1 I)^\dagger\big]/n} - \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_2 I)^\dagger\big]/n} \right| \frac{1}{n}\sum_{i=1}^n \left| \Big[\big\{(XX^\top/n+\lambda_1 I)^\dagger - (XX^\top/n+\lambda_2 I)^\dagger\big\} y\Big]_i \right| \\
&\le L \left| \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_1 I)^\dagger\big]/n} - \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_2 I)^\dagger\big]/n} \right| \frac{1}{n} \left\| \big\{(XX^\top/n+\lambda_1 I)^\dagger - (XX^\top/n+\lambda_2 I)^\dagger\big\} y \right\|_1
\end{aligned}
\tag{B.28}
$$

Since the map $\lambda \mapsto \operatorname{tr}\big[(XX^\top + \lambda I)^\dagger\big]/n$ is non-increasing over $[\underline{\lambda}, \overline{\lambda}]$, we can bound the first term of (B.28) using

$$
\left| \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_1 I)^\dagger\big]/n} - \frac{1}{\operatorname{tr}\big[(XX^\top/n+\lambda_2 I)^\dagger\big]/n} \right| \le 2 \left| \frac{1}{\operatorname{tr}\big[(XX^\top/n+\underline{\lambda} I)^\dagger\big]/n} \right|. \tag{B.29}
$$

For bounding the second term of (B.28), note that

$$
\begin{aligned}
\left\| \big\{(XX^\top/n+\lambda_1 I)^\dagger - (XX^\top/n+\lambda_2 I)^\dagger\big\} y \right\|_1 /n &\le \left\| \big\{(XX^\top/n+\lambda_1 I)^\dagger - (XX^\top/n+\lambda_2 I)^\dagger\big\} y \right\|_2 /\sqrt{n} \\
&\le \left\| (XX^\top/n+\lambda_1 I)^\dagger - (XX^\top/n+\lambda_2 I)^\dagger \right\|_{\mathrm{op}} \|y\|_2/\sqrt{n} \\
&\le C\,|\lambda_1 - \lambda_2|,
\end{aligned}
\tag{B.30}
$$

where we used the bound from (B.26), along with the fact that $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough from the strong law of large numbers under Assumption 2.2. Plugging (B.29) and (B.30) into (B.28) then finishes the proof. □

B.4 Proof of Theorem 2.3.1

Let $\widehat{F}_\lambda^{\mathrm{gcv}}$ and $\widehat{F}_\lambda^{\mathrm{loo}}$ denote the CDFs associated with the plug-in distributions $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$ of the GCV and LOOCV reweighted errors, respectively. Recall that F_λ denotes the CDF of the out-of-sample error distribution P_λ. To prove Theorem 2.3.1, for all $z \in \mathbb{R}$ that are continuity points of F_λ for n sufficiently large, we will sandwich $F_\lambda(z)$ such that, almost surely, $\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \le F_\lambda(z)$ along with $F_\lambda(z) \le \liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z)$. This then yields the desired result that $\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F_\lambda(z) \xrightarrow{\text{a.s.}} 0$. Similar argument shows $\widehat{F}_\lambda^{\mathrm{loo}}(z) - F_\lambda(z) \xrightarrow{\text{a.s.}} 0$. The idea of the proof is similar to that used in the proof of the Portmanteau theorem, with the main difference being that the target distribution in our case is also a random distribution. We will make use of Theorem 2.4.1 to deduce the desired inequalities in each direction using suitably chosen error functions.

Fix $\epsilon > 0$ and $z \in \mathbb{R}$. For the first direction, let $t_{z,\epsilon}$ be an error function defined as

$$
t_{z,\epsilon}(r) = \begin{cases} 1 & r \le z \\ 1 + (z-r)/\epsilon & z \le r \le z+\epsilon \\ 0 & r \ge z+\epsilon. \end{cases}
$$

Observe that $\mathbb{I}\{r \le z\} \le t_{z,\epsilon}(r)$ for all $r \in \mathbb{R}$. Here $\mathbb{I}$ denotes the indicator function. This allow us to write

$$
\widehat{F}_\lambda^{\mathrm{gcv}}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{ \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} \le z \right\} \le \frac{1}{n}\sum_{i=1}^n t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} \right). \tag{B.31}
$$

Furthermore, $t_{r,\epsilon}$ is Lipschitz continuous and satisfies Assumption 2.3. Hence, invoking Theorem 2.4.1, we have that

$$
\frac{1}{n}\sum_{i=1}^n t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \operatorname{tr}[L_\lambda]/n} \right) - \mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{B.32}
$$

In addition, observe that $t_{z,\epsilon}(r) \le \mathbb{I}\{r \le z + \epsilon\}$ for all $r \in \mathbb{R}$. This gives us

$$\mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \le \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon \mid X, y\big]. \tag{B.33}$$

Thus, combining (B.31) to (B.33), we get that almost surely

$$\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \le \limsup_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon \mid X, y\big] = \limsup_{n\to\infty} F_\lambda(z + \epsilon). \tag{B.34}$$

Now sending $\epsilon \to 0$, we obtain the desired inequality $\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \le F_\lambda(z)$ almost surely.

We proceed analogously on the other side. Again fix $\epsilon > 0$ and let $z \in \mathbb{R}$ be a continuity point of F_λ for n sufficiently large. We will now use the function $t_{z-\epsilon,\epsilon}$. Explicitly, the evaluation map of $t_{z-\epsilon,\epsilon}$ is given by

$$t_{z-\epsilon,\epsilon}(r) = \begin{cases} 1 & r \le z - \epsilon \\ (z - r)/\epsilon & z - \epsilon \le r \le z \\ 0 & r \ge z. \end{cases}$$

Noting that $t_{z-\epsilon,\epsilon}(r) \le \mathbb{I}\{r \le z\}$ for all $r \in \mathbb{R}$, we obtain

$$\widehat{F}_\lambda^{\mathrm{gcv}}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \le z\right\} \ge \frac{1}{n}\sum_{i=1}^{n} t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right). \tag{B.35}$$

Again, since $t_{z-\epsilon,\epsilon}$ is Lipschitz continuous and satisfies Assumption 2.3, application of Theorem 2.4.1 yields

$$\frac{1}{n}\sum_{i=1}^{n} t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{B.36}$$

Finally, because $t_{z-\epsilon,\epsilon}(r) \ge \mathbb{I}\{r \le z - \epsilon\}$ for $r \in \mathbb{R}$, we have that

$$\mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \ge \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\big]. \tag{B.37}$$

Combining (B.35) to (B.37), we have almost surely,

$$\liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \ge \liminf_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\big] = \liminf_{n\to\infty} F_\lambda(z - \epsilon). \tag{B.38}$$

Since z is a continuity point of F_λ, sending $\varepsilon \to 0$, we get the desired inequality $\liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \ge F_\lambda(z)$ almost surely.

Combining (B.34) and (B.38), we conclude that almost surely $\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) - \liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \to 0$, and $\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F(z) \to 0$, completing the proof.

B.5 Proofs related to Theorem 2.5.1

B.5.1 Proof of Theorem 2.5.1

The proof of Theorem 2.5.1 mainly builds on the result of Theorem 2.4.1. We will use Theorem 2.4.1 to certify pointwise convergence (in v) of $\widehat{T}_\lambda^{\mathrm{gcv}}(v)$ and $\widehat{T}_\lambda^{\mathrm{loo}}(v)$ to $T_\lambda(v)$. Then using the equicontinuity of $\mathcal{T}_{\mathcal{V}}$ and appealing to Lemma B.6.6, we will prove the convergence of the minimizers $\widehat{V}_\lambda^{\mathrm{gcv}}$ and V_λ^{loo} to V_λ.

First observe that each $t(\cdot, v) : \mathbb{R} \to \mathbb{R}$ is a continuos function since $\mathcal{T}_{\mathcal{V}}$ is an equicontinous family of functions. In addition, each $t(\cdot, v)$ satisfies Assumption 2.3. Thus, for each $v \in \mathcal{V}$, Theorem 2.4.1 implies

$$\widehat{T}_\lambda^{\mathrm{gcv}}(v) - T_\lambda(v) \xrightarrow{\text{a.s.}} 0.$$

Next note that for any $\delta > 0$,

$$\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big|T_\lambda(v_1) - T_\lambda(v_2)\big|$$

$$
\begin{aligned}
&= \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \left|\mathbb{E}\big[t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) \mid X, y\big] - \mathbb{E}\big[t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2) \mid X, y\big]\right| \\
&= \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \left|\mathbb{E}\big[t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2) \mid X, y\big]\right| \\
&\le \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \mathbb{E}\Big[\big|t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2)\big| \mid X, y\Big] \\
&\le \mathbb{E}\left[\sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \big|t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2)\big| \;\Big|\; X, y\right], \qquad (B.39)
\end{aligned}
$$

where the third line follows from Jensen's inequality, the last inequality follows because for any $v_1, v_2 \in \mathcal{V}$ such that $|v_1 - v_2| \le \delta$, we have that

$$
\big|t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2)\big| \le \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \big|t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top\widehat{\beta}_\lambda, v_2)\big|,
$$

which after taking expectation and taking sup gives the desired inequality. Similarly, for any $\delta > 0$,

$$
\begin{aligned}
&\sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \big|\widehat{T}_\lambda^{\mathrm{gcv}}(v_1) - \widehat{T}_\lambda^{\mathrm{gcv}}(v_2)\big| \\
&\quad= \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \left|\frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1\right) - \frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2\right)\right| \\
&\quad\le \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \frac{1}{n}\sum_{i=1}^n \left|t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1\right) - t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2\right)\right| \\
&\quad\le \frac{1}{n}\sum_{i=1}^n \sup_{|v_1-v_2|\le\delta,\, v_1,v_2\in\mathcal{V}} \left|t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1\right) - t\left(\frac{y_i - x_i^\top\widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2\right)\right|. \qquad (B.40)
\end{aligned}
$$

Note that the exact argument holds for the case of $\lambda = 0$ by replacing replacing the first argument of t with the modified GCV errors. Since the family $\{t(\cdot, v) : v \in \mathcal{V}\}$ is pointwise equicontinous, (B.39) and (B.40) imply equicontinuity of $\{T_\lambda(v) : v \in \mathcal{V}\}$ and $\{\widehat{T}_\lambda^{\mathrm{gcv}}(v) : v \in \mathcal{V}\}$. Moreover, as $\mathcal{V}$ is compact and V_λ is assumed to be unique, Lemma B.6.6 yields

$$
\widehat{V}_\lambda^{\mathrm{gcv}} - V_\lambda \xrightarrow{\text{a.s.}} 0.
$$

Analogous argument shows the convergence for $\widehat{V}_\lambda^{\mathrm{loo}}$ by using the LOOCV part of Theorem 2.4.1.

B.5.2 Proof of Corollary 2.5.2

We verify that the conditions of Theorem 2.5.1 are satisfied. For $\tau \in (0, 1)$ and compact set $\mathcal{U} \subseteq \mathbb{R}$, the family of error functions under consideration is $\mathcal{T}_{\mathcal{U}} = \{t_\tau(\cdot, u) : u \in \mathcal{U}\}$, where each function $t_\tau(\cdot, u)$ is such that for $r \in \mathbb{R}$

$$
t_\tau(r, u) = (r - u)(\tau - \mathbb{I}\{r - u < 0\}.
$$

In other words, the evaluation map is given by

$$
t_\tau(r, u) = \begin{cases} (r - u)\tau & \text{if } r \ge u \\ (u - r)(1 - \tau) & \text{if } u > r. \end{cases}
$$

A sufficient condition to establish equicontinuity of $\mathcal{T}_{\mathcal{U}}$ is to show that the functions in the family are Lipschitz continuous with uniformly bounded Lipschitz constant (see, e.g., Section 1.8 of Tao, 2010). It is easy to check that each function in the family $\mathcal{T}_{\mathcal{U}}$ is Lipschitz continuous with uniformly bounded constant $L = \max\{\tau, 1 - \tau\}$. Thus, the family $\mathcal{T}_{\mathcal{U}}$ is equicontinous over compact set $\mathcal{U}$. Furthermore, since $\mathcal{U}$ is assumed to contain the true quantile, $Q_\lambda(\tau)$ is unique. Therefore, invoking Theorem 2.5.1 we obtain the desired conclusion.

B.6 Useful results

In this section, we record statements of various results adapted from other sources that are used in the proofs throughout the supplement.

The following inequality bounding q-th moment of sum of random variables is by Burkholder (1973). See also Bai and Silverstein (2010, Lemma 2.13).

Lemma B.6.1 (Burkholder's inequality)**.** *Let $\{Z_k\}$ be a martingale difference sequence with respect to the increasing σ-field $\{\mathcal{F}_k\}$. Then, for $q \ge 2$,*

$$\mathbb{E}\Big[\Big|\sum_k Z_k\Big|^q\Big] \le C_q \left\{\mathbb{E}\left[\Big(\sum_k \mathbb{E}\Big[\big|Z_k\big|^2 \mid \mathcal{F}_{k-1}\Big]\Big)^{q/2}\right] + \mathbb{E}\Big[\sum_k \big|Z_k\big|^q\Big]\right\}$$

for a constant C_q that only depends on q.

The following inequality bounding L_p norm of an inner product is from Erdos and Yau (2017, Lemma 7.8).

Lemma B.6.2 (L_q norm of an inner product)**.** *Let $u \in \mathbb{R}^p$ be a random vector consisting of independent entries u_i with $\mathbb{E}[u_i] = 0$, $\mathbb{E}[u_i^2] = 1$, and $\|u_i\|_{L_q} \le K_q$ for $i = 1, \ldots, p$. Let $a \in \mathbb{R}^p$ be a deterministic vector. Then,*

$$\|a^\top u\|_{L_q} \le C_q K_q \|a\|_2$$

for a constant C_q depending only on q.

The following lemma bounding q-th moment of a quadratic form is from Bai and Silverstein (2010, Lemma B.26). See also Dobriban and Wager (2018, Lemma 7.10).

Lemma B.6.3 (Centered moment a quadratic form)**.** *Let $W \in \mathbb{R}^{p\times p}$ be a deterministic matrix. Let $v \in \mathbb{R}^p$ be a random vector of independent entries v_i for $i = 1, \ldots, p$ with each $\mathbb{E}[v_i] = 0$, $\mathbb{E}[v_i^2] = 1$, and $\mathbb{E}[|v_i|^r] \le M_r$. Then, for any $q \ge 1$,*

$$\mathbb{E}\Big[\big|v^\top W v - \operatorname{tr}[W]\big|^q\Big] \le C_q \left\{\big(M_4 \operatorname{tr}[WW^\top]\big)^{q/2} + M_{2q} \operatorname{tr}\big[(WW^\top)^{q/2}\big]\right\}$$

for a constant C_q that only depends on q.

The following equivalence lemma for the denominator arising from GCV is adapted from Patil et al. (2021, Lemma S.3.1).

Lemma B.6.4 (GCV denominator lemma)**.** *Suppose Assumption 2.1 holds. Then, for $\lambda \in (\lambda_{\min}, \infty) \setminus \{0\}$*

$$1 + \operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger \Sigma\big]/n - \frac{1}{1 - \operatorname{tr}\big[(X^\top X/n + \lambda I)^\dagger X^\top X/n\big]/n} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, and for the case of $\lambda = 0$,

$$\operatorname{tr}\big[(I - (X^\top X/n)^\dagger X^\top X/n)\Sigma\big]/n - \frac{1}{\operatorname{tr}\big[(X^\top X/n)^\dagger\big]/n} \xrightarrow{\text{a.s.}} 0,$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

The following results are standard results on stochastic uniform convergence. See, e.g., Chapter 21 of Davidson (1994).

Lemma B.6.5 (Stochastic uniform convergence)**.** *Let $f_n(\theta)$, $\theta \in \Theta$ be a family of stochastic functions. Suppose Θ is a compact, and for every $\theta \in \Theta$, $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$. Further, assume that $\{f_n(\theta)\}$ is strongly stochastic equicontinous. Then, as $n \to \infty$,*

$$\sup_{\theta\in\Theta} |f_n(\theta) - f(\theta)| \xrightarrow{\text{a.s.}} 0.$$

A corollary of Lemma B.6.5 is the following statement.

Lemma B.6.6 (Convergence of minimizers)**.** *Assume the setting of Lemma B.6.5. Let $\widehat{\xi}_n$ and ξ be minimizers of f_n and f over $\theta \in \Theta$, respectively. Moreover, assume that f has a unique minimizer over Θ. Then, as $n \to \infty$,*

$$\widehat{\xi} \xrightarrow{\text{a.s.}} \xi.$$

The following lemma is a simple application of Markov's inequality along with the Borel-Cantelli lemma.

Lemma B.6.7 (Moment version of the Borel-Cantelli lemma)**.** *Let $\{S_n\}$ be a sequence of random variables. Suppose $\{\mathbb{E}[|S_n|^p]\}$ forms a summable sequence for some $p > 0$. Then, as $n \to \infty$, $S_n \xrightarrow{\text{a.s.}} 0$.*

B.7 Additional numerical illustrations

In this section, we provide additional numerical illustrations to complement those included in the main chapter. The details of feature and response models used throughout different experiments are described next.

Feature model. The feature $x_i \in \mathbb{R}^p$ is generated according to

$$x_i = \Sigma^{1/2} z_i, \tag{B.41}$$

where $z_i \in \mathbb{R}^p$ contains independently sampled entires from a common distribution, and $\Sigma \in \mathbb{R}^{p \times p}$ is a positive semidefinite feature covariance matrix. The different distributions that we use for the components of z_i include: (1) Gaussian distribution, (2) Student's t-distribution, and (3) Bernoulli distribution. These represent a mix of both continuous and discrete, and light- and heavy-tailed distributions. We standardize the distributions so that the mean is zero and the variance is one. The different feature covariance matrix structures that we use include: (1) Identity ($\Sigma_{ij} = 1$ when $i = j$ and $\Sigma_{ij} = 0$ when $i \neq j$) and (2) Autoregressive with parameter ρ ($\Sigma_{ij} = \rho^{|i-j|}$ for all i, j).

Response model. Given x_i, the response $y_i \in \mathbb{R}$ is generated according to

$$y_i = \beta_0^\top x_i + \left(x_i^\top A x_i - \operatorname{tr}[A\Sigma]\right)/p + \varepsilon_i, \tag{B.42}$$

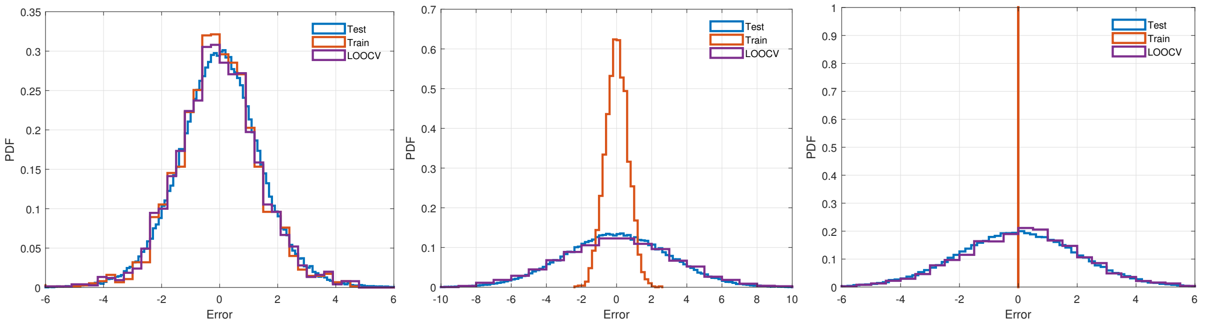
where $\beta_0 \in \mathbb{R}^p$ is a fixed signal vector, $A \in \mathbb{R}^{p \times p}$ is a fixed matrix, and $\varepsilon_i \in \mathbb{R}$ is a random noise variable. Note that we have subtracted the mean from the squared nonlinear component and scaled it to keep the variance of the nonlinear component at the same order as the noise variance (see Mei and Montanari (2022) for more details, for example). We again use either Gaussian, Student's t, or Bernoulli distribution for the random noise component, which is again standardized so that the mean is zero and the variance is one. We refer to the value of $\beta_0^\top \Sigma \beta_0$ as the effective signal energy.

Train and test set sizes. In all of our experiments, the sample size for the train set is fixed at $n = 2500$. To compute various out-of-sample quantities, we use a test set of 100000 indepedent observations. We use three feature sizes of $p = 100$, $p = 2000$, and $p = 5000$ that represent low, moderate, and high-dimensional settings (with aspect ratios p/n of 0.04, 0.8, and 2), respectively.

B.7.1 Distribution estimation

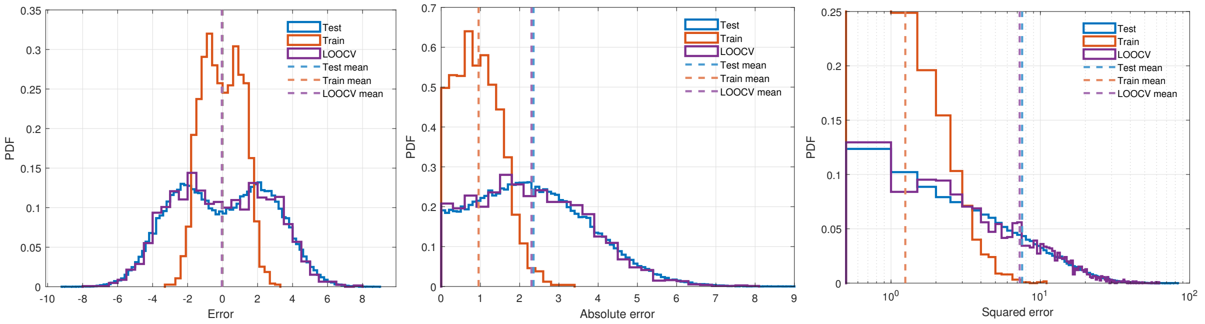
We first present illustrations with LOOCV reweighted errors for Figures 2.1 and 2.2 in Figures B.1 and B.2, respectively.

Note that both in Figures 2.1 and 2.2 as well as Figures B.1 and B.2, the out-of-sample error distributions and the associated GCV and LOOCV reweighted error distributions are all symmetric distributions. This need not be the case. In Figure B.3, we consider a case in which the out-of-sample error distribution and the estimated distributions based on GCV and LOOCV reweighted errors are negatively skewed.



(a) Low dimension ($p/n = 0.04$) (b) Moderate dimension ($p/n = 0.8$) (c) High dimension ($p/n = 2$)

Figure B.1: A simulation with $n = 2500$ and $p \in \{100, 2000, 5000\}$ features with a different p per panel above. In each setting, the feature vectors x_i are generated as in (B.41) with identity covariance with components of z_i sampled from a t-distribution with 5 degrees of freedom, and the responses y_i are generated as in (B.42). We fit the min-norm least squares solution, as in (2.1) with $\lambda = 0$. The blue curve in each panel is a histogram of the true prediction error distribution, computed from 10^5 independent test samples. The red curve is a histogram of the training errors; when $p > n$, this is just a point mass at zero. The purple curve is a histogram of LOOCV reweighted training errors, as in (2.12) (when $p < n$ in the first two panels) and (2.14) (when $p > n$ in the last panel). This tracks the blue curve very well in all three settings again. Empirical results for GCV are provided in Figure 2.1.



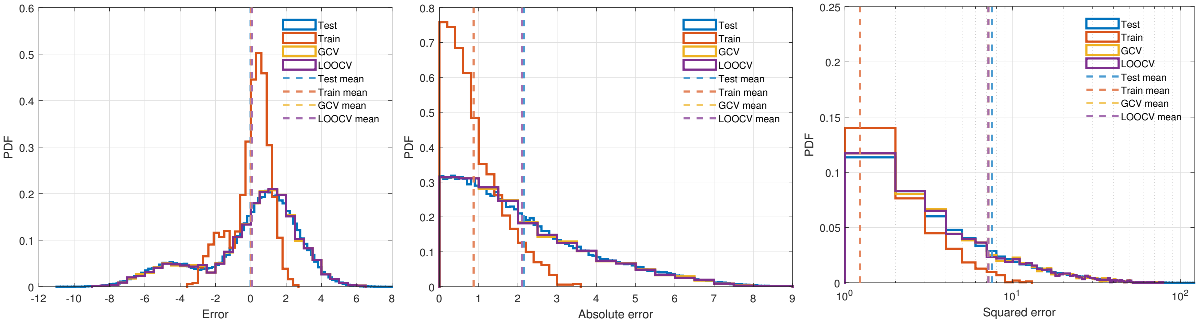
(a) $h(e) = e$ (b) $h(e) = |e|$ (c) $h(e) = e^2$

Figure B.2: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (B.41) with identity covariance with the components of z_i sampled from a symmetric Bernoulli distribution, and each response y_i is generated according to (B.42). The ridge parameter was fixed at $\lambda = 1$. Each panel above examines weak convergence per (2.17) for a different function h of the error variable (identity, absolute value, and square, from left to right). In each case, the LOOCV estimate (purple) tracks the true distribution (blue) closely. Empirical results for GCV are in Figure 2.2.

B.7.2 Quantile estimation

We first provide further details on the setup used in Figure 2.3. We use a special "latent" space data model, in which the true signal component lies in a small eigenspace of the feature covariance matrix. Such setup was investigated in the context of ridge regression by Kobak et al. (2020); Wu and Xu (2020); Richards et al. (2020); Hastie et al. (2022), who study the optimality of zero (or even negative) ridge regularization for expected squared out-of-sample error under special cases. We verify empirically that such behavior continues to hold even for general functionals of the out-of-sample error distribution and their plug-in estimators based on GCV and LOOCV such as the length of prediction intervals, and even under nonlinear model.

For numerical illustration, we consider an extreme case where the signal vector is aligned with the



(a) $h(e) = e$ (b) $h(e) = |e|$ (c) $h(e) = e^2$

Figure B.3: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (B.41) with identity covariance and components of z_i sampled from a Gaussian distribution, and each response y_i according to (B.42) with noise variable ε_i distributed according to a Bernoulli random variable with success probability 0.8. The ridge parameter was fixed at $\lambda = 1$. Each panel above examines weak convergence per (2.17) for a different function h of the error variable (identity, absolute value, and square, from left to right). In each case, the GCV estimate (yellow) and LOOCV estimate (purple) track the true distribution (blue) closely.

eigenvector of the covariance matrix corresponding to the largest eigenvalue. More precisely, let $\Sigma = WRW^\top$ denote the eigenvalue decomposition of the covariance matrix Σ, where $W \in \mathbb{R}^{p \times p}$ is a orthogonal matrix whose columns $w_1, \ldots, w_p$ are eigenvectors of Σ and $R \in \mathbb{R}^{p \times p}$ is a diagonal matrix whose entries $r_1 \geq \cdots \geq r_p$ are eigenvalues of Σ in descending order. We then let $\beta_0 = \zeta w_1$, where ζ controls the effective signal energy. Figure B.4 illustrate the coverage and length of prediction intervals (2.30) computed using the LOOCV reweighted error distribution.

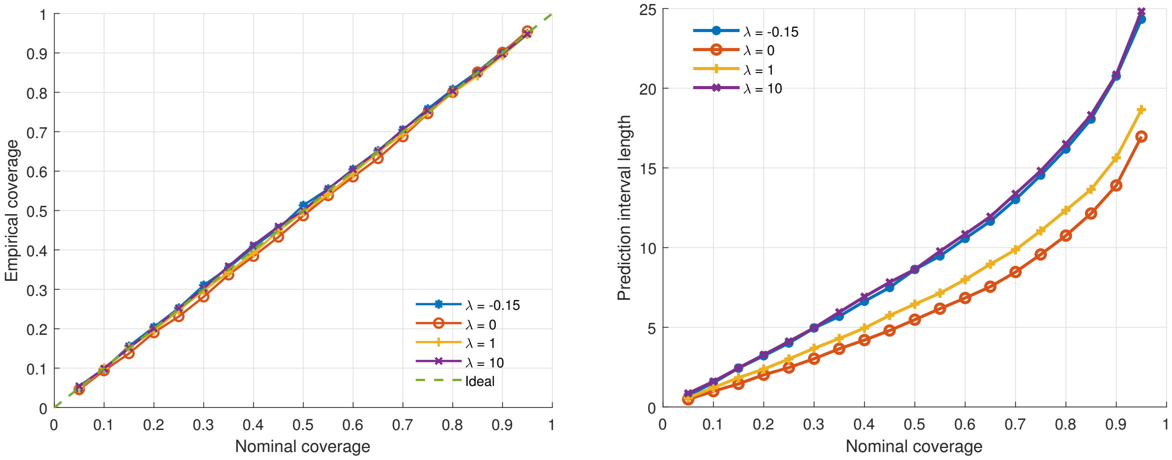


Figure B.4: Illustration of empirical coverage and length of LOOCV prediction intervals constructed using (2.30) against nominal coverage, where $n = 2500$, $p = 5000$. We generated features x_i according to (B.41) with autoregressive covariance structure (with $\rho = 0.25$) and t-distributed components of z_i with 5 degrees of freedom. The responses y_i are generated according to (B.42) where the signal β_0 is aligned with the top eigenvector of the covariance matrix and the effective signal energy is 50. We see that intervals for any λ have excellent finite-sample coverage (left), and the case of $\lambda = 0$ provides the smallest interval lengths (right). Empirical results for GCV prediction intervals are in Figure B.4.

Finally, as a contrast we consider a "regular" setting in Figure B.5 where the signal does not have any special structure, and the signal covariance is identity, where we see that regularization does in fact help indicating the subtle interplay between the signal vector and feature covariance that causes the near optimality of ridgeless estimator for various functionals of the out-of-sample error distribution.

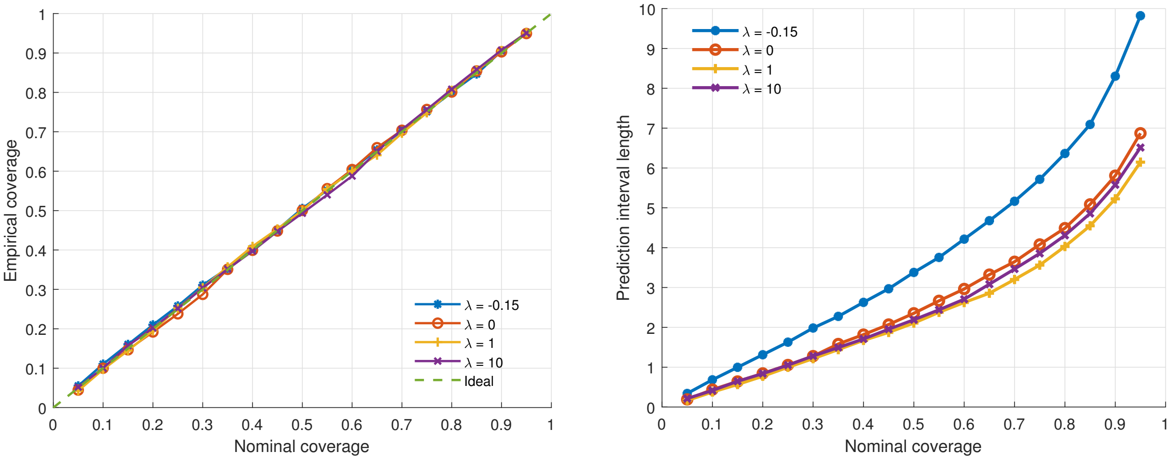


Figure B.5: Illustration of empirical coverage and length of LOOCV prediction intervals (2.30) against nominal coverage, where $n = 2500$, $p = 5000$. The features x_i are generated according to (B.41) with identity covariance and components of z_i having Gaussian distribution. The responses y_i are generated according to (B.42) with the nonlinearity component set to 0 (thus a well-specified linear model) and a random signal vector. We see again that the intervals for any λ have excellent finite-sample coverage (left) and now the case of $\lambda = 1$ provides the smallest interval lengths (right). Similar trend holds for GCV prediction intervals, and hence we do not present the corresponding figure for GCV.

Appendix C

Supplement for Chapter 3

This supplement contains proofs and additional details for Chapter 3. The content of the supplement is organized as follows.

- In Appendix C.1, we present proofs of results related to general cross-validation and model selection from Sections 3.2.1 to 3.2.3.
- In Appendix C.2, we present proofs of results related to risk monotonization behavior of the zero-step procedure from Section 3.3.3.
- In Appendix C.3, we present proofs for the verification of the deterministic risk profile assumption for the MN2LS and MN1LS prediction procedures from Section 3.3.3.2.
- In Appendix C.4, we present proofs of results related to risk monotonization behavior of the one-step procedure from Section 3.4.3.1.
- In Appendix C.5, we present proofs for the verification of the deterministic risk profile assumption for arbitrary linear prediction procedures, and the MN2LS and MN1LS prediction procedures from Section 3.4.3.2.
- In Appendix C.6, we collect various technical helper lemmas and their proofs that are used in proofs in Appendices C.2 to C.5, and other miscellaneous details.
- In Appendix C.7, we list calculus rules for a certain notion of asymptotic equivalence of sequences of matrices that are used in proofs in Appendices C.3 and C.5.
- In Appendix C.8, we record statements of useful concentration results available in the literature that are used in proofs in Appendices C.1, C.3 and C.5.
- In Appendix C.9, we list some of the main notation used in this work.

C.1 Proofs related to general cross-validation and model selection

C.1.1 Proof of Proposition 3.2.1

Additive form. We will first prove the oracle risk inequalities (3.7) in additive form. Recall Algorithm 1 returns $\widehat{f}^{\mathrm{cv}} = \widehat{f}^{\widehat{\xi}}$. Adding and subtracting $\min_{\xi \in \Xi} R(\widehat{f}^{\xi})$ and $\min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi})$ to $R(\widehat{f}^{\mathrm{cv}})$, we can break $R(\widehat{f}^{\mathrm{cv}})$ into the following additive form:

$$R(\widehat{f}^{\mathrm{cv}}) = \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) + \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) - \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) - \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) + R(\widehat{f}^{\widehat{\xi}}). \tag{C.1}$$

An application of triangle inequality then lets us upper bound $R(\widehat{f}^{\mathrm{cv}})$ into sum of three terms:

$$R(\widehat{f}^{\mathrm{cv}}) \le \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) + \underbrace{\left| \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) - \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \right|}_{(a)} + \underbrace{\left| R(\widehat{f}^{\widehat{\xi}}) - \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) \right|}_{(b)}. \tag{C.2}$$

We will next upper bound both terms (a) and (b) by Δ_n^{add} to finish the first inequality of (3.7).

By definition (3.6a) of Δ_n^{add}, for every $\xi \in \Xi$, we can write

$$R(\widehat{f}^{\xi}) \le \widehat{R}(\widehat{f}^{\xi}) + \Delta_n^{\mathrm{add}} \quad \text{and} \quad \widehat{R}(\widehat{f}^{\xi}) \le R(\widehat{f}^{\xi}) + \Delta_n^{\mathrm{add}}. \tag{C.3}$$

Taking minimum on both sides of the inequalities in (C.3) then yields

$$\min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) \le \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) + \Delta_n^{\mathrm{add}} \quad \text{and} \quad \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \le \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) + \Delta_n^{\mathrm{add}}.$$

Combining the two inequalities, we arrive at the desired bound for term (a):

$$\left| \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) - \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \right| \le \Delta_n^{\mathrm{add}}. \tag{C.4}$$

Since $\widehat{\xi} \in \arg\min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi})$, we can obtain the following upper bound for term (b):

$$\left| R(\widehat{f}^{\widehat{\xi}}) - \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) \right| = \left| R(\widehat{f}^{\widehat{\xi}}) - \widehat{R}(\widehat{f}^{\widehat{\xi}}) \right| \le \Delta_n^{\mathrm{add}}, \tag{C.5}$$

where the inequality follows from the definition of Δ_n^{add}.

Substituting the bounds (C.4) and (C.5) into (C.2), we conclude that

$$\left| R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \right| \le 2\Delta_n^{\mathrm{add}}. \tag{C.6}$$

This implies the first inequality of (3.7). Taking expectations on the both sides of the first inequality of (3.7), we obtain

$$\mathbb{E}\big[R(\widehat{f}^{\mathrm{cv}})\big] \le \mathbb{E}\big[\min_{\xi \in \Xi} R(\widehat{f}^{\xi})\big] + 2\mathbb{E}\big[\Delta_n^{\mathrm{add}}\big]. \tag{C.7}$$

It is clear that the first term on the right hand side is bounded above by $\min_{\xi \in \Xi} \mathbb{E}[R(\widehat{f}^{\xi})]$, and thus we obtain the second inequality of (3.7). This completes the proof of the oracle risk inequalities in additive form.

Multiplicative form. We now turn to prove the oracle risk inequality (3.8) in multiplicative form. Recall again that Algorithm 1 returns $\widehat{f}^{\mathrm{cv}} = \widehat{f}^{\widehat{\xi}}$. In contrast to the proof of Proposition 3.2.1, we now break $R(\widehat{f}^{\mathrm{cv}})$ into the following multiplicative form:

$$\begin{aligned}
R(\widehat{f}^{\mathrm{cv}}) = \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \widehat{R}(\widehat{f}^{\mathrm{cv}}) \; &= \; \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \widehat{R}(\widehat{f}^{\widehat{\xi}}) \\
&\overset{(i)}{=} \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \min_{\xi \in \Xi} \widehat{R}(\widehat{f}^{\xi}) \\
&= \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \min_{\xi \in \Xi} \left[\frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \cdot R(\widehat{f}^{\xi}) \right] \\
&\overset{(ii)}{\le} \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \min_{\xi \in \Xi} \left[\left(\max_{\rho \in \Xi} \frac{\widehat{R}(\widehat{f}^{\rho})}{R(\widehat{f}^{\rho})} \right) \cdot R(\widehat{f}^{\xi}) \right]
\end{aligned}$$

$$
\begin{aligned}
&\le \frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} \cdot \left(\max_{\xi \in \Xi} \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \right) \cdot \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \\
&\overset{(iii)}{\le} \frac{1}{\min\limits_{\xi \in \Xi} \dfrac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})}} \cdot \left(\max_{\xi \in \Xi} \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \right) \cdot \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \\
&= \frac{\max\limits_{\xi \in \Xi} \dfrac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})}}{\min\limits_{\xi \in \Xi} \dfrac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})}} \cdot \min_{\xi \in \Xi} R(\widehat{f}^{\xi}). \qquad \text{(C.8)}
\end{aligned}
$$

In the chain above, equality (i) follows from the definition of $\widehat{\xi}$ in Algorithm 1, inequality (ii) follows from the inequality $a_i b_i \le (\max_j a_j) b_i$ for any two sequences $a_i, b_i, 1 \le i \le m$, and inequality (iii) follows by noting that

$$
\frac{R(\widehat{f}^{\mathrm{cv}})}{\widehat{R}(\widehat{f}^{\mathrm{cv}})} = \frac{1}{\dfrac{\widehat{R}(\widehat{f}^{\mathrm{cv}})}{R(\widehat{f}^{\mathrm{cv}})}} = \frac{1}{\dfrac{\widehat{R}(\widehat{f}^{\widehat{\xi}})}{R(\widehat{f}^{\widehat{\xi}})}} \le \frac{1}{\min\limits_{\xi \in \Xi} \dfrac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})}}.
$$

Now, from the definition of Δ_n^{mul}, for all $\xi \in \Xi$, we have

$$
1 - \Delta_n^{\mathrm{mul}} \le \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \le 1 + \Delta_n^{\mathrm{mul}}.
$$

In addition, since the loss function is assumed to be non-negative, both $R(\widehat{f}^{\xi})$ and $\widehat{R}(\widehat{f}^{\xi})$ are non-negative for all ξ. Hence, we can bound

$$
(1 - \Delta_n^{\mathrm{mul}})_+ \le \min_{\xi \in \Xi} \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \le \max_{\xi \in \Xi} \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} \le 1 + \Delta_n^{\mathrm{mul}}. \qquad \text{(C.9)}
$$

Using (C.9) in (C.8) then implies the desired upper bound:

$$
R(\widehat{f}^{\mathrm{cv}}) \le \frac{1 + \Delta_n^{\mathrm{mul}}}{(1 - \Delta_n^{\mathrm{mul}})_+} \cdot \min_{\xi \in \Xi} R(\widehat{f}^{\xi}).
$$

This completes the proof of the oracle risk inequality in multiplicative form.

C.1.2 Proof of Lemma 3.2.4

Tail bound. We begin by applying the Bernstein inequality (see Lemma C.8.1 for the exact statement) on the random variables $\ell(Y_j, \widehat{f}^{\xi}(X_j)), j \in \mathcal{I}_{\mathrm{te}}$ with mean $R(\widehat{f}^{\xi})$ conditionally on $\mathcal{D}_{\mathrm{tr}}$. (Note that the random variables are i.i.d. conditionally on $\mathcal{D}_{\mathrm{tr}}$.) For any $0 < \eta < 1$ and $\xi \in \Xi$, we have the tail bound

$$
\mathbb{P}\left(\left| \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{j \in \mathcal{I}_{\mathrm{te}}} \ell(Y_j, \widehat{f}^{\xi}(X_j)) - R(\widehat{f}^{\xi}) \right| \ge C_1 \max\left\{ \sqrt{\widehat{\sigma}_{\xi}^2 \frac{\log(2/\eta)}{|\mathcal{D}_{\mathrm{te}}|}}, \widehat{\sigma}_{\xi} \frac{\log(2/\eta)}{|\mathcal{D}_{\mathrm{te}}|} \right\} \,\middle|\, \mathcal{D}_{\mathrm{tr}} \right) \le \eta. \qquad \text{(C.10)}
$$

Taking expectation on both sides, we get that the unconditional probability is also bounded by η. Denoting the prediction risk estimate by $\widehat{R}(\widehat{f}^{\xi})$, and choosing $\eta = \eta/|\Xi|$, for any $\xi \in \Xi$, we can equivalently write the bound as

$$
\mathbb{P}\left(\left| \widehat{R}(\widehat{f}^{\xi}) - R(\widehat{f}^{\xi}) \right| \ge C_1 \widehat{\sigma}_{\xi} \max\left\{ \sqrt{\frac{\log(2|\Xi|/\eta)}{n_{\mathrm{te}}}}, \frac{\log(2|\Xi|/\eta)}{n_{\mathrm{te}}} \right\} \right) \le \frac{\eta}{|\Xi|}.
$$

Applying union bound over $\xi \in \Xi$, for any $0 < \eta < 1/|\Xi|$, we get uniform bound

$$\mathbb{P}\left(\max_{\xi\in\Xi}\left|\widehat{R}(\widehat{f}^{\xi}) - R(\widehat{f}^{\xi})\right| \ge C_1 \max_{\xi\in\Xi}\widehat{\sigma}_{\xi}\max\left\{\sqrt{\frac{\log\left(2|\Xi|/\eta\right)}{n_{\mathrm{te}}}}, \frac{\log\left(2|\Xi|/\eta\right)}{n_{\mathrm{te}}}\right\}\right) \le \eta.$$

Using the definition of Δ_n^{add}, and setting $\widehat{\sigma}_{\Xi} := \max_{k\in\Xi}\widehat{\sigma}_{\xi}$, so far we have that

$$\mathbb{P}\left(\Delta_n^{\mathrm{add}} \ge C_1\widehat{\sigma}_{\Xi}\max\left\{\sqrt{\frac{\log\left(2|\Xi|/\eta\right)}{n_{\mathrm{te}}}}, \frac{\log\left(2|\Xi|/\eta\right)}{n_{\mathrm{te}}}\right\}\right) \le \eta. \tag{C.11}$$

Choosing $\eta = n^{-A}$ for $A > 0$ provides the desired tail bound (for a modified constant $C_1 > 0$)

$$\mathbb{P}\left(\Delta_n^{\mathrm{add}} \ge C_1\widehat{\sigma}_{\Xi}\max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\}\right) \le n^{-A}.$$

Expectation bound. We now turn to bounding $\mathbb{E}[\Delta_n^{\mathrm{add}}]$. Define the event

$$\mathcal{B}_n^{\complement} := \left\{\Delta_n^{\mathrm{add}} \ge C_1 C_2\max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\}\right\}.$$

Since $\mathbb{P}(\widehat{\sigma}_n \ge C_2) \le n^{-A}$, combining this with (C.11), we conclude that $\mathbb{P}(\mathcal{B}_n^{\complement}) \le 2n^{-A}$. For the case of $\mathtt{CEN} = \mathtt{MOM}$, the proof follows from that of Lemma 3.2.5. This follows because bounded ψ_1 norm implies bounded L_2 norm.

We can bound $\mathbb{E}[\Delta_n^{\mathrm{add}}]$ by breaking the expected value as

$$\begin{aligned}
\mathbb{E}[\Delta_n^{\mathrm{add}}] &= \mathbb{E}[\Delta_n^{\mathrm{add}}\mathbb{1}_{\mathcal{B}_n}] + \mathbb{E}[\Delta_n^{\mathrm{add}}\mathbb{1}_{\mathcal{B}_n^{\complement}}] \\
&\le C_1 C_2\max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\} + \left(\mathbb{E}[(\Delta_n^{\mathrm{add}})^t]\right)^{1/t}\left(\mathbb{P}(\mathcal{B}_n^c)\right)^{1/r} \\
&\le C_1 C_2\max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\} + \left(\mathbb{E}[(\Delta_n^{\mathrm{add}})^t]\right)^{1/t}(2n^{-A})^{1/r},
\end{aligned} \tag{C.12}$$

for Hölder conjugates $t, r \ge 2$ satisfying $1/t + 1/r = 1$. Observe now that

$$\begin{aligned}
\mathbb{E}[(\Delta_n^{\mathrm{add}})^t] &\le |\Xi|\max_{\xi\in\Xi}\mathbb{E}\left[\left|\widehat{R}(\widehat{f}^{\xi}) - R(\widehat{f}^{\xi})\right|^t\right] \\
&\le |\Xi|\max_{\xi\in\Xi}\mathbb{E}\left[\mathbb{E}\left[\left|\widehat{R}(\widehat{f}^{\xi}) - R(\widehat{f}^{\xi})\right|^t \mid \mathcal{D}_{\mathrm{tr}}\right]\right] \\
&\le C_3|\Xi|\max_{\xi\in\Xi}\mathbb{E}\left[\widehat{\sigma}_{\xi}^t\max\left\{\left(\frac{t}{n_{\mathrm{te}}}\right)^{t/2}, \left(\frac{t}{n_{\mathrm{te}}}\right)^{t}\right\}\right],
\end{aligned}$$

where the last inequality follows from integrating the quantile bound in (C.10) and C_3 is a constant potentially larger than C_1. Substituting this bound in (C.12), we obtain the desired expectation bound

$$\mathbb{E}[\Delta_n^{\mathrm{add}}] \le C_1 C_2\max\left\{\sqrt{\frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}}, \frac{\log\left(|\Xi|n^A\right)}{n_{\mathrm{te}}}\right\} + C_3 n^{-A/r}|\Xi|^{1/t}\max\left\{\sqrt{\frac{t}{n_{\mathrm{te}}}}, \frac{t}{n_{\mathrm{te}}}\right\}\max_{\xi\in\Xi}\left(\mathbb{E}[\widehat{\sigma}_{\xi}^t]\right)^{1/t}.$$

for $t, r \ge 2$ such that $1/r + 1/t = 1$. This completes the proof.

C.1.3 Proof of Lemma 3.2.5

Tail bound. The proof is similar to the proof of Lemma 3.2.4. Our main workhorse is going to be Lemma C.8.2. We use $\eta = \left(|\Xi| n^A\right)^{-1}$ in Algorithm 1. Applying the lemma with such η on the random variables $\ell(Y_j, \widehat{f}^\xi(X_j)), j \in \mathcal{I}_{\mathrm{te}}$ conditionally on $\mathcal{D}_{\mathrm{tr}}$, for each $\xi \in \Xi$ we get the tail bound

$$\mathbb{P}\left(\left| \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{j \in \mathcal{I}_{\mathrm{te}}} \ell(Y_j, \widehat{f}^\xi(X_j)) - R(\widehat{f}^\xi) \right| \geq C_1 \widehat{\sigma}_\xi \sqrt{\frac{\log(|\Xi| n^A)}{|\mathcal{D}_{\mathrm{te}}|}} \;\middle|\; \mathcal{D}_{\mathrm{tr}} \right) \leq \frac{n^{-A}}{|\Xi|}$$

for some absolute constant $C_1 > 0$. In other words,

$$\mathbb{P}\left(\left| \widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi) \right| \geq C_1 \widehat{\sigma}_\xi \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \;\middle|\; \mathcal{D}_{\mathrm{tr}} \right) \leq \frac{n^{-A}}{|\Xi|}.$$

Integrating out $\mathcal{D}_{\mathrm{tr}}$ and applying union bound over $\xi \in \Xi$ then leads to the uniform bound

$$\mathbb{P}\left(\max_{\xi \in \Xi} \left| \widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi) \right| \geq C_1 \max_{\xi \in \Xi} \widehat{\sigma}_\xi \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right) \leq n^{-A}. \tag{C.13}$$

Substituting for the definitions of Δ_n^{add} and $\widehat{\sigma}_\Xi$ gives the desired tail bound

$$\mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1 \widehat{\sigma}_\Xi \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right) \leq n^{-A}. \tag{C.14}$$

Expectation bound. For bounding $\mathbb{E}[\Delta_n^{\mathrm{add}}]$, we again follow similar strategy as in the proof of Lemma 3.2.4. In order to bound certain expectations, we begin by extending the tail bound (C.14). From the assumption, $\mathbb{P}(\widehat{\sigma}_\Xi \geq C_2) \leq n^{-A}$ for a constant $C_2 > 0$. For such a constant, consider the event

$$\mathcal{B}_n^{\complement} := \left\{ \Delta_n^{\mathrm{add}} \geq C_1 C_2 \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right\}.$$

Conditioning on the event $\{\widehat{\sigma}_\Xi \geq C_2\}$, we can bound the probability of $\mathcal{B}_n^{\complement}$ as follows:

$$\begin{aligned} \mathbb{P}(\mathcal{B}_n^{\complement}) &= \mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1 C_2 \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}}, \widehat{\sigma}_\Xi \leq C_2 \right) + \mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1 C_2 \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}}, \widehat{\sigma}_\Xi \geq C_2 \right) \\ &\leq \mathbb{P}\left(\Delta_n^{\mathrm{add}} \geq C_1 \widehat{\sigma}_\Xi \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right) + \mathbb{P}\left(\widehat{\sigma}_n \geq C_2 \right) \leq \frac{2}{n^A}, \end{aligned}$$

where we used the bound from (C.14). We are now ready to bound $\mathbb{E}[\Delta_n^{\mathrm{add}}]$ by splitting using the event $\mathcal{B}_n^{\complement}$. We have

$$\begin{aligned} \mathbb{E}\left[\Delta_n^{\mathrm{add}}\right] &= \mathbb{E}\left[\Delta_n^{\mathrm{add}} \mathbb{1}_{\mathcal{B}_n}\right] + \mathbb{E}\left[\Delta_n^{\mathrm{add}} \mathbb{1}_{\mathcal{B}_n^{\complement}}\right] \\ &\leq C_1 C_2 \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} + \left(\mathbb{P}(\mathcal{B}_n^{\complement})\right)^{1/2} \left(\mathbb{E}[|\Delta_n^{\mathrm{add}}|^2]\right)^{1/2} \\ &\leq C_1 C_2 \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} + \left(2 n^{-A}\right)^{1/2} \left(\mathbb{E}[|\Delta_n^{\mathrm{add}}|^2]\right)^{1/2} \end{aligned} \tag{C.15}$$

where in the first inequality, we used Cauchy-Schwartz inequality for the second term. It remains to bound $\mathbb{E}[|\Delta_n^{\text{add}}|^2]$, which we do below. We have

$$\mathbb{E}[\left|\Delta_n^{\text{add}}\right|^2] = \mathbb{E}\left[\max_{\xi\in\Xi}\left|\widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi)\right|^2\right] \le |\Xi| \max_{\xi\in\Xi}\mathbb{E}\left[|\widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi)|^2\right].$$

For bounding the second term, recall that the `MOM` procedure computes $\widehat{R}(\widehat{f}^\xi)$ as the median of empirical means computed on B partitions of the test data. For each of the B partitions, the variance of the empirical mean is $\widehat{\sigma}_\xi^2/(n_{\text{te}}/B)$. To bound the variance of the median of means on B partitions, we invoke Theorem 1 of Gribkova (2020) (with $k = 2$, $\rho = 1$, and i corresponding to the median position). Note that each of the B empirical means are independent and identically distributed. This provides

$$\mathbb{E}\left[\left|\widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi)\right|^2 \,\middle|\, \mathcal{D}_{\text{tr}}\right] \le C\left(\frac{\widehat{\sigma}_\xi^2}{n_{\text{te}}/B}\right) \le C\frac{B\widehat{\sigma}_\xi^2}{n_{\text{te}}}.$$

for some absolute constant C. Thus,

$$\begin{aligned}\left(\mathbb{E}\left[|\Delta_n^{\text{add}}|^2\right]\right)^{1/2} &\le C\left(|\Xi|\frac{B}{n_{\text{te}}}\max_{\xi\in\Xi}\mathbb{E}[\widehat{\sigma}_\xi^2]\right)^{1/2} \\ &\le C|\Xi|^{1/2}\sqrt{\frac{B}{n_{\text{te}}}}\max_{\xi\in\Xi}\left(\mathbb{E}[\widehat{\sigma}_\xi^2]\right)^{1/2}\end{aligned}$$

Recalling $B = \lceil 8\log(|\Xi|n^A)\rceil$ and combining this bound with (C.15), we finally have the desired expectation bound

$$\mathbb{E}\left[\Delta_n^{\text{add}}\right] \le C_1C_2\sqrt{\frac{\log(|\Xi|n^A)}{n_{\text{te}}}} + C_3n^{-A/2}|\Xi|^{1/2}\sqrt{\frac{\log(|\Xi|n^A)}{n_{\text{te}}}}\max_{\xi\in\Xi}\left(\mathbb{E}[\widehat{\sigma}_\xi^2]\right)^{1/2}.$$

for some absolute constant $C_3 > 0$. This completes the proof.

C.1.4 Proof of Lemma 3.2.9

As argued in the proof of Lemma 3.2.4, using Lemma C.8.1, for any $A > 0$, we have the tail bound:

$$\mathbb{P}\left(\left|\widehat{R}(\widehat{f}^\xi) - R(\widehat{f}^\xi)\right| \ge C\widehat{\sigma}_\xi \max\left\{\sqrt{\frac{\log(|\Xi|n^A)}{|\mathcal{D}_{\text{te}}|}}, \frac{\log(|\Xi|n^A)}{|\mathcal{D}_{\text{te}}|}\right\} \,\middle|\, \mathcal{D}_{\text{tr}}\right) \le \frac{n^{-A}}{|\Xi|}$$

for some universal constant $C > 0$. By diving $R(\widehat{f}^\xi)$ on the both side of error event, and denoting $\widehat{\sigma}_\xi/R(\widehat{f}^\xi)$ by $\widehat{\kappa}_\xi$, equivalently we have

$$\mathbb{P}\left(\left|\frac{\widehat{R}(\widehat{f}^\xi)}{R(\widehat{f}^\xi)} - 1\right| \ge C\widehat{\kappa}_\xi \max\left\{\sqrt{\frac{\log(|\Xi|n^A)}{|\mathcal{D}_{\text{te}}|}}, \frac{\log(|\Xi|n^A)}{|\mathcal{D}_{\text{te}}|}\right\} \,\middle|\, \mathcal{D}_{\text{tr}}\right) \le \frac{n^{-A}}{|\Xi|}.$$

Integrating over randomness in $\mathcal{D}_{\text{tr}}$, and applying union bound over $\xi \in \Xi$, we obtain

$$\mathbb{P}\left(\max_{\xi\in\Xi}\left|\frac{\widehat{R}(\widehat{f}^\xi)}{R(\widehat{f}^\xi)} - 1\right| \ge C\max_{\xi\in\Xi}\widehat{\kappa}_\xi \max\left\{\sqrt{\frac{\log(|\Xi|n^A)}{n_{\text{te}}}}, \frac{\log(|\Xi|n^A)}{n_{\text{te}}}\right\}\right) \le n^{-A}.$$

In other words, in terms Δ_n^{mul} and $\widehat{\kappa}_\Xi$, we have

$$\mathbb{P}\left(\Delta_n^{\text{mul}} \ge C\widehat{\kappa}_\Xi \max\left\{\sqrt{\frac{\log(|\Xi|n^A)}{n_{\text{te}}}}, \frac{\log(|\Xi|n^A)}{n_{\text{te}}}\right\}\right) \le n^{-A},$$

as desired. This completes the proof.

C.1.5 Proof of Lemma 3.2.10

As argued in the proof of Lemma 3.2.5, using Lemma C.8.2, for any $A > 0$, we have the following tail bound:

$$\mathbb{P}\left(\left| \widehat{R}(\widehat{f}^{\xi}) - R(\widehat{f}^{\xi}) \right| \geq C\widehat{\sigma}_{\xi} \sqrt{\frac{\log(|\Xi| n^A)}{|\mathcal{D}_{\mathrm{te}}|}} \;\middle|\; \mathcal{D}_{\mathrm{tr}} \right) \leq \frac{n^{-A}}{|\Xi|}$$

for some universal constant $C > 0$. By diving $R(\widehat{f}^{\xi})$ on the both side of error event, and denoting $\widehat{\sigma}_{\xi}/R(\widehat{f}^{\xi})$ by $\widehat{\kappa}_{\xi}$, we obtain

$$\mathbb{P}\left(\left| \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} - 1 \right| \geq C\widehat{\kappa}_{\xi} \sqrt{\frac{\log(|\Xi| n^A)}{|\mathcal{D}_{\mathrm{te}}|}} \;\middle|\; \mathcal{D}_{\mathrm{tr}} \right) \leq \frac{n^{-A}}{|\Xi|}.$$

Integrating over randomness in $\mathcal{D}_{\mathrm{tr}}$, and applying union bound over $\xi \in \Xi$, this implies that

$$\mathbb{P}\left(\max_{\xi \in \Xi} \left| \frac{\widehat{R}(\widehat{f}^{\xi})}{R(\widehat{f}^{\xi})} - 1 \right| \geq C \max_{\xi \in \Xi} \widehat{\kappa}_{\xi} \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right) \leq n^{-A}.$$

Writing in terms Δ_n^{mul} and $\widehat{\kappa}_{\Xi}$, we arrive at the desired bound:

$$\mathbb{P}\left(\Delta_n^{\mathrm{mul}} \geq C\widehat{\kappa}_{\Xi} \sqrt{\frac{\log(|\Xi| n^A)}{n_{\mathrm{te}}}} \right) \leq n^{-A}.$$

This finishes the proof.

C.1.6 Proof of Proposition 3.2.14

Part 1. For the first part, observe that $|\ell(Y_0, \widehat{f}(X_0))| = \max\{0, 1 - Y_0 \widehat{f}(X_0)\} \leq 2$ assuming $|Y_0| \leq 1$ and $|\widehat{f}(X_0)| \leq 1$. For a bounded random variable Z, $\|Z\|_{\psi_2} \lesssim \|Z\|_{\infty}$ (see, e.g., Example 2.5.8 of Vershynin (2018)). Thus, the random variable $\ell(Y_0, \widehat{f}(X_0))$ is conditionally sub-Gaussian with sub-Gaussian norm 2 (up to constants), and consequently sub-exponential with the same sub-exponential norm upper bound. The conditional L_2 norm bound follows similarly.

Part 2. The second part follows in the same vein by noting that $\ell(Y_0, \widehat{f}(X_0)) = \mathbb{1}_{Y_0 \neq \widehat{f}(X_0)}$ only takes values 0 or 1, and Bernoulli random variables are sub-Gaussian with sub-Gaussian norm 1 (up to constants) and hence sub-exponential with the same sub-exponential norm upper bound. The bound on the conditional L_2 norm follows analogously.

C.1.7 Proof of Theorem 3.2.15

An outline for the proof is already provided in Section 3.2.3. The theorem follows by combining the additive form of the oracle inequality from Proposition 3.2.1, along with the probabilistic bounds on Δ^{add} from Lemmas 3.2.4 and 3.2.5, and the bounds on conditional ψ_1 and L_2 norm bounds from Proposition 3.2.14.

C.1.8 Proof of Proposition 3.2.16

Part 1. For the first part, we bound the ψ_1 norm of the squared error by the squared ψ_2 norm of the error to get

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} = \|(Y_0 - X_0^\top \widehat{\beta})^2\|_{\psi_1|\mathcal{D}_n} \leq \|Y_0 - X_0^\top \widehat{\beta}\|_{\psi_2|\mathcal{D}_n}^2, \tag{C.16}$$

where the inequality follows by Lemma 2.7.7 of Vershynin (2018). Note that for any $\beta \in \mathbb{R}^p$, we have

$$(Y_0 - X_0^\top \widehat{\beta}) = (Y_0 - X_0^\top \beta) + X_0^\top (\beta - \widehat{\beta}). \tag{C.17}$$

Because $\|Z_1 + Z_2\|_{\psi_2} \le \|Z_1\|_{\psi_2} + \|Z\|_{\psi_2}$ we can bound

$$\|Y_0 - X_0^\top \widehat{\beta}\|_{\psi_2|\mathcal{D}_n} \le \|Y_0 - X_0^\top \beta\|_{\psi_2} + \|X_0^\top(\beta - \widehat{\beta})\|_{\psi_2|\mathcal{D}_n}. \tag{C.18}$$

Noting that $Y_0 - X_0^\top \beta = (Y_0, X_0)^\top(1, -\beta)$ and $(\beta - \widehat{\beta})$ is a fixed vector conditioned on $\mathcal{D}_n$, by using $\psi_2 - L_2$ equivalence on (X_0, Y_0), we have

$$\|Y_0 - X_0^\top \beta\|_{\psi_2} \le \tau \|Y_0 - X_0^\top \beta\|_{L_2} \quad \text{and} \quad \|X_0^\top(\beta - \widehat{\beta})\|_{\psi_2|\mathcal{D}_n} \le \tau \|X_0^\top(\beta - \widehat{\beta})\|_{L_2|\mathcal{D}_n} = \tau\|\widehat{\beta} - \beta\|_\Sigma, \tag{C.19}$$

where in the last inequality we used the fact that $\mathbb{E}[X_0] = 0$ and $\mathbb{E}[X_0 X_0^\top] = \Sigma$. Thus, combining (C.16), (C.18), and (C.19), for $\beta \in \mathbb{R}^p$, we have

$$\|\ell(Y_0 - X_0^\top \widehat{\beta})\|_{\psi_1|\mathcal{D}_n} \le (\|Y_0 - X_0^\top \beta\|_{\psi_2} + \|\widehat{\beta} - \beta\|_\Sigma)^2.$$

Taking infimum over β, we have that for squared loss

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n} \le \tau^2 \inf_{\beta \in \mathbb{R}^p} (\|Y_0 - X_0^\top \beta\|_{\psi_2} + \|\widehat{\beta} - \beta\|_\Sigma)^2,$$

as desired. This completes the proof of the first inequality in (3.15). For the second inequality in (3.15), using the $\psi_2 - L_2$ equivalence on the vector (X_0, Y_0), observe that

$$\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n] = \mathbb{E}[(Y_0 - X_0^\top \widehat{\beta})^2 \mid \mathcal{D}_n] = \|Y_0 - X_0^\top\|^2_{L_2|\mathcal{D}_n}. \tag{C.20}$$

Hence, from (C.16) and (C.20), we have

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \frac{\|Y_0 - X_0^\top \widehat{\beta}\|^2_{\psi_2|\mathcal{D}_n}}{\|Y_0 - X_0^\top \widehat{\beta}\|^2_{L_2|\mathcal{D}_n}} = \left(\frac{\|(Y_0, X_0)(1, -\widehat{\beta})\|_{\psi_2|\mathcal{D}_n}}{\|(Y_0, X_0)(1, -\widehat{\beta})\|_{L_2|\mathcal{D}_n}}\right)^2 \le \tau^2,$$

as desired. This completes the proof of the first part.

Part 2. We now turn to the second part to bound the conditional L_2 norm of the square loss. For the square loss, note that

$$\|\ell(Y_0, \widehat{f}(X_0))\|^2_{L_2|\mathcal{D}_n} = \mathbb{E}[(Y_0 - \widehat{f}(X_0))^4 \mid \mathcal{D}_n]. \tag{C.21}$$

Using the decomposition (C.17) and triangle inequality with respect to the L_4 norm, we have

$$\mathbb{E}[(Y_0 - X_0^\top \widehat{\beta})^4 \mid \mathcal{D}_n]^{1/4} \le \mathbb{E}[(Y_0 - X_0^\top \beta)^4 \mid \mathcal{D}_n]^{1/4} + \mathbb{E}[X_0^\top(\beta - \widehat{\beta})^4 \mid \mathcal{D}_n]^{1/4} \tag{C.22}$$

Using the $L_4 - L_2$ equivalence for (Y_0, X_0), we can bound

$$\|Y_0 - X_0^\top \beta\|_{L_4} \le \tau \|Y_0 - X_0^\top \beta\|_{L_2} \quad \text{and} \quad \|X_0^\top(\beta - \widehat{\beta})\|_{L_4|\mathcal{D}_n} \le \tau \|X_0^\top(\beta - \widehat{\beta})\|_{L_2|\mathcal{D}_n}. \tag{C.23}$$

Thus, combining (C.21), (C.22), and (C.23), we have for any $\beta \in \mathbb{R}^p$,

$$\|(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n} \le (\tau\|Y_0 - X_0^\top \beta\|_{L_2} + \tau\|\widehat{\beta} - \beta\|_\Sigma)^2 \le \tau^2(\|Y_0 - X_0^\top \beta\|_{L_2} + \|\widehat{\beta} - \beta\|_\Sigma)^2.$$

This completes the proof of first inequality in (3.16). For the second inequality of (3.16), note that

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{L_2|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \frac{\|Y_0 - \widehat{f}(X_0)\|^2_{L_4|\mathcal{D}_n}}{\|Y_0 - \widehat{f}(X_0)\|^2_{L_2|\mathcal{D}_n}} = \left(\frac{\|(Y_0, X_0)(1, -\widehat{\beta})\|_{L_4|\mathcal{D}_n}}{\|(Y_0, X_0)(1, -\widehat{\beta})\|_{L_2|\mathcal{D}_n}}\right)^2 \le \tau^2.$$

This concludes the proof of the second part.

C.1.9 Proof of Proposition 3.2.17

The proof is similar to that of Proposition 3.2.16.

Part 1. From the decomposition (C.17) and the triangle inequality on ψ_1 norm, we have for any $\beta \in \mathbb{R}^p$,

$$\|Y_0 - X_0^\top \widehat{\beta}\|_{\psi_1 \mid \mathcal{D}_n} \le \|Y_0 - X_0^\top \beta\|_{\psi_1} + \|X_0^\top (\beta - \widehat{\beta})\|_{\psi_1 \mid \mathcal{D}_n}. \tag{C.24}$$

Using the $\psi_1 - L_1$ equivalence of (X_0, Y_0), note that

$$\|Y_0 - X_0^\top \beta\|_{\psi_1} \le \tau \|Y_0 - X_0^\top \beta\|_{L_1} \quad \text{and} \quad \|X_0^\top (\beta - \widehat{\beta})\|_{\psi_1 \mid \mathcal{D}_n} \le \tau \|X_0^\top (\beta - \widehat{\beta})\|_{\psi_1 \mid \mathcal{D}_n}. \tag{C.25}$$

Thus, from (C.24) and (C.25), for any $\beta \in \mathbb{R}^p$, we have

$$\|Y_0 - X_0^\top \widehat{\beta}\|_{\psi_1 \mid \mathcal{D}_n} \le \tau(\|Y_0 - X_0^\top \beta\|_{L_1} + \|X_0^\top (\widehat{\beta} - \beta)\|_{L_1 \mid \mathcal{D}_n}).$$

Now taking infimum over $\beta \in \mathbb{R}^p$ yields the first inequality of (3.18). To show the second inequality, observe that

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1 \mid \mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \frac{\|Y_0 - X_0^\top \widehat{\beta}\|_{\psi_1 \mid \mathcal{D}_n}}{\|Y_0 - X_0^\top \widehat{\beta}\|_{L_1 \mid \mathcal{D}_n}} \le \tau,$$

as desired. This finishes the proof.

Part 2. The second part follows analogously to the first part by using the $L_2 - L_1$ equivalence on (X_0, Y_0).

C.1.10 Proof of Proposition 3.2.18

We start by writing the loss as

$$\begin{aligned} \ell(Y_0, \widehat{f}(X_0)) &= Y_0 \log(1 + e^{-X_0^\top \widehat{\beta}}) + (1 - Y_0) \log(1 + e^{X_0^\top \widehat{\beta}}) \\ &= \mathrm{KL}(Y_0,\, (1 + \exp(-X_0^\top \widehat{\beta}))^{-1}). \end{aligned}$$

Observe that the loss is non-negative since $\log(1 + e^t) \ge 0$ for all t.

Upper bounds on ψ_1 and L_2 norms. We will first obtain an upper on the loss and consequently on the ψ_1 and L_2 norms of the loss. Because Y_0 takes values 0 or 1, we have that

$$\begin{aligned} \ell(Y_0, \widehat{f}(X_0)) &\le \max\left\{ \log(1 + e^{-X_0^\top \widehat{\beta}}), \log(1 + e^{X_0^\top \widehat{\beta}}) \right\} \\ &\le \log(1 + e^{|X_0^\top \widehat{\beta}|}), \end{aligned}$$

where the second inequality follows since $t \mapsto e^t$ is monotonically increasing in t. Now using the following bound on $\log(1 + e^{|t|})$:

$$\log(1 + e^{|t|}) \le \begin{cases} \log 2 & \text{if } e^{|t|} \le 1 \\ \log(2e^{|t|}) = \log 2 + |t| & \text{otherwise,} \end{cases}$$

we can upper bound the loss by

$$\ell(Y_0, \widehat{f}(X_0)) \le |X_0^\top \widehat{\beta}| + \log 2.$$

Hence, we can upper bound the ψ_1 and L_2 norm of the loss as follows:

$$\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1 \mid \mathcal{D}_n} \le \log(2) + \|X_0^\top \widehat{\beta}\|_{\psi_1 \mid \mathcal{D}_n}, \tag{C.26}$$

$$(\mathbb{E}[\ell^2(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n])^{1/2} \le \log(2) + (\mathbb{E}[|X_0^\top \widehat{\beta}|^2 \mid \mathcal{D}_n])^{1/2}. \tag{C.27}$$

Lower bound on expectation. Next we obtain a lower bound on $\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]$. Setting $p(x) = \mathbb{E}[Y_0 | X_0 = x]$, it is clear that

$$\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n, X_0] = p(X_0)\log(1 + \exp(-X_0^\top \widehat{\beta})) + (1 - p(X_0))\log(1 + \exp(X_0^\top \widehat{\beta})).$$

Because $0 < p_{\min} \le \min\{p(x), 1 - p(x)\}$ for all x, we have

$$\begin{aligned}
\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n] &\ge p_{\min}\, \mathbb{E}[\max\{\log(1 + \exp(-X_0^\top \widehat{\beta})), \log(1 + \exp(X_0^\top \widehat{\beta}))\} \mid \mathcal{D}_n] \\
&= p_{\min}\, \mathbb{E}[\log(1 + \exp(|X_0^\top \widehat{\beta}|)) \mid \mathcal{D}_n] \\
&\ge \frac{p_{\min}}{2}\, \mathbb{E}[\log(2) + |X_0^\top \widehat{\beta}| \mid \mathcal{D}_n] = \frac{p_{\min}}{2}(\log(2) + \mathbb{E}|X_0^\top \widehat{\beta}|),
\end{aligned} \tag{C.28}$$

where the second equality follows since $t \mapsto e^t$ is monotonically increasing in $t \in \mathbb{R}$, and the last inequality follows from the fact that $1/2 \le \log(1 + \exp(x))/(\log(2) + x) \le 1$ for all $x \ge 0$.

Using (C.26) and (C.28), we have

$$\frac{\|\ell(Y_0, \widehat{f}(X_0))\|_{\psi_1|\mathcal{D}_n}}{\mathbb{E}[\ell(Y_0, \widehat{f}(X_0)) \mid \mathcal{D}_n]} \le \frac{\|X_0^\top \widehat{\beta}\|_{\psi_1|\mathcal{D}_n} + \log(2)}{p_{\min}(\mathbb{E}[|X_0^\top \widehat{\beta}| \mid \mathcal{D}_n] + \log(2))/2} \le \frac{\tau\|X_0^\top \widehat{\beta}\|_{L_1|\mathcal{D}_n} + \log(2)}{p_{\min}(\tau\|X_0^\top \widehat{\beta}\|_{L_1|\mathcal{D}_n} + \log(2))/2} = 2\tau p_{\min}^{-1}.$$

This proves the first part of Proposition 3.2.18. A similar bound holds for the second inequality of Proposition 3.2.18 using upper bound from (C.27) and lower bound (C.28). This completes the proof.

C.1.11 Proof of Theorem 3.2.22

An outline for the proof is provided in Section 3.2.3. The theorem follows by combining the multiplicative form of the oracle inequality from Proposition 3.2.1, along with probabilistic bounds on Δ^{mul} from Lemmas 3.2.9 and 3.2.10, and the bounds on ratio of conditional ψ_1 and L_1 norms, and L_2 and L_1 norms from Proposition 3.2.16.

C.2 Proofs related to risk monotonization for zero-step procedure

C.2.1 Proof of Theorem 3.3.4

An outline for the proof is already provided in Section 3.3.3. For the sake of completeness, we briefly summarize the main steps below.

The deterministic additive and multiplicative oracle risk inequalities from Proposition 3.2.1, along with probabilistic bounds from Lemmas 3.2.4, 3.2.5, 3.2.9 and 3.2.10, provide the following bound on the risk of the zero-step predictor

$$R(\widehat{f}^{\mathrm{zs}}) = \begin{cases} \min_{\xi \in \Xi_n} R(\widehat{f}^{\xi}) + O_p(1)\sqrt{\log n / n_{\mathrm{te}}} & \text{if } \widehat{\sigma}_{\Xi} = O_p(1), \\ \min_{\xi \in \Xi_n} R(\widehat{f}^{\xi})\big(1 + O_p(1)\sqrt{\log n / n_{\mathrm{te}}}\big) & \text{if } \widehat{\kappa}_{\xi} = O_p(1). \end{cases} \tag{C.29}$$

Depending on the value of M, we now bound the term $\min_{\xi \in \Xi_n} R(\widehat{f}^{\xi})$ under the assumptions (DET*) or (DET).

Case of $M = 1$. Under (DET*), we have from (3.33),

$$\min_{\xi \in \Xi_n} R(\widehat{f}^{\xi}) = \min_{\xi \in \Xi_n} R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi,1})) = R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})(1 + o_p(1)). \tag{C.30}$$

Combining (C.30) with (C.29) yields

$$\begin{aligned}
R(\widehat{f}^{\mathrm{zs}}) &= \begin{cases} R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})(1 + o_p(1)) + O_p(1)\sqrt{\log n / n_{\mathrm{te}}} & \text{if } \widehat{\sigma}_{\Xi} = O_p(1) \\ R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})(1 + o_p(1)) & \text{if } \widehat{\kappa}_{\Xi} = O_p(1) \end{cases} \\
&= R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f}) \begin{cases} 1 + o_p(1) + \sqrt{\log n / n_{\mathrm{te}}} / R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f}) & \text{if } \widehat{\sigma}_{\Xi} = O_p(1) \\ 1 + o_p(1) & \text{if } \widehat{\kappa}_{\Xi} = O_p(1). \end{cases}
\end{aligned} \tag{C.31}$$

Thus, under (O1) or (O2), we have $|R(\widehat{f}^{\rm zs}) - R^{\rm det}_{\nearrow}(n; \widetilde{f})| / R^{\rm det}_{\nearrow}(n; \widetilde{f}) = o_p(1)$ as desired.

Case of $M > 1$. Under (DET), we have from (3.32),

$$\min_{\xi \in \Xi_n} R(\widehat{f}^{\xi}) \le R^{\rm det}_{\nearrow}(n; \widetilde{f})(1 + o_p(1)). \tag{C.32}$$

Now similar to the case of $M = 1$, combining (C.32) with (C.29), and under (O1) or (O2), we have that $(R(\widehat{f}^{\rm zs}) - R^{\rm det}_{\nearrow}(n; \widetilde{f}))_+ / R^{\rm det}_{\nearrow}(n; \widetilde{f}) = o_p(1)$ as claimed. This finishes the proof.

C.2.2 Proof of Lemma 3.3.8

Our goal is to verify (DETPA-0), i.e., existence of a deterministic profile $R^{\rm det}(\cdot; \widetilde{f})$ such that for all non-stochastic sequences $\xi_n^\star \in \arg\min_{\xi \in \Xi_n} R^{\rm det}(p_n/n_\xi; \widetilde{f})$ and $1 \le j \le M$,

$$\frac{R(\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n^\star, j})) - R^{\rm det}(p_n/n_{\xi_n^\star}; \widetilde{f})}{R^{\rm det}(p_n/n_{\xi_n^\star}; \widetilde{f})} \xrightarrow{\rm p} 0,$$

as $n \to \infty$ under (PA(γ)). Recall here $\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n^\star, j})$, $1 \le j \le M$, is a predictor trained on the dataset $\mathcal{D}_{\rm tr}^{\xi_n^\star, j}$ of sample size $n_{\xi_n^\star} = n_{\rm tr} - \xi_n^\star \lfloor n^\nu \rfloor$ and feature dimension p_n. We will make a series of reductions to verify (DETPA-0) from the assumptions of Lemma 3.3.8.

First, note that $R(\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n, j}))$ for $1 \le j \le M$ are identically distributed. It thus suffices to pick $j = 1$, which we will do below and drop the index for notational brevity. Second, since $R(\widetilde{f}(\cdot; \mathcal{D}_{k_m})) > 0$ for all k_m, it suffices to show that as $n \to \infty$ under (PA(γ)),

$$R(\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n^\star})) - R^{\rm det}(p_n/n_{\xi_n^\star}; \widetilde{f}) \xrightarrow{\rm p} 0, \quad \text{where} \quad \xi_n^\star \in \arg\min_{\xi \in \Xi_n} R^{\rm det}(p_n/n_\xi; \widetilde{f}).$$

More explicitly, that for all $\epsilon > 0$, it suffices to verify that as $n \to \infty$ under (PA(γ)),

$$\mathbb{P}\big(|R(\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n^\star})) - R^{\rm det}(p_n/n_{\xi_n^\star}; \widetilde{f})| \ge \epsilon\big) \to 0, \quad \text{where} \quad \xi_n^\star \in \arg\min_{\xi \in \Xi_n} R^{\rm det}(p_n/n_\xi; \widetilde{f}).$$

Now, we will do our final reduction. Fix $\epsilon > 0$. Define a sequence $\{h_n(\epsilon)\}_{n \ge 1}$ as follows:

$$h_n(\epsilon) := \mathbb{P}\big(|R(\widetilde{f}(\cdot; \mathcal{D}_{\rm tr}^{\xi_n^\star})) - R^{\rm det}(p_n/n_{\xi_n^\star}; \widetilde{f})| \ge \epsilon\big).$$

From the discussion in Section 3.3.3.1, we know that $p_n/n_{\xi_n^\star}$ may not necessarily converge as $n \to \infty$. But applying Lemma C.6.3 on the sequence $\{h_n(\epsilon)\}_{n \ge 1}$, in order to verify that $h_n(\epsilon) \to 0$ as $n \to \infty$, it suffices to show that for any index subsequence $\{n_k\}_{k \ge 1}$, there exists a further subsequence $\{n_{k_l}\}_{l \ge 1}$ such that $h_{n_{k_l}}(\epsilon) \to 0$ as $l \to 0$. Towards that goal, fix an arbitrary index subsequence $\{n_k\}_{k \ge 1}$. We will appeal to Lemma C.6.5 to construct the desired subsequence $\{n_{k_l}\}_{l \ge 1}$ along which we will argue that $h_{n_{k_l}} \to 0$ provided the assumptions of Lemma 3.3.8 are satisfied. In particular, from Lemma C.6.1, note that since $n_{\rm tr}/n \to 1$ as $n \to \infty$, we have $\Pi_{\Xi_n}(\zeta) \to \zeta$ for any $\zeta \in [\gamma, \infty]$ as $n \to \infty$. Now applying Lemma C.6.5 on $R^{\rm det}(\cdot; \widetilde{f})$ and the grid Ξ_n guarantees that for any subsequence $\{p_{n_k}/n_{\xi_{n_k}^\star}\}_{k \ge 1}$, there exists a subsequence $\{p_{n_{k_l}}/n_{\xi_{n_{k_l}}^\star}\}_{l \ge 1}$ such that as $l \to \infty$,

$$\frac{p_n}{n_{\xi_{n_{k_l}}^\star}} \to \phi \in \arg\min_{\zeta \in [\gamma, \infty]} R^{\rm det}(\zeta; \widetilde{f}). \tag{C.33}$$

We will now show that $h_{n_{k_l}}(\epsilon) \to 0$ as $l \to \infty$ if the profile convergence assumption (DETPAR-0) of Lemma 3.3.8 is satisfied, i.e., for a dataset $\mathcal{D}_{k_m}$ with k_m observations and p_m features, there exists $R^{\rm det}(\cdot; \widetilde{f})$ such that

$$R(\widetilde{f}(\cdot; \mathcal{D}_{k_m})) \xrightarrow{\rm p} R^{\rm det}(\phi; \widetilde{f}) \quad \text{whenever} \quad \frac{p_m}{k_m} \to \phi \in \arg\min_{\zeta \in [\gamma, \infty]} R^{\rm det}(\zeta; \widetilde{f}). \tag{C.34}$$

This follows easily because the profile convergence condition (C.34) implies that as $l \to \infty$,

$$\mathbb{P}\left(\left|R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi^\star_{n_{k_l}}})) - R^{\mathrm{det}}(\phi; \widetilde{f})\right| \geq \epsilon\right) \to 0 \quad \text{whenever} \quad \frac{p_n}{n_{\xi^\star_{n_{k_l}}}} \to \phi \in \operatorname*{arg\,min}_{\zeta \in [\gamma, \infty]} R^{\mathrm{det}}(\zeta; \widetilde{f}).$$

But since $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is continuous at ϕ, and $p_n / n_{\xi^\star_{n_{k_l}}} \to \phi \in \arg\min_{\zeta \in [\gamma, \infty]} R^{\mathrm{det}}(\zeta; \widetilde{f})$ as $l \to \infty$ from (C.33) this implies that, as $l \to \infty$,

$$\mathbb{P}\left(\left|R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi^\star_{n_{k_l}}})) - R^{\mathrm{det}}(p_n / n_{\xi^\star_{n_{k_l}}}; \widetilde{f})\right| \geq \epsilon\right) = h(n_{k_l}) \to 0.$$

This concludes the proof.

C.2.3 Proof of Proposition 3.3.9

In order to verify lower semicontinuity of h, if suffices to show that for any $t \in \mathbb{R}_{\geq 0}$, the set $\{x : h(x) \leq t\}$ is closed. Because $\lim_{x \to b^-} h(x) = \infty$ and h continuous on $[a, b)$, there exists $b_-(t) < b$ such that $h(x) > t$ for all $x > b_-(t)$. Similarly, there exists $b_+(t) > b$ such that $h(x) > t$ for all $x < b_+(t)$. Note that

$$\{x : h(x) \leq t\} = \{x : h|_{[a, b_-(t)]}(x) \leq t\} \cup \{x : h|_{[b_+(t), c]}(x) \leq t\}.$$

Because h is continuous on $[a, b_-(t)]$ and $[b_+(t), c]$, it is also lower semicontinuous on these intervals, and hence the corresponding level sets are closed. Because the intersection of two closed sets is closed, the statement follows.

C.2.4 Proof of Proposition 3.3.10

The proof builds on similar idea as that in the proof of Lemma C.6.7 and employs a proof by contradiction. However, since the random functions in this case (which are conditional prediction risks) are not simply indexed by n (but also by other properties of the data distributions), we will need to do a bit more work.

We wish to show that $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is continuous on $\mathcal{I} \in (0, \infty)$. We will first show that $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is $\mathbb{Q}$-continuous (see Definition C.6.8) on $\mathcal{I}$ and use Lemma C.6.9 to lift $\mathbb{Q}$-continuity to $\mathbb{R}$-continuity. Towards showing $\mathbb{Q}$-continuity, for the sake of contradiction, suppose $R^{\mathrm{det}}(\cdot; \widetilde{f})$ is $\mathbb{Q}$-discontinuous at some point $\phi_\infty \in \mathcal{I}$. This implies that there exists a sequence $\{\phi_r\}_{r \geq 1}$ in $\mathbb{Q}_{>0}$ such that $\phi_r \to \phi_\infty$, but for some $\epsilon > 0$ and all $r \geq 1$,

$$R^{\mathrm{det}}(\phi_r; \widetilde{f}) \notin [R^{\mathrm{det}}(\phi_\infty; \widetilde{f}) - 2\epsilon, R^{\mathrm{det}}(\phi_\infty; \widetilde{f}) + 2\epsilon]. \tag{C.35}$$

(Note that $R^{\mathrm{det}}(\phi_r; \widetilde{f}) \not\to R^{\mathrm{det}}(\phi_\infty; \widetilde{f})$ as $\phi_r \to \phi_\infty$.) The proof strategy is now to construct a sequence of datasets $\{\mathcal{D}'_{k_m}\}_{m \geq 1}$ whose aspects ratios p_m / k_m converge to ϕ_∞, but the conditional prediction risks $R(\widetilde{f}(\cdot; \mathcal{D}'_{k_m}))$ of predictors $\widetilde{f}(\cdot; \mathcal{D}'_{k_m})$ trained on these datasets do not converge to $R^{\mathrm{det}}(\phi_\infty; \widetilde{f})$, thereby supplying a contradiction to the hypothesis of continuous convergence of $R(\widetilde{f}(\cdot; \mathcal{D}'_{k_m}))$ to $R^{\mathrm{det}}(\phi_\infty; \widetilde{f})$. We will construct such a sequence of datasets below.

For every $r \geq 1$, construct a sequence of datasets $\{\mathcal{D}^{\phi_r}_{k_m}\}_{m \geq 1}$ with k_m observations and $p_m = \phi_i k_m$ features. (Since $\phi_r \in \mathbb{Q}_{>0}$, the resulting p_m is a positive integer.) See Figure C.1 for a visual illustration. For every $r \geq 1$, from the assumption of Proposition 3.3.10, we have that

$$R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_r}_{k_m})) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi_r; \widetilde{f}) \tag{C.36}$$

as $k_m, p_m \to \infty$ because $p_m / k_m \to \phi_r$ as $m \to \infty$. Now, fix $p \in (0, 1)$. For $r = 1$, the convergence in (C.36) guarantees that there exists an integer $m_1 \geq 1$ such that the event

$$\Omega_{m_1} := \{|R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_1}_{k_{m_1}})) - R^{\mathrm{det}}(\phi_1; \widetilde{f})| \leq \epsilon\} \tag{C.37}$$

has probability at least p. In addition, on the event Ω_{m_1}, by the triangle inequality we have that

$$|R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_1}_{k_{m_1}})) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| \geq |R^{\mathrm{det}}(\phi_1; \widetilde{f}) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| - |R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_1}_{k_{m_1}})) - R^{\mathrm{det}}(\phi_1; \widetilde{f})| > \epsilon, \tag{C.38}$$

where the second inequality follows by using (C.35) and (C.37). Next, for $r \geq 2$, let $m_r > m_{r-1}$ be an integer such that the event

$$\Omega_{m_r} := \{|R(\widetilde{f}(\cdot; \mathcal{D}_{k_{m_r}}^{\phi_r})) - R^{\mathrm{det}}(\phi_r; \widetilde{f})| \leq \epsilon\} \tag{C.39}$$

has probability at least p. Such sequence of integers $\{m_r\}_{r \geq 2}$ and the associated events $\{\Omega_{m_r}\}_{r \geq 2}$ indeed exist as a consequence of the convergence in (C.36) for $r \geq 2$. On each Ω_{m_r}

$$|R(\widetilde{f}(\cdot; \mathcal{D}_{k_{m_r}}^{\phi_r})) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| > \epsilon$$

by similar reasoning as that for (C.38) using (C.35) and (C.39) for $r \geq 2$. Moreover, note that since $m_r > m$, $m_r \to \infty$ as $r \to \infty$.

Consider now a sequence of datasets $\{\mathcal{D}'_{k_m}\}_{m \geq 1}$ such that:

1. The first m_1 datasets are $\{\mathcal{D}_{k_m}^{\phi_1}\}_{m=1}^{m_1}$ that have k_m number of observations and $p_m = \phi_1 k_m$ number of features for $m = 1, \dots, m_1$.
2. The next $m_2 - m_1$ datasets are $\{\mathcal{D}_{k_m}^{\phi_2}\}_{m=m_1+1}^{m_2}$ that have k_m number of observations and $p_m = \phi_2 k_m$ number of features for $m = m_1 + 1, \dots, m_2$.
3. The next $m_3 - m_2$ datasets are $\{\mathcal{D}_{k_m}^{\phi_3}\}_{m=m_2+1}^{m_3}$ that have k_m number of observations and $p_m = \phi_3 k_m$ number of features for $m = m_2 + 1, \dots, m_3$.
4. And so on ...

We will argue now that the sequence of datasets $\{\mathcal{D}'_{k_m}\}_{m \geq 1}$ works for our promised contradiction. Observe that in the construction above the aspect ratios $p_m/k_m \to \phi_\infty$ because $\phi_r \to \phi_\infty$. However, we have that for all $r \geq 1$,

$$\mathbb{P}(|R(\widetilde{f}(\cdot; \mathcal{D}'_{k_{m_r}})) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| > \epsilon) = \mathbb{P}(|R(\widetilde{f}(\cdot; \mathcal{D}_{k_{m_r}})) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| > \epsilon) \geq p.$$

Therefore, there exists an $\epsilon > 0$ for which there is no $M \geq 1$ such that for $m \geq M$,

$$\mathbb{P}(|R(\widetilde{f}(\cdot; \mathcal{D}'_{k_m})) - R^{\mathrm{det}}(\phi_\infty; \widetilde{f})| > \epsilon) < p/2.$$

Hence, we get the desired contraction that

$$R(\widetilde{f}(\cdot; \mathcal{D}'_{k_m})) \overset{\mathrm{p}}{\not\to} R^{\mathrm{det}}(\phi_\infty, \widetilde{f})$$

as $k_m, p_m \to \infty$ and $p_m/k_m \to \phi_\infty$. This completes the proof.

It is worth pointing out that the proof above bears similarity to the proof of Lemma C.6.9. It is possible to combine the two and not have to go through the route of $\mathbb{Q}$-continuity. We, however, find it easier to break them so that the main ideas are easier to digest even though it leads to some repetition of overall proof strategies.

C.2.5 Proof of Theorem 3.3.11

We will split the proof depending on the value of M.

Case of $M = 1$. Consider first the case when $M = 1$. In this case, for every $\xi \in \Xi$, $\widehat{f}^\xi = \widetilde{f}_1^\xi$ (and thus, $\widetilde{f}^\star = \widehat{f}^{\mathrm{cv}}$), which we denote by $\widetilde{f}^\xi$ for simplicity of notation. To bound the desired difference, we break it into three terms:

$$\begin{aligned}\left(R(\widehat{f}^{\mathrm{cv}}) - \min_{\zeta \geq p/n} R^{\mathrm{det}}(\widetilde{f}; \zeta)\right)_+ &= \left(R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widetilde{f}^\xi)\right)_+ \\ &\quad + \left(\min_{\xi \in \Xi} R(\widetilde{f}^\xi) - \min_{\xi \in \Xi} R^{\mathrm{det}}\left(\widetilde{f}; \frac{p_n}{n_\xi}\right)\right)_+ \\ &\quad + \left(\min_{\xi \in \Xi} R^{\mathrm{det}}\left(\widetilde{f}; \frac{p_n}{n_\xi}\right) - \min_{\zeta \geq p/n} R^{\mathrm{det}}(\widetilde{f}; \zeta)\right)_+ .\end{aligned} \tag{C.40}$$

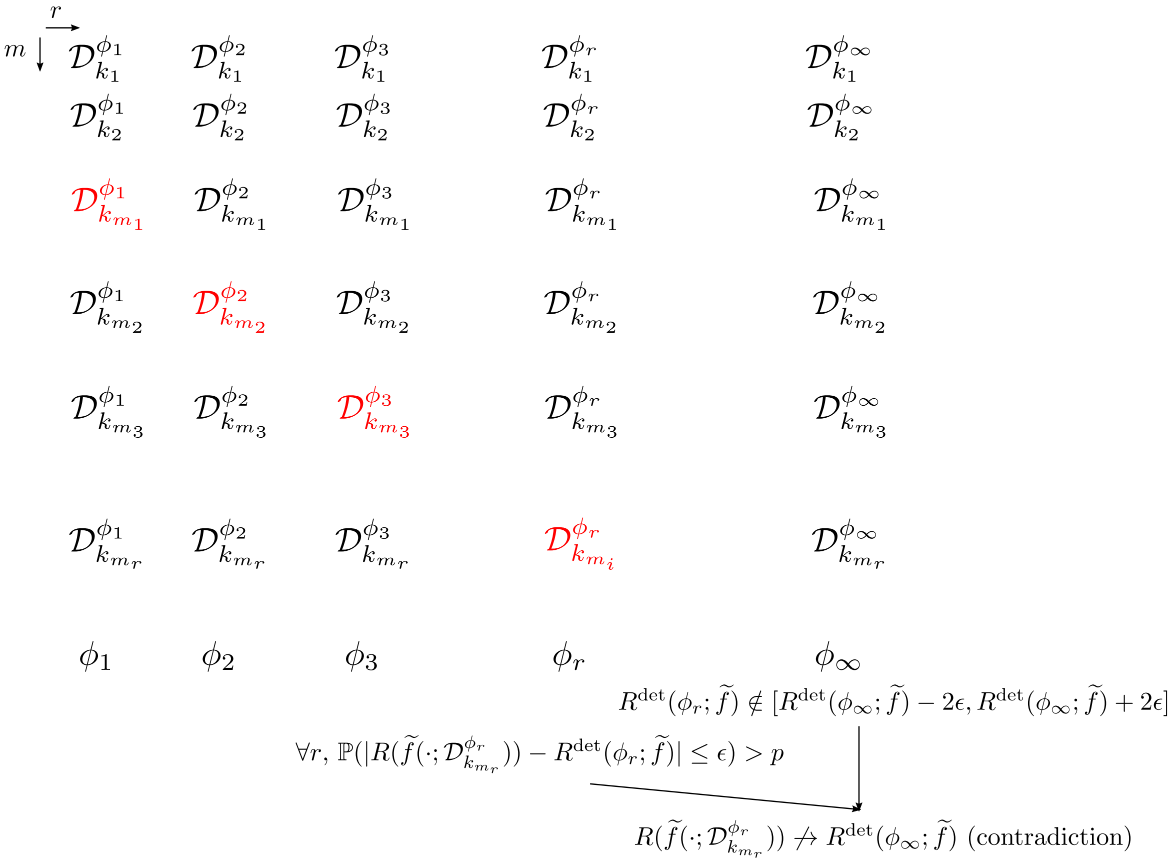

r
m
$\mathcal{D}_{k_1}^{\phi_1}$ $\mathcal{D}_{k_1}^{\phi_2}$ $\mathcal{D}_{k_1}^{\phi_3}$ $\mathcal{D}_{k_1}^{\phi_r}$ $\mathcal{D}_{k_1}^{\phi_\infty}$
$\mathcal{D}_{k_2}^{\phi_1}$ $\mathcal{D}_{k_2}^{\phi_2}$ $\mathcal{D}_{k_2}^{\phi_3}$ $\mathcal{D}_{k_2}^{\phi_r}$ $\mathcal{D}_{k_2}^{\phi_\infty}$
$\mathcal{D}_{k_{m_1}}^{\phi_1}$ $\mathcal{D}_{k_{m_1}}^{\phi_2}$ $\mathcal{D}_{k_{m_1}}^{\phi_3}$ $\mathcal{D}_{k_{m_1}}^{\phi_r}$ $\mathcal{D}_{k_{m_1}}^{\phi_\infty}$
$\mathcal{D}_{k_{m_2}}^{\phi_1}$ $\mathcal{D}_{k_{m_2}}^{\phi_2}$ $\mathcal{D}_{k_{m_2}}^{\phi_3}$ $\mathcal{D}_{k_{m_2}}^{\phi_r}$ $\mathcal{D}_{k_{m_2}}^{\phi_\infty}$
$\mathcal{D}_{k_{m_3}}^{\phi_1}$ $\mathcal{D}_{k_{m_3}}^{\phi_2}$ $\mathcal{D}_{k_{m_3}}^{\phi_3}$ $\mathcal{D}_{k_{m_3}}^{\phi_r}$ $\mathcal{D}_{k_{m_3}}^{\phi_\infty}$
$\mathcal{D}_{k_{m_r}}^{\phi_1}$ $\mathcal{D}_{k_{m_r}}^{\phi_2}$ $\mathcal{D}_{k_{m_r}}^{\phi_3}$ $\mathcal{D}_{k_{m_i}}^{\phi_r}$ $\mathcal{D}_{k_{m_r}}^{\phi_\infty}$
ϕ_1 ϕ_2 ϕ_3 ϕ_r ϕ_∞
$R^{\mathrm{det}}(\phi_r; \widetilde{f}) \notin [R^{\mathrm{det}}(\phi_\infty; \widetilde{f}) - 2\epsilon, R^{\mathrm{det}}(\phi_\infty; \widetilde{f}) + 2\epsilon]$
$\forall r$, $\mathbb{P}(|R(\widetilde{f}(\cdot; \mathcal{D}_{k_{m_r}}^{\phi_r})) - R^{\mathrm{det}}(\phi_r; \widetilde{f})| \le \epsilon) > p$
$R(\widetilde{f}(\cdot; \mathcal{D}_{k_{m_r}}^{\phi_r})) \not\to R^{\mathrm{det}}(\phi_\infty; \widetilde{f})$ (contradiction)

Figure C.1: Illustration of construction of grid of datasets used in the proof of Proposition 3.3.10. (Side note: as can be seen from the figure, the argument bears similarity to the standard diagonalization argument.)

This inequality follows from the fact that $(a + b + c)_+ \le (a)_+ + (b)_+ + (c)_+$ for any $a, b, c \in \mathbb{R}$. We show below that each of the three terms asymptotically vanish in probability as $n \to \infty$ with $p/n \le \Gamma$.

Term 1: Because $|\Xi| \le n^{1-\nu} \le n$, and $\widehat{\sigma}_\Xi = o_p(\sqrt{n^\nu / \log(n)})$, following Remark 3.2.8, under the assumptions of Lemma 3.2.4 or Lemma 3.2.5, we have

$$\left| R(\widetilde{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widetilde{f}^{\xi}) \right| = o_p(1), \tag{C.41}$$

which proves that the first term on the right hand side of (C.40) converges to zero in probability.

Term 2: To deal with the second term on the right hand side of (C.40), define

$$\xi_n^\star \in \operatorname*{arg\,min}_{\xi \in \Xi} R^{\mathrm{det}} \left(\widetilde{f}; \frac{p_n}{n_\xi} \right).$$

Because $R^{\mathrm{det}}(\cdot; \cdot)$ is a non-stochastic function, $\{\xi_n^\star\}_{n \ge 1}$ is a non-stochastic sequence and further, trivially, $\xi_i^\star \in \Xi$ for all $n \ge 1$. Observe now that

$$\begin{aligned} \min_{\xi \in \Xi} R(\widetilde{f}^{\xi}) &\le R(\widetilde{f}^{\xi_n^\star}) \\ &= R(\widetilde{f}^{\xi_n^\star}) - R^{\mathrm{det}} \left(\widetilde{f}; \frac{p_n}{n_{\xi_n^\star}} \right) + \min_{\xi \in \Xi} R^{\mathrm{det}} \left(\widetilde{f}; \frac{p_n}{n_\xi} \right). \end{aligned} \tag{C.42}$$

Hence, assumption (DETPA-0) implies that

$$\left(\min_{\xi\in\Xi} R(\widetilde{f}^{\xi}) - \min_{\xi\in\Xi} R^{\mathrm{det}}\left(\widetilde{f};\, \frac{p_n}{n_\xi}\right)\right)_+ = o_p(1), \tag{C.43}$$

as $n \to \infty$.

Term 3: Finally, because the risk profile $\zeta \mapsto R^{\mathrm{det}}(\widetilde{f};\zeta)$ is assumed to be continuous at $\zeta^\star$, Lemma C.6.1 with the grid Ξ yields

$$\left|\min_{\xi\in\Xi} R^{\mathrm{det}}\left(\widetilde{f};\frac{p_n}{n_\xi}\right) - \inf_{\zeta\geq\gamma} R^{\mathrm{det}}(\widetilde{f};\zeta)\right| = o(1). \tag{C.44}$$

Combining (C.41), (C.43), and (C.44), we have the desired result that

$$\left|R(\widehat{f}^{\mathrm{cv}}) - \min_{\zeta\geq\gamma} R^{\mathrm{det}}(\widetilde{f};\zeta)\right| \xrightarrow{\mathrm{p}} 0.$$

Case of $M > 1$. Consider now the case when $M > 1$. Note that $(x+y)_+ \leq (x)_+ + (y)_+$ since $\max\{z, 0\}$ is a convex function of z. Thus, we can break and bound the desired difference as:

$$\begin{aligned}
&\left(R(\widehat{f}^{\mathrm{cv}}) - \min_{\zeta\geq p/n} R^{\mathrm{det}}(\widetilde{f};\zeta)\right)_+ \\
&\leq \left(R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi\in\Xi} R(\widehat{f}^{\xi})\right)_+ + \left(\min_{\xi\in\Xi} R(\widehat{f}^{\xi}) - \min_{\xi\in\Xi} \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi})\right)_+ \\
&\quad + \left(\min_{\xi\in\Xi} \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi}) - \min_{\xi\in\Xi} R^{\mathrm{det}}\left(\widetilde{f}^{\xi};\frac{p_n}{n_\xi}\right)\right)_+ \\
&\quad + \left(\min_{\xi\in\Xi} R^{\mathrm{det}}\left(\widetilde{f};\frac{p_n}{n_\xi}\right) - \min_{\zeta\geq\gamma} R^{\mathrm{det}}(\widetilde{f};\zeta)\right)_+ .
\end{aligned}$$

As before, we show below that each of these terms are asymptotically vanishing in probability.

Term 1: Note that $\widehat{\sigma}_\Xi \leq \widetilde{\sigma}_\Xi$ (from the triangle inequality for L_2 and ψ_1 norms). Thus, as argued above for the case of $m = 1$, the first term is $o_p(1)$.

Term 2: For the second term, observe that, for all $\xi \in \Xi$,

$$\begin{aligned}
R\left(\widehat{f}^{\xi}\right) = R\left(\frac{1}{M}\sum_{j=1}^{M}\widetilde{f}_j^{\xi}\right) &= \mathbb{E}\left[\ell\left(Y_0, \frac{1}{M}\sum_{i=1}^{M}\widetilde{f}_j^{\xi}(X_0)\right) \,\Big|\, \mathcal{D}_1\right] \\
&\leq \frac{1}{M}\sum_{j=1}^{M}\mathbb{E}\left[\ell(Y_0, \widetilde{f}_j^{\xi}(X_0)) \,\big|\, \mathcal{D}_1\right] \\
&\leq \frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi}).
\end{aligned}$$

Therefore, we have

$$\min_{\xi\in\Xi} R(\widehat{f}^{\xi}) \leq \min_{\xi\in\Xi}\frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi})$$

and the second term is 0.

Term 3: For the third term, as before, note that

$$\left(\min_{\xi\in\Xi}\frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi}) - \min_{\xi\in\Xi} R^{\mathrm{det}}\left(\widetilde{f}^{\xi};\frac{p}{n_\xi}\right)\right)_+ \leq \left(\frac{1}{M}\sum_{j=1}^{M} R(\widetilde{f}_j^{\xi_n^\star}) - R^{\mathrm{det}}\left(\widetilde{f};\frac{p_n}{n_{\xi_n^\star}}\right)\right)_+,$$

with the right hand side being $o_p(1)$ because of (DETPA-0).

Term 4: Analogous to the argument for the $m = 1$ case, the fourth term is $o(1)$.

Combined together, we have the final result. This completes the proof. For an overview, a schematic for the proof of Theorem 3.3.11 is provided in Figure C.2.

$$\begin{tikzcd}
R(\widehat{f}^{\mathrm{cv}}) \arrow[r] \arrow[ddrr] & \min\limits_{\xi \in \Xi_n} R(\widehat{f}^{\xi}) \arrow[r] & R(\widehat{f}^{\xi_n^\star}) \arrow[d] \\
 & & R^{\det}(p_n/n_{\xi_n^\star}; \widetilde{f}) \arrow[d] \\
 & & \min\limits_{\zeta \in [\gamma, \infty]} R^{\det(\zeta; \widetilde{f})}
\end{tikzcd}$$

Figure C.2: Schematic of the proof of Theorem 3.3.11.

C.3 Proofs related to deterministic profile verification for zero-step procedure

In this section, we verify the assumption (DETPAR-0) for the MN2LS and MN1LS prediction procedures.

C.3.1 Proof of Proposition 3.3.14

Recall $\mathcal{D}_{k_m}$ is a dataset with k_m observations and p_m features. Theorem 3 of Hastie et al. (2022) assumes the following distributional assumptions on the dataset $\mathcal{D}_{k_m}$.

(ℓ_2A1) The observations (X_i, Y_i), $1 \le i \le k_m$, are sampled i.i.d. from the model $Y_i = X_i^\top \beta_0 + \varepsilon_i$ for some (deterministic) unknown signal vector $\beta_0 \in \mathbb{R}^{p_m}$ and (random) unobserved error ε_i, assumed to be independent of $X_i \in \mathbb{R}^{p_m}$, with mean 0, variance σ^2, and bounded moment of order $4 + \delta$ for some $\delta > 0$.

(ℓ_2A2) The feature vector X_i, $1 \le i \le k_m$, decomposes as $X_i = \Sigma^{1/2} Z_i$, where $\Sigma \in \mathbb{R}^{p_m \times p_m}$ is a positive semidefinite (covariance) matrix and $Z_i \in \mathbb{R}^{p_m \times 1}$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4 + \delta$ for some $\delta > 0$.

(ℓ_2A3) The norm of the signal vector $\|\beta_0\|_2$ is uniformly bounded in p, and $\lim_{p_m \to \infty} \|\beta_0\|_2^2 = \rho^2 < \infty$.

(ℓ_2A4) There exist real numbers $r_{\min}$ and $r_{\max}$ with $0 < r_{\min} \le r_{\max} < \infty$ such that $r_{\min} I_{p_m} \preceq \Sigma \preceq r_{\max} I_{p_m}$.

(ℓ_2A5) Let $\Sigma = W R W^\top$ denote the eigenvalue decomposition of the covariance matrix Σ, where $R \in \mathbb{R}^{p_m \times p_m}$ is a diagonal matrix containing eigenvalues (in non-increasing order) $r_1 \ge r_2 \ge \cdots \ge r_{p_m} \ge 0$, and $W \in \mathbb{R}^{p_m \times p_m}$ is an orthonormal matrix containing the associated eigenvectors $w_1, w_2, \ldots, w_{p_m} \in \mathbb{R}^{p_m}$. Let H_{p_m} denote the empirical spectral distribution of Σ (supposed on $\mathbb{R}_{>0}$) whose value at any $r \in \mathbb{R}$ is given by

$$H_{p_m}(r) = \frac{1}{p_m} \sum_{i=1}^{p_m} \mathbb{I}_{\{r_i \le r\}}.$$

Let G_{p_m} denote a certain distribution (supported on $\mathbb{R}_{>0}$) that encodes the components of the signal vector β_0 in the eigenbasis of Σ via the distribution of (squared) projection of β_0 along the eigenvectors $w_j, 1 \leq j \leq p_m$, whose value any $r \in \mathbb{R}$ is given by

$$G_{p_m}(r) = \frac{1}{\|\beta_0\|_2^2} \sum_{i=1}^{p_m} (\beta_0^\top w_i)^2 \, \mathbb{I}_{\{r_i \leq r\}}.$$

Assume there exist fixed distributions H and G (supported on $\mathbb{R}_{>0}$) such that $H_{p_m} \xrightarrow{d} H$ and $G_{p_m} \xrightarrow{d} G$ as $p_m \to \infty$.

Under assumptions (ℓ_2A1)–(ℓ_2A5), we will verify that, for the MN2LS base prediction procedure $\widetilde{f}_{\mathrm{mn2}}$, there exists a deterministic risk approximation $R^{\mathrm{det}}(\cdot; \widetilde{f}_{\mathrm{mn2}}) : (0, \infty] \to [0, \infty]$ that satisfy the two conditions stated in Proposition 3.3.14. In particular, we will show that the function $R^{\mathrm{det}}(\cdot; \widetilde{f}_{\mathrm{mn2}})$ defined below satisfies the required conditions:

$$R^{\mathrm{det}}(\phi; \widetilde{f}_{\mathrm{mn2}}) = \begin{cases} \sigma^2 \dfrac{1}{1-\phi} & \text{if } \phi \in (0,1) \\ \infty & \text{if } \phi = 1 \\ \rho^2 (1 + \widetilde{v}_g(0;\phi)) \displaystyle\int \frac{r}{(1+v(0;\phi)r)^2} \, \mathrm{d}G(r) & \\ \quad + \sigma^2 \left(\phi \widetilde{v}(0;\phi) \displaystyle\int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}H(r) + 1 \right) & \text{if } \phi = (1, \infty) \\ \rho^2 \displaystyle\int r \, \mathrm{d}G(r) + \sigma^2 & \text{if } \phi = \infty, \end{cases} \tag{C.45}$$

where the scalars $v(0;\phi)$, $\widetilde{v}(0;\phi)$, and $\widetilde{v}_g(0;\phi)$, for $\phi \in (1, \infty)$, are defined as follows:

- $v(0;\phi)$ is the unique solution to the fixed-point equation:

$$\frac{1}{\phi} = \int \frac{v(0;\phi) r}{1 + v(0;\phi) r} \, \mathrm{d}H(r), \tag{C.46}$$

- $\widetilde{v}(0;\phi)$ is defined through $v(0;\phi)$ by the equation:

$$\widetilde{v}(0;\phi) = \left(\frac{1}{v(0;\phi)^2} - \phi \int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}H(r) \right)^{-1}, \tag{C.47}$$

- $\widetilde{v}_g(0;\phi)$ is defined through $v(0;\phi)$ and $\widetilde{v}(0;\phi)$ by the equation:

$$\widetilde{v}_g(0;\phi) = \widetilde{v}(0;\phi) \phi \int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}H(r). \tag{C.48}$$

We will verify the two conditions of Proposition 3.3.14 below.

The limiting risk for the MN2LS predictor provided in (C.45), although in a different notation, matches the one obtained in Theorem 3 of Hastie et al. (2022). We believe our notation makes the subsequent analysis for the one-step procedure easy to follow for the reader. It is worth mentioning, however, that Hastie et al. (2022) only explicitly consider $\phi \in (0,1) \cup (1, \infty)$. We extend the analysis to show that the risk continuously diverges to ∞ as $\phi \to 1$ and also continuously converges to the null risk as $\phi \to \infty$. In addition, as mentioned in Remark 3.3.16, we analyze the prediction risk conditioned on both $(\boldsymbol{X}, \boldsymbol{Y})$ as opposed to only on $\boldsymbol{X}$ as done in Hastie et al. (2022). Furthermore, we also establish continuity properties of the deterministic risk approximation in the aspect ratio that is needed for our analysis.

Condition 1: Continuous convergence of conditional risk over $\phi \in (0,1) \cup (1,\infty]$.

Let $\boldsymbol{X} \in \mathbb{R}^{k_m \times p_m}$ denote the design matrix and $\boldsymbol{Y} \in \mathbb{R}^{k_m}$ denote the response vector associated with the dataset $\mathcal{D}_{k_m}$. Let $\boldsymbol{\varepsilon} \in \mathbb{R}^{k_m}$ denote the error vector containing errors ε_i, $1 \le i \le k_m$. Write the data model from assumption (ℓ_2A1) as $\boldsymbol{Y} = \boldsymbol{X}^\top \beta_0 + \boldsymbol{\varepsilon}$, and the MN2LS estimator (3.20) as

$$\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) = (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{Y}/k_m. \tag{C.49}$$

The associated predictor $\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})$ is given by (3.22). Recall the prediction risk $R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m}))$ (where we use the subscripts $\boldsymbol{X}, \boldsymbol{Y}$ to explicitly indicate the dependence of $R(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m}))$ on the training data $(\boldsymbol{X}, \boldsymbol{Y})$) under the squared error loss is given by

$$R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) = \mathbb{E}[(Y_0 - \widetilde{f}_{\mathrm{mn2}}(X_0; \mathcal{D}_{k_m}))^2 \mid \boldsymbol{X}, \boldsymbol{Y}], \tag{C.50}$$

where (X_0, Y_0) is sampled independently from the same distribution as the training data $(\boldsymbol{X}, \boldsymbol{Y})$.

Our goal is to show that as $k_m, p_m \to \infty$, if $p_m/k_m \to \phi \in (0,1) \cup (1,\infty]$, $R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) \xrightarrow{\text{a.s.}} R^{\det}(\phi; \widetilde{f}_{\mathrm{mn2}})$. The proof follows by combining Propositions C.3.1 to C.3.3. Specifically:

1. Propositions C.3.1 and C.3.2 combined together imply that $R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) \xrightarrow{\text{a.s.}} R^{\det}(\phi; \widetilde{f}_{\mathrm{mn2}})$ as $p_m, k_m \to \infty$ and $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$.

2. Proposition C.3.3 imply that $R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) \xrightarrow{\text{a.s.}} R^{\det}(\infty; \widetilde{f}_{\mathrm{mn2}})$ as $p_m, k_m \to \infty$ and $p_m/k_m \to \infty$.

Below we prove Propositions C.3.1 to C.3.3.

In preparation for the statements to follow, denote by $\widehat{\boldsymbol{\Sigma}} := \boldsymbol{X}^\top \boldsymbol{X}/k_m$ the sample covariance matrix. Let the singular value decomposition of $\boldsymbol{X}/\sqrt{k_m}$ be $\boldsymbol{X}/\sqrt{k_m} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{k_m \times k_m}$ and $\boldsymbol{V} \in \mathbb{R}^{p_m \times p_m}$ are orthonormal matrices, and $\boldsymbol{S} \in \mathbb{R}^{k_m \times p}$ is a diagonal matrix containing singular values in non-increasing order $s_1 \ge s_2 \ge \dots$.

The proposition below provides conditional convergence for the prediction risk (C.50) when $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$ as $p_m, k_m \to \infty$.

Proposition C.3.1 (Conditional convergence of squared prediction risk of MN2LS predictor)**.** *Suppose assumptions (ℓ_2A1)–(ℓ_2A4) hold. Then, as $k_m, p_m \to \infty$, if $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$, then*

$$R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) - \beta_0^\top (I_{p_m} - \widehat{\boldsymbol{\Sigma}}^\dagger \widehat{\boldsymbol{\Sigma}}) \Sigma (I_{p_m} - \widehat{\boldsymbol{\Sigma}}^\dagger \widehat{\boldsymbol{\Sigma}}) \beta_0 - \sigma^2 \operatorname{tr}[\widehat{\boldsymbol{\Sigma}}^\dagger \Sigma]/k_m - \sigma^2 \xrightarrow{\text{a.s.}} 0. \tag{C.51}$$

Proof. Under assumption (ℓ_2A1), the squared prediction risk (C.50) decomposes into

$$R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) = (\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) - \beta_0)^\top \Sigma (\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) - \beta_0) + \sigma^2. \tag{C.52}$$

Similarly, under assumption (ℓ_2A1), the estimator (C.49) decomposes into

$$\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) = (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{X}/k_m \, \beta_0 + (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{\varepsilon}/k_m.$$

Consequently, the difference between the estimator and the true parameter decomposes as

$$\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) - \beta_0 = \big\{(\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{X}/k_m - I_{p_m}\big\}\beta_0 + (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{\varepsilon}/k_m. \tag{C.53}$$

Substituting (C.53) into (C.52), we can split the first term on the right hand side of (C.52) into three component terms:

$$(\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) - \beta_0)^\top \Sigma (\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m}) - \beta_0) = \boldsymbol{B}_0 + \boldsymbol{V}_0 + \boldsymbol{C}_0,$$

where the component terms are given by:

$$\boldsymbol{B}_0 = \beta_0^\top \big\{(\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{X}/k_m - I_{p_m}\big\} \Sigma \big\{(\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{X}/k_m - I_{p_m}\big\}\beta_0$$

$$
\begin{aligned}
&= \beta_0^\top (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \Sigma (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \beta_0, \\
\boldsymbol{C}_0 &= \beta_0^\top \{ (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{X}/k_m - I_{p_m} \} \Sigma (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{\varepsilon}/k_m \\
&= -\beta_0^\top (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \Sigma \widehat{\mathbf{\Sigma}}^\dagger \boldsymbol{X}^\top \boldsymbol{\varepsilon}/k_m, \\
\boldsymbol{V}_0 &= \boldsymbol{\varepsilon}^\top \boldsymbol{X}/k_m (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \Sigma (\boldsymbol{X}^\top \boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top \boldsymbol{\varepsilon}/k_m \\
&= \boldsymbol{\varepsilon}^\top (\boldsymbol{X} \widehat{\mathbf{\Sigma}}^\dagger \Sigma \widehat{\mathbf{\Sigma}}^\dagger \boldsymbol{X}^\top / k_m) \boldsymbol{\varepsilon}/k_m.
\end{aligned}
$$

To finish the proof, we will show concentration of the terms $\boldsymbol{C}_0$ and $\boldsymbol{V}_0$ below.

Term $\boldsymbol{C}_0$: We will show that $\boldsymbol{C}_0 \xrightarrow{\text{a.s.}} 0$ as $k_m, p_m \to \infty$ such that $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$. Note that

$$
\begin{aligned}
\| \boldsymbol{X} \widehat{\mathbf{\Sigma}}^\dagger \Sigma (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \beta_0 \|_2^2 / k_m &= \beta_0^\top (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \Sigma \widehat{\mathbf{\Sigma}}^\dagger \boldsymbol{X}^\top \boldsymbol{X} \widehat{\mathbf{\Sigma}}^\dagger \Sigma (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \beta_0 / k_m \\
&\le \|\beta_0\|_2^2 \| \| (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \Sigma \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Sigma}}^\dagger \Sigma (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \|_{\text{op}} \\
&\le \|\beta_0\|_2^2 \| \cdot r_{\max}^2 \cdot \| \widehat{\mathbf{\Sigma}}^\dagger \|_{\text{op}}, \qquad \text{(C.54)}
\end{aligned}
$$

where in the last inequality (C.54), we used the fact that $\| I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}} \|_{\text{op}} \le 1$, $\|\Sigma\|_{\text{op}} \le r_{\max}$, and that $\widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Sigma}}^\dagger = \widehat{\mathbf{\Sigma}}^\dagger$, along with the submultiplicativity of the operator norm. Now, note that $\liminf \min_{1 \le i \le p} s_i^2 \ge r_{\min}(1 - \sqrt{\phi})^2$ almost surely from Bai and Silverstein (2010) for $\phi \in (0,1) \cup (1,\infty)$. Therefore, $\limsup \| \widehat{\mathbf{\Sigma}}^\dagger \|_{\text{op}} \le C$ for some constant $C < \infty$ almost surely. Applying Lemma C.8.5, we thus have that $\boldsymbol{C}_0 \xrightarrow{\text{a.s.}} 0$.

Term $\boldsymbol{V}_0$: We will show that $\boldsymbol{V}_0 - \operatorname{tr}[\widehat{\mathbf{\Sigma}}^+ \Sigma]/k_m \xrightarrow{\text{a.s.}} 0$ as $k_m, p_m \to \infty$ such that $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$. Observe that

$$
\| \boldsymbol{X} \widehat{\mathbf{\Sigma}}^\dagger \Sigma \widehat{\mathbf{\Sigma}}^\dagger \boldsymbol{X}^\top / k_m \|_{\text{op}} \le r_{\max} \| \widehat{\mathbf{\Sigma}} \|_{\text{op}} \| \widehat{\mathbf{\Sigma}}^\dagger \|_{\text{op}}^2. \qquad \text{(C.55)}
$$

Now, note that $\limsup \| \widehat{\mathbf{\Sigma}} \|_{\text{op}} \le \limsup \max_{1 \le i \le p} s_i^2 \le r_{\max}(1 + \sqrt{\phi})^2$, almost surely for $\phi \in (0,1) \cup (1,\infty)$ from Bai and Silverstein (2010). In addition, as argued above, $\| \widehat{\mathbf{\Sigma}}^\dagger \|_{\text{op}} \le C$ almost surely for some constant $C < \infty$. Thus, using Lemma C.8.6, it follows that $\boldsymbol{V}_0 - \sigma^2 \operatorname{tr}[\boldsymbol{X} \widehat{\mathbf{\Sigma}}^+ \Sigma \widehat{\mathbf{\Sigma}}^+ \boldsymbol{X}^\top]/k_m^2 \xrightarrow{\text{a.s.}} 0$. Finally, since $\operatorname{tr}[\boldsymbol{X} \widehat{\mathbf{\Sigma}}^+ \Sigma \widehat{\mathbf{\Sigma}}^+ \boldsymbol{X}^\top]/k_m^2 = \operatorname{tr}[\widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}} \widehat{\mathbf{\Sigma}}^\dagger \Sigma]/k_m = \operatorname{tr}[\widehat{\mathbf{\Sigma}}^\dagger \Sigma]/k_m$, we obtain that $\boldsymbol{V}_0 - \sigma^2 \operatorname{tr}[\widehat{\mathbf{\Sigma}}^\dagger \Sigma]/k_m \xrightarrow{\text{a.s.}} 0$. □

The next proposition provides deterministic limits of the conditional risk functionals in Proposition C.3.1 when $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$ as $k_m, p_m \to \infty$.

Proposition C.3.2 (Limits of conditional risk functionals over $\phi \in (0,1) \cup (1,\infty)$)**.** *Suppose assumptions (ℓ_2A2)–(ℓ_2A5) hold. Then, as $k_m, p_m \to \infty$, and $p_m/k_m \to \phi \in (0,1) \cup (1,\infty)$, the following holds:*

- *Bias functional:*

$$
\beta_0^\top (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \Sigma (I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger \widehat{\mathbf{\Sigma}}) \beta_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \textit{if } \phi \in (0,1) \\ \rho^2 (1 + \widetilde{v}_g(0;\phi)) \displaystyle\int \frac{r}{(1 + v(0;\phi) r)^2} \, \mathrm{d}G(r) & \textit{if } \phi \in (1,\infty), \end{cases}
$$

- *Variance functional:*

$$
\sigma^2 \operatorname{tr}[\widehat{\mathbf{\Sigma}}^\dagger \Sigma]/k_m \xrightarrow{\text{a.s.}} \begin{cases} \sigma^2 \dfrac{\phi}{1 - \phi} & \textit{if } \phi \in (0,1) \\ \sigma^2 \phi \widetilde{v}(0;\phi) \displaystyle\int \frac{r^2}{(1 + v(0;\phi) r)^2} \, \mathrm{d}H(r) & \textit{if } \phi \in (1,\infty), \end{cases}
$$

where $v(0;\phi)$, $\widetilde{v}(0;\phi)$, and $\widetilde{v}_g(0;\phi)$ are as defined in (C.46), (C.47), *and* (C.48), *respectively.*

Proof. We will consider the bias and functionals separately below.

Bias functional. Consider first the bias functional $\beta_0^\top(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\Sigma(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\beta_0$. Since $r_{\min} > 0$, the smallest eigenvalue of $\widehat{\mathbf{\Sigma}}^\dagger$ is almost surely positive, and the matrix $\widehat{\mathbf{\Sigma}}$ is almost surely invertible as $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in (0,1)$. Therefore, in this case, $\widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}} = I_{p_m}$ almost surely, and $\beta_0^\top(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\Sigma(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\beta_0 \xrightarrow{\text{a.s.}} 0$. For the case when $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in (1,\infty)$, from the second part of Corollary C.6.12 by taking $f(\Sigma) = \Sigma$, we have

$$(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\Sigma(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}}) \simeq (1 + \widetilde{v}_g(0;\phi))(v(0;\phi)\Sigma + I_{p_m})^{-1}\Sigma(v(0;\phi)\Sigma + I_{p_m})^{-1},$$

where $v(0;\phi)$ and $\widetilde{v}_g(0)$ are as defined by (C.46) and (C.48), respectively. Note that from Lemma C.6.13 (1) $v(0;\phi)$ is bounded for $\phi \in (1,\infty)$, and the function $r \mapsto r/(1 + rv(0;\phi))^2$ is continuous. Hence, under (ℓ_2A3) and (ℓ_2A5), using Lemma C.7.2 (4), we have

$$\begin{aligned}
\beta_0^\top(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\Sigma(I_{p_m} - \widehat{\mathbf{\Sigma}}^\dagger\widehat{\mathbf{\Sigma}})\beta_0 &\xrightarrow{\text{a.s.}} \lim_{p_m\to\infty}\sum_{i=1}^{p_m}(1 + \widetilde{v}_g(0;\phi))\frac{r_i}{(1 + r_i v(0;\phi))^2}(\beta_0^\top w_i)^2 \\
&= \lim_{p_m\to\infty}\|\beta_0\|_2^2(1 + \widetilde{v}_g(0;\phi))\int\frac{r}{(1 + rv(0;\phi))^2}\,\mathrm{d}G_{p_m}(r) \\
&= \rho^2(1 + \widetilde{v}_g(0;\phi))\int\frac{r}{(1 + rv(0;\phi))^2}\,\mathrm{d}G(r),
\end{aligned}$$

where in the last line we used the fact that G_{p_m} and G have compact supports, and $\lim_{p_m\to\infty}\|\beta_0\|_2^2 = \rho^2$. This completes the proof of the first part.

Variance functional. Consider next the variance functional $\mathrm{tr}[\widehat{\mathbf{\Sigma}}^\dagger\Sigma]/k_m$. As $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in (0,1)$, $\widehat{\mathbf{\Sigma}}$ is almost surely invertible as explained above. In this case, $\mathrm{tr}[\widehat{\mathbf{\Sigma}}^\dagger\Sigma]/k_m - \mathrm{tr}[(\boldsymbol{Z}^\top\boldsymbol{Z}/k_m)^{-1}]/k_m \xrightarrow{\text{a.s.}} 0$, where $\boldsymbol{Z} \in \mathbb{R}^{k_m\times p_m}$ is matrix with rows Z_i, $1 \le i \le k_m$. From the proof of Proposition 2 of Hastie et al. (2022), this limit is given by $\phi/(1-\phi)$. In the case when $k_m, p_m \to \infty$ and $p_m/k_m \to \phi \in (1,\infty)$, from Corollary C.6.12, we have

$$\widehat{\mathbf{\Sigma}}^\dagger\Sigma \simeq \widetilde{v}(0;\phi)(v(0;\phi)\Sigma + I_p)^{-2}\Sigma^2.$$

Along the same lines as above, from Lemma C.6.13 (1), $v(0;\phi)$ is bounded for $\phi \in (1,\infty)$, and the the function $r \mapsto r^2/(1 + v(0;\phi)r)^2$ is continuous. Thus, under (ℓ_2A5), using Lemma C.7.2 (4), we have

$$\begin{aligned}
\sigma^2\,\mathrm{tr}[\widehat{\mathbf{\Sigma}}^\dagger\Sigma]/k_m &\xrightarrow{\text{a.s.}} \lim_{p_m\to\infty}\frac{p_m}{k_m}\frac{1}{p_m}\widetilde{v}(0;\phi)\sum_{i=1}^{p_m}\frac{r_i^2}{(1 + v(0;\phi)r_i)^2} \\
&= \lim_{p_m\to\infty}\frac{p_m}{k_m}\widetilde{v}(0;\phi)\int\frac{r^2}{(1 + v(0;\phi)r)^2}\,\mathrm{d}H(r) \\
&= \phi\widetilde{v}(0;\phi)\int\frac{r^2}{(1 + v(0;\phi)r)^2}\,\mathrm{d}H(r).
\end{aligned}$$

This completes the proof of the second part. □

We remark that Corollary C.6.12 used in the proof of Proposition C.3.2 assumes existence of moments of order $8 + \alpha$ for some $\alpha > 0$ on the entries of Z_i, $1 \le i \le k_m$, mentioned in assumption (ℓ_2A1). As done in the proof of Theorem 6 of Hastie et al. (2022) (in Appendix A.1.4 therein), this can be relaxed to only requiring existence of moments of order $4 + \alpha$. This being a simple truncation argument, we omit the details and refer the readers to Hastie et al. (2022).

The proposition below covers the case when $p_m/k_m \to \infty$ as $p_m, k_m \to \infty$.

Proposition C.3.3 (Limits of risk and deterministic risk approximation as $\phi \to \infty$)**.** *Suppose assumptions (ℓ_2A1)–(ℓ_2A5) hold. Then, as $k_m, p_m \to \infty$ and $p_m/k_m \to \infty$, we have*

$$R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot;\mathcal{D}_{k_m})) - \beta_0^\top\Sigma\beta_0 - \sigma^2 \xrightarrow{\text{a.s.}} 0.$$

In addition,

$$\lim_{\phi\to\infty} R^{\mathrm{det}}(\cdot; \widetilde{f}_{\mathrm{mn2}}) = \lim_{p_m\to\infty} \beta_0\Sigma\beta_0 + \sigma^2 = \rho^2 \int r \, \mathrm{d}G(r) + \sigma^2.$$

Proof. From (C.52), note that

$$\begin{aligned} R_{\boldsymbol{X},\boldsymbol{Y}}(\widetilde{f}_{\mathrm{mn2}}(\cdot; \mathcal{D}_{k_m})) - (\|\beta_0\|_\Sigma^2 + \sigma^2) &= \|\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})\|_\Sigma^2 - 2\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})^\top \Sigma\beta_0 \\ &\le r_{\min}^{-1}\|\widetilde{\beta}_{\mathrm{mn2}}\|_2^2 + 2\|\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})\|_2\|\Sigma\beta_0\|_2 \\ &\le r_{\min}^{-1}\|\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})\|_2^2 + 2r_{\max} r\|\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})\|_2, \end{aligned}$$

where the first inequality follows by using the lower bound $r_{\min}$ on the smallest eigenvalue of Σ, and the Cauchy-Schwarz inequality, and the second inequality follows by using the upper bound $r_{\max}$ on the largest eigenvalue of Σ. Thus, for the first part it suffices to show that $\|\widetilde{\beta}_{\mathrm{mn2}}\|_2 \to 0$ as $k_m, p \to 0$ and $p/k_m \to \infty$. Towards that end, note that

$$\begin{aligned} \|\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_m})\|_2 &= \|(\boldsymbol{X}^\top\boldsymbol{X}/k_m)^\dagger \boldsymbol{X}^\top\boldsymbol{Y}/k_m\|_2 \\ &\le \|(\boldsymbol{X}^\top\boldsymbol{X}/k_m)^\dagger \boldsymbol{X}/\sqrt{k_m}\|_{\mathrm{op}}\|\boldsymbol{Y}/\sqrt{k_m}\|_2 \\ &\le C\|(\boldsymbol{X}^\top\boldsymbol{X}/k_m)^\dagger \boldsymbol{X}/\sqrt{k_m}\|_{\mathrm{op}}\sqrt{\rho^2+\sigma^2}, \end{aligned}$$

where the last inequality holds eventually almost surely since (ℓ_2A1) and (ℓ_2A3) imply that the entries of $\boldsymbol{Y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{Y}/\sqrt{k_m}\|_2$ is eventually almost surely bounded above by $\sqrt{\mathbb{E}[Y^2]} = \sqrt{\rho^2+\sigma^2}$. Observe that operator norm of the matrix $(\boldsymbol{X}^\top\boldsymbol{X}/k_m)^\dagger\boldsymbol{X}/\sqrt{k_m}$ is upper bounded by the inverse of the smallest non-zero singular value $s_{\min}$ of $\boldsymbol{X}$. As $k_m, p_m \to \infty$ such that $p_m/k_m \to \infty$, $s_{\min} \to \infty$ almost surely (e.g., from results in Bloemendal et al. (2016)) and therefore, $\|\beta\|_2 \to 0$ almost surely. This completes the proof of first part.

Now, from Lemma C.6.13 (1) $\lim_{\phi\to\infty} v(0;\phi) = 0$, and from Lemma C.6.13 (4) $\lim_{\phi\to\infty} \widetilde{v}_g(0;\phi) = 0$. Thus,

$$\lim_{\phi\to\infty} \rho^2(1+\widetilde{v}_g(0;\phi)) \int \frac{r}{(1+v(0;\phi)r)^2)} \, \mathrm{d}G(r) = \rho^2 \int r \, \mathrm{d}G(r).$$

On the other hand, from Lemma C.6.13 (4),

$$\lim_{\phi\to\infty} \sigma^2\phi\widetilde{v}(0;\phi) \int \frac{r}{(1+v(0;\phi)r)^2} \, \mathrm{d}H(r) = 0.$$

This proves the second part, and finishes the proof. □

<u>**Condition 2**</u>**: Left and right limits of deterministic risk approximation as $\phi \to 1$.**

Next we verify that $\lim_{\phi\to1} R^{\mathrm{det}}(\phi; \widetilde{f}_{\mathrm{mn2}}) = \infty$. First note that $\lim_{\phi\to1^-} R^{\mathrm{det}}(\phi; \widetilde{f}_{\mathrm{mn2}}) = \lim_{\phi\to1^-} 1/(1-\phi) = \infty$. Now, from Lemma C.6.13 (4), observe that

$$\lim_{\phi\to1^+} \phi\widetilde{v}(0;\phi) \int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}H(r) = \infty.$$

Since $\lim_{\phi\to1^-} R^{\mathrm{det}}(\phi) = \lim_{\phi\to1^+} R^{\mathrm{det}}(\phi) = \infty$, we have that $\lim_{\phi\to1} R^{\mathrm{det}}(\phi) = \infty$, as claimed. This finishes the verification.

C.3.2 Proof of Proposition 3.3.15

Recall that $\mathcal{D}_{k_m}$ is a dataset with k_m observations and p_m features. Li and Wei (2021) makes the following distributional assumptions on the dataset $\mathcal{D}_{k_m}$. We adapt the scalings of Li and Wei (2021) to match the current work for easy comparisons.

(ℓ_1A1) (X_i, Y_i) for $1 \le i \le k_m$ are i.i.d. observations from the model: $Y = X^\top \beta_0 + \varepsilon$ for some fixed unknown vector $\beta_0 \in \mathbb{R}^{p_m \times 1}$ and unobserved error ε where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ independent of X.

(ℓ_1A2) Each design vector is independently drawn by $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p)$.

(ℓ_1A3) The signal vector β_0 is random such that the scaled coordinates $\{\sqrt{p_m} \cdot \beta_0^i\}_{i=1}^{p_m}$ converge weakly to a probability measure P_Θ, where $\mathbb{E}[\Theta^2] < \infty$ and $\mathbb{P}(\Theta \neq 0) > 0$.

Under these assumptions, Theorem 2 of Li and Wei (2021) demonstrates that the prediction risk of the MN1LS estimator obeys [1]

$$\lim_{\substack{p/n \to \phi \\ n,\, p \to \infty}} R(\widetilde{f}_{\text{mn1}}(\cdot; \mathcal{D}_{k_m})) \;=\; \tau^{\star 2}, \tag{C.56}$$

almost surely with respect to X and Y. Here, $(\tau^\star, \alpha^\star)$ stands for the unique solution to the following system of equations

$$\tau^2 = \sigma^2 + \mathbb{E}\left[\left(\eta(\Theta + \tau Z; \alpha\tau) - \Theta\right)^2\right], \tag{C.57a}$$

$$\phi^{-1} = \mathbb{P}\big(|\Theta + \tau Z| > \alpha\tau\big), \tag{C.57b}$$

where $\Theta \sim P_\Theta$, and $Z \sim \mathcal{N}(0,1)$ and is independent of Θ. Here, $\eta(\cdot; b)$ is the soft-thresholding function at level $b \ge 0$ that maps $x \in \mathbb{R}$ to

$$\eta(x; b) = (|x| - b)_+ \operatorname{sgn}(x).$$

The existence and uniqueness of the equation set (C.57) is established in Li and Wei (2021). To facilitate accurate characterization of $\tau^\star$ as a function of ϕ, we make assumption on how the ground true is generated as follows.

(ℓ_1A4) Suppose that each coordinate of $\beta_0 = [\beta_0^i]_{1 \le i \le p}$ is identically and independently drawn as follows

$$\beta_0^i \overset{\text{i.i.d.}}{\sim} \epsilon \mathcal{P}_{M/\sqrt{p_m}} + (1 - \epsilon)\mathcal{P}_0, \tag{C.58}$$

where $\mathcal{P}_c$ corresponds to the Dirac measure at point $c \in \mathbb{R}$, and $M > 0$ is some given scalar that determines the magnitude of a non-zero entry.

Under the above four assumptions, it is proved in Lemma 2 (p. 50) of Li and Wei (2021) that

$$\lim_{\phi \to 1^+} \tau^{\star 2}(\phi) = \infty, \tag{C.59}$$

and Lemma 1 (p. 51) of Li and Wei (2021) that

$$\lim_{\phi \to \infty} \tau^{\star 2}(\phi) = \sigma^2 + \mathbb{E}\|\beta_0\|_2^2 = \sigma^2 + \epsilon M^2.$$

We remark that the above results are stated slight differently therein due to a different scaling, where a global $1/\sqrt{k_m}$ is applied to the design matrix and $\sqrt{p_m}$ is applied to the ground truth parameter β_0. Here, we adapt a global scaling to allow for convenient comparisons with the MN2LS estimator.

From the discussion above, it is therefore clear that, one can set

$$R^{\text{det}}(\cdot; \widetilde{f}_{\text{mn1}}) = \begin{cases} \sigma^2 \dfrac{1}{1-\phi} & \text{if } \phi \in (0,1) \\ \infty & \text{if } \phi = 1 \\ \tau^{\star 2} & \text{if } \phi \in (1, \infty) \\ \sigma^2 + \epsilon M^2 & \text{if } \phi = \infty \end{cases} \tag{C.60}$$

[1] Li and Wei (2021) assumes $p/n = \phi$ for simplicity, but the proof goes through literatim as $p/n \to \phi$.

which satisfies the conditions of Proposition 3.3.15.

In order to see this, first recognizing that the convergence (C.56) holds almost surely, the first condition of Proposition 3.3.15 is satisfied naturally. Additionally, as established in Section C.3.1 and in (C.59), one has

$$\lim_{\phi \to 1^+} R^{\mathrm{det}}(\phi; \widetilde{f}_{\mathrm{mn1}}) = \infty, \quad \text{and} \quad \lim_{\phi \to 1^-} R^{\mathrm{det}}(\phi; \widetilde{f}_{\mathrm{mn1}}) = \infty, \tag{C.61}$$

which validates the second condition of Proposition 3.3.15. Putting everything together completes the proof of Proposition 3.3.15.

C.4 Proofs related to risk monotonization for one-step procedure

C.4.1 Proof of Lemma 3.4.1

The idea of the proof is similar to proof of Lemma 3.3.8. We wish to verify that there exists a deterministic approximation $R^{\mathrm{det}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ to the conditional prediction risk of the predictor $\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n},j}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n},j})$, $1 \le j \le M$ that satisfy

$$\left| R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n}^\star,j}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n}^\star,j})) - R^{\mathrm{det}}\left(\frac{p_n}{n_{1,\xi_{1,n}^\star}}, \frac{p_n}{n_{2,\xi_{2,n}^\star}}; \widetilde{f}\right) \right| = o_p(1) R^{\mathrm{det}}\left(\frac{p_n}{n_{1,\xi_{1,n}^\star}}, \frac{p_n}{n_{2,\xi_{2,n}^\star}}; \widetilde{f}\right)$$

as $n \to \infty$ under (PA(γ)), where $(\xi_{1,n}^\star, \xi_{2,n}^\star)$ are indices such that

$$(\xi_{1,n}^\star, \xi_{2,n}^\star) \in \operatorname*{arg\,min}_{(\xi_1,\xi_2)\in\Xi_n} R^{\mathrm{det}}\left(\frac{p_n}{n_{1,\xi_1}}, \frac{p_n}{n_{2,\xi_2}}; \widetilde{f}\right).$$

Following the arguments in the proof of Lemma 3.3.8, using the lower bound on $R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n},j}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n},j}))$ and identical distribution across j, it suffices to show that for all $\epsilon > 0$,

$$\mathbb{P}\left(\left| R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n}^\star}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n}^\star})) - R^{\mathrm{det}}\left(\frac{p_n}{n_{1,\xi_{1,n}^\star}}, \frac{p_n}{n_{2,\xi_{2,n}^\star}}; \widetilde{f}\right) \right| \ge \epsilon \right) \to 0$$

as $n \to \infty$ under (PA(γ)). Note that here we have dropped the superscript j for brevity. Now we will show that (DETPAR-1) along with the assumed continuity behavior of $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$ implies desired conclusion. Fix $\varepsilon > 0$ and define a sequence $h_n(\epsilon)$ as follows:

$$h_n(\epsilon) := \mathbb{P}\left(\left| R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n}^\star}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n}^\star})) - R^{\mathrm{det}}\left(\frac{p_n}{n_{1,\xi_{1,n}^\star}}, \frac{p_n}{n_{2,\xi_{2,n}^\star}}; \widetilde{f}\right) \right| \ge \epsilon \right).$$

We want to show that $h_n(\epsilon) \to \infty$ as $n \to \infty$ under (PA(γ)). We first note that using Lemma C.6.3, it suffices to show that for an arbitrary subsequence $\{n_k\}_{k\ge1}$, there exists further subsequence $\{n_{k_l}\}_{l\ge1}$ such that $h_{n_{k_l}} \to 0$ as $n \to \infty$. Also, note that since $n_{\mathrm{tr}}/n \to 1$, the grid Ξ_n satisfies the space-filling property from Lemma C.6.2 that $\Pi_{\Xi_n}(\zeta_1, \zeta_2) \to (\zeta_1, \zeta_2)$ for any (ζ_1, ζ_2) that satisfy $\zeta_1^{-1} + \zeta_2^{-1} \le \gamma^{-1}$ and the set of (ζ_1, ζ_2) that satisfy this condition is compact. Now, we apply Lemma C.6.5 on the function $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$ and the grid Ξ_n. Let sequence $\{x_n\}_{n\ge1}$ be such that $x_n := (p_n/n_{1,\xi_{1,n}^\star}, p_n/n_{2,\xi_{2,n}^\star})$ for $n \ge 1$. Lemma C.6.5 guarantees that for any arbitrary subsequence $\{x_{n_k}\}_{k\ge1}$, there exists a further subsequence $\{x_{n_{k_l}}\}_{l\ge1}$ such that

$$x_{n_{k_l}} \to (\phi_1, \phi_2) \in \operatorname*{arg\,min}_{\zeta_1^{-1}+\zeta_2^{-1}\le\gamma^{-1}} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}). \tag{C.62}$$

We will now show that $h_{n_{k_l}} \to 0$ as $l \to \infty$ if assumption (DETPAR-1) Lemma 3.4.1 is satisfied. It is easy to see that the assumption implies

$$R(\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_{1,n}^\star}, \mathcal{D}_{\mathrm{tr}}^{\xi_{2,n}^\star})) \xrightarrow{\mathrm{p}} R^{\mathrm{det}}(\phi_1, \phi_2; \widetilde{f})$$

as $n, p_n, \xi^\star_{1,n}, \xi^\star_{2,n} \to \infty$, whenever

$$(p_n/n_{1,\xi^\star_{1,n}}, p_n/n_{2,\xi^\star_{2,n}}) \to (\phi_1, \phi_2) \in \operatorname*{arg\,min}_{\zeta_1^{-1}+\zeta_2^{-1}\le\gamma^{-1}} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}).$$

But using the continuity of $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$ on the set $\operatorname{arg\,min}_{\zeta_1^{-1}+\zeta_2^{-1}\le\gamma^{-1}} R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f})$ and the fact that the sequence $\{x_{n_{k_l}}\}_{l\ge1}$ converges to a point in this minimizing set from (C.62), it follows that that $h_{n_{k_l}} \to 0$ as $l \to \infty$ as desired. This finishes the proof.

C.4.2 Proof of Proposition 3.4.2

Fix $t < \infty$. We will verify that the set $C_t := \{x : h(x) \le t\}$ is closed. Note that $C_t \subseteq M \setminus C$ because $h(x) < \infty$ for $x \in C_t$. Now consider any converging sequence $\{x_n\}_{n\ge1}$ in C_t with limit point p. We will argue that $p \in C_t$. First note that the function h is continuous over C_t because $C_t \subseteq M \setminus C$. Note that $p \notin C$, because if it does then $h(x_n) \to \infty$ as $n \to \infty$, which in turn implies that for infinitely many $k \ge 1$, $h(x_k) > t$, contradicting $x_n \in C_t$ for all $n \ge 1$. Hence, $p \in M \setminus C$ and $x_n \in M \setminus C$ for all $n \ge 1$. Therefore, continuity of h on $M \setminus C$ yields $h(x_n) \to h(p)$. Moreover, $h(x_n) \le t$ implies that $\lim_{n\to\infty} h(x_n) \le t$, which in turn implies that $h(p) \le t$. Hence $p \in C$, finishing the proof.

C.4.3 Proof of Proposition 3.4.3

The proof uses a similar contradiction strategy employed in the proof of Proposition 3.3.10. We only sketch the proof, and omit the details.

Suppose $R^{\mathrm{det}}(\cdot, \cdot; \widetilde{f})$ is discontinuous at some point $(\phi_{1,\infty}, \phi_{2,\infty})$. This gives us a sequence $\{(\phi_{1,r}, \phi_{2,r})\}_{r\ge1}$ such that for some $\epsilon > 0$ and all $r \ge 1$,

$$R^{\mathrm{det}}(\phi_{1,r}, \phi_{2,r}; \widetilde{f}) \notin [R^{\mathrm{det}}(\phi_{1,\infty}, \phi_{2,\infty}; \widetilde{f}) - 2\epsilon, R^{\mathrm{det}}(\phi_{1,\infty}, \phi_{2,\infty}; \widetilde{f}) + 2\epsilon], \tag{C.63}$$

while $(\phi_{1,r}, \phi_{2,r}) \to (\phi_{1,\infty}, \phi_{2,\infty})$ as $r \to \infty$. From the continuous convergence hypothesis, for each $r \ge 1$, one can then construct a sequence of datasets $\{(\mathcal{D}^{\phi_{1,r}}_{k_{1,m}}, \mathcal{D}^{\phi_{2,r}}_{k_{2,m}})\}_{m\ge1}$ with p_m features and $(k_{1,m}, k_{2,m})$ observations for which

$$R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_{1,r}}_{k_{1,m}}, \mathcal{D}^{\phi_{2,r}}_{k_{2,m}})) \xrightarrow{\text{p}} R^{\mathrm{det}}(\phi_{1,r}, \phi_{2,r}; \widetilde{f}) \tag{C.64}$$

as $p_m, k_{1,m}, k_{2,m} \to \infty$ and $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_{1,r}, \phi_{2,r})$. From (C.63) and (C.64), one can obtain a sequence of increasing integers $\{m_r\}_{r\ge1}$ such that for each $r \ge 1$, with probability $0 < p < 1$,

$$|R(\widetilde{f}(\cdot; \mathcal{D}^{\phi_{1,r}}_{k_{1,m}}, \mathcal{D}^{\phi_{2,r}}_{k_{2,m}})) - R^{\mathrm{det}}(\phi_{1,\infty}, \phi_{2,\infty}; \widetilde{f})| > \epsilon.$$

This then lets us construct a sequence of datasets $\{(\mathcal{D}'_{k_{1,m}}, \mathcal{D}'_{k_{2,m}})\}_{m\ge1}$ similar as done in the proof of Proposition 3.3.10 for which

$$R(\widetilde{f}(\cdot; \mathcal{D}'_{k_{1,m}}, \mathcal{D}'_{k_{2,m}})) \not\xrightarrow{\text{p}} R^{\mathrm{det}}(\phi_{1,\infty}, \phi_{2,\infty}; \widetilde{f})$$

as $p_m, k_{1,m}, k_{2,m} \to \infty$ and $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_{1,\infty}, \phi_{2,\infty})$. This supplies the required contradiction to the continuous convergence hypothesis.

C.4.4 Proof of Theorem 3.4.4

The idea of the proof is similar to that of the proof of Theorem 3.3.11. We will break the proof in two cases.

Case of $M = 1$. Consider first the case when $m = 1$. In this case, $\widehat{f}^{\mathrm{cv}} = \widetilde{f}_1^{\xi}$, which we denote by $\widetilde{f}^{\xi}$ for notational simplicity. Bound the desired difference as

$$\begin{aligned}
&\left| R(\widehat{f}^{\mathrm{cv}}) - \min_{1/\zeta_1 + 1/\zeta_2 \le n/p} R^{\det}(\widetilde{f}; \zeta_1, \zeta_2) \right| \\
&\le \left| R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widetilde{f}^{\xi}) \right| + \left| \min_{\xi \in \Xi} R(\widetilde{f}^{\xi}) - \min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) \right| \\
&\quad + \left| \min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) - \min_{1/\zeta_1 + 1/\zeta_1 \le n/p} R^{\det}(\widetilde{f}; \zeta_1, \zeta_2) \right|
\end{aligned}$$

We show below that each of the terms asymptotically go to zero. Observe that

$$|\Xi| = \sum_{\xi_1 = 2}^{\lceil n / \lfloor n^{\nu} \rfloor - 2 \rceil} (\xi_1 - 1) \le n^2.$$

Since $\widehat{\sigma}_{\Xi} = \widetilde{\sigma}_{\Xi} = o_p(\sqrt{n^{\nu} / \log(n)})$, under the setting of Lemma 3.2.4 or Lemma 3.2.5, Remark 3.2.8 hold so that

$$\left| R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widetilde{f}) \right| = o_p(1).$$

The assumption on the asymptotic risk profile (DETPA-1) leads to

$$\left| \min_{\xi \in \Xi} R(\widetilde{f}^{\xi}) - \min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) \right| = o_p(1).$$

Since the risk profile $R^{\det}(\widetilde{f}; \zeta_1, \zeta_2)$ is assumed be continuous at its minimizer, applying Lemma C.6.2 we get

$$\min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) \to \min_{1/\zeta_1 + 1/\zeta_2 \le n/p} R^{\det}(\widetilde{f}; \zeta_1, \zeta_2).$$

Combining the above three convergences, we have the desired conclusion.

Case of $M > 1$. When $m > 1$, we bound the desired difference as

$$\begin{aligned}
&\left(R(\widehat{f}^{\mathrm{cv}}) - \min_{1/\zeta_1 + 1/\zeta_2 \le n/p} R^{\det}(\widetilde{f}; \zeta_1, \zeta_2) \right)_+ \\
&\le \left(R(\widehat{f}^{\mathrm{cv}}) - \min_{\xi \in \Xi} R(\widehat{f}^{\xi}) \right)_+ + \left(\min_{\xi \in \Xi} R(\widehat{f}^{\xi}) - \frac{1}{M} \sum_{j=1}^{M} \min_{\xi \in \Xi} R(\widetilde{f}_j^{\xi}) \right)_+ \\
&\quad + \left(\frac{1}{M} \sum_{j=1}^{M} \min_{\xi \in \Xi} R(\widetilde{f}_j^{\xi}) - \min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}^{\xi}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) \right)_+ \\
&\quad + \left(\min_{\xi \in \Xi} R^{\det}\left(\widetilde{f}; \frac{p_n}{n - \xi_1 \lfloor n^{\nu} \rfloor}, \frac{p_n}{\xi_2 \lfloor n^{\nu} \rfloor} \right) - \min_{1/\zeta_1 + 1/\zeta_2 \le n/p} R^{\det}(\widetilde{f}; \zeta_1, \zeta_2) \right)_+
\end{aligned}$$

As before, we show below that each of the terms asymptotically vanish. Noting that $\widehat{\sigma}_{\Xi} \le \widetilde{\sigma}_{\Xi}$, application of Remark 3.2.8 shows that the first term is $o_p(1)$. The second term is 0 exactly as argued in the proof of Theorem 3.3.11. The third term is $o_p(1)$ by noting that (DETPA-1) holds for all $j = 1, \ldots, m$. Finally, the fourth term is 0 as argued for the case of $m = 1$.

C.5 Proofs related to deterministic profile verification for one-step procedure

In this section, we verify the assumption (DETPAR-1) for the one-step procedure, where the base prediction procedure is linear, under some regularity conditions. We also specifically consider the cases of MN2LS and MN1LS base prediction procedures.

C.5.1 Predictor simplifications and risk decompositions

In this section, we first provide preparatory lemmas that will be useful in the proofs of Lemma 3.4.8 and Corollary 3.4.9.

Let $\boldsymbol{X}_1 \in \mathbb{R}^{k_{1,m} \times p_m}$ and $\boldsymbol{Y}_1 \in \mathbb{R}^{k_{1,m}}$ denote the feature matrix and response vector corresponding to the first split dataset $\mathcal{D}_{k_{1,m}}$. Similarly, let $\boldsymbol{X}_2 \in \mathbb{R}^{k_{2,m} \times p_m}$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{k_{2,m}}$ denote the feature matrix and response vector corresponding to the second split dataset $\mathcal{D}_{k_{2,m}}$.

The following lemma gives an alternative representation for the ingredient one-step predictor assuming that the base prediction procedure is linear.

Lemma C.5.1 (Alternate representation for the ingredient one-step predictor)**.** *Suppose the base prediction procedure $\widetilde{f}$ is linear such that $\widetilde{f}(x; \mathcal{D}_{k_{1,m}}) = x^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}})$ for some estimator $\widetilde{\beta}(\mathcal{D}_{k_{1,m}})$ trained on $\mathcal{D}_{k_{1,m}}$. Let $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ denote the ingredient one-step predictor (3.51). Then, $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ is a linear predictor such that $\widetilde{f}(x; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) = x^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ with the corresponding ingredient one-step estimator $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{2,m})$ given by*

$$\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) = \big\{ I_p - (\boldsymbol{X}_2^T \boldsymbol{X}_2 / k_{2,m})^\dagger (\boldsymbol{X}_2^T \boldsymbol{X}_2 / k_{2,m}) \big\} \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + \widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_{2,m}}), \tag{C.65}$$

where $\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_{2,m}})$ is the MN2LS estimator fit on $\mathcal{D}_{k_{2,m}}$. Furthermore, suppose assumption (ℓ_2A1) holds true for $\mathcal{D}_{k_{2,m}}$. Then, the error between $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ and β_0 can be expressed as

$$\begin{aligned} &\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0 \\ &\quad = \big\{ I_p - (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m}) \big\} (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2 / k_{2,m}. \end{aligned} \tag{C.66}$$

Proof. For the first part, start by re-arranging the ingredient one-step predictor (3.51) as follows:

$$\begin{aligned} \widetilde{f}(x; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) &= \widetilde{f}(x; \mathcal{D}_{k_{1,m}}) + x^\top (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top (\boldsymbol{Y}_2 - \boldsymbol{X}_2 \widetilde{\beta}(\mathcal{D}_{k_{1,m}})) / k_{2,m} \\ &= x^\top \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + x^\top (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top (\boldsymbol{Y}_2 - \boldsymbol{X}_2 \widetilde{\beta}(\mathcal{D}_{k_{1,m}})) / k_{2,m} \\ &= x^\top \big\{ I_p - (\boldsymbol{X}_2^\top \boldsymbol{X} / k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2) / k_{2,m} \big\} \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + x^\top (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{Y}_2 / k_{2,m} \\ &= x^\top \big\{ I_p - (\boldsymbol{X}_2^\top \boldsymbol{X} / k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2) / k_{2,m} \big\} \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + x^\top \widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_{2,m}}), \end{aligned}$$

where $\widetilde{\beta}_{\mathrm{mn2}}(\mathcal{D}_{k_{2,m}}) = (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{Y}_2 / k_{2,m}$ is the MN2LS estimator fit on $\mathcal{D}_{k_{2,m}}$. Thus, $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ is a linear predictor with the corresponding ingredient one-step estimator $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{2,m})$ given by (C.65). This completes the proof of the first part.

For the second part, note that under linear model $\boldsymbol{Y}_2 = \boldsymbol{X}_2 \beta_0 + \boldsymbol{\varepsilon}_2$ (from (ℓ_2A1) for $\mathcal{D}_{k_{2,m}}$), the ingredient one-step estimator $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ can be further simplified to

$$\begin{aligned} &\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) \\ &= \big\{ I_p - (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m}) \big\} \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m}) \beta_0 \\ &\quad + (\boldsymbol{X}_2^\top \boldsymbol{X}_2 / k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2 / k_{2,m}. \end{aligned}$$

Hence, the error between $\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ and β_0 can be expressed as

$$\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0$$

$$\begin{aligned}
&= \{I_p - (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})\}\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})\beta_0 \\
&\quad + (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2/k_{2,m} - \beta_0 \\
&= \{I_p - (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})\}\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + \{(\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m}) - I_p\}\beta_0 \\
&\quad + (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2/k_{2,m} \\
&= \{I_p - (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})\}(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2/k_{2,m}.
\end{aligned}$$

This completes the proof of the second part. □

Recall that we are interested in the conditional squared prediction risk of $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$:

$$R_{\boldsymbol{X}_1,\boldsymbol{Y}_1,\boldsymbol{X}_2,\boldsymbol{Y}_2}(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})) = \mathbb{E}[(Y_0 - \widetilde{f}(X_0; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}))^2 \mid \boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2], \tag{C.67}$$

where (X_0, Y_0) is sampled independently and from the same distribution as the training data $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ and $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$. We are being explicit about the dependence of $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}))$ on $(\boldsymbol{X}_1, \boldsymbol{Y}_1, \boldsymbol{X}_2, \boldsymbol{Y}_2)$ as we will consider concentration of $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}))$ conditional on $(\boldsymbol{X}_1, \boldsymbol{Y}_1)$ first, followed by that on $(\boldsymbol{X}_2, \boldsymbol{Y}_2)$. For notational convenience, let $\widehat{\boldsymbol{\Sigma}}_1 := \boldsymbol{X}_1^T \boldsymbol{X}_1/k_{1,m}$ and $\widehat{\boldsymbol{\Sigma}}_2 := \boldsymbol{X}_2^T \boldsymbol{X}_2/k_{2,m}$ denote the sample covariance matrices for the two data splits $\mathcal{D}_{k_{1,m}}$ and $\mathcal{D}_{k_{2,m}}$, respectively. The next lemma gives conditional concentration of the squared prediction risk (C.67) of the one-step ingredient predictor under the additional assumptions (ℓ_2A2)–(ℓ_2A4) on $\mathcal{D}_{k_{2,m}}$.

Lemma C.5.2 (Conditional concentration of squared prediction risk of one-step ingredient predictor)**.** *Assume the setting of Lemma C.5.1. In addition, suppose assumptions (ℓ_2A2)–(ℓ_2A4) hold for $\mathcal{D}_{k_{2,m}}$. Let $k_{1,m}, k_{2,m}, p_m \to \infty$ such that $p_m/k_{2,m} \to \phi_2 \in (0,1) \cup (1,\infty)$ and assume* $\limsup \|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0\|_2 < \infty$ *almost surely. Then, we have*

$$\begin{aligned}
&R_{\boldsymbol{X}_1,\boldsymbol{Y}_1,\boldsymbol{X}_2,\boldsymbol{Y}_2}(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})) \\
&\quad - (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\Sigma(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) - \sigma^2 \operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma]/k_{2,m} - \sigma^2 \xrightarrow{\text{a.s.}} 0.
\end{aligned}$$

Proof. The proof follows similar steps as those in the proof of Proposition C.3.1. We start by decomposing the squared prediction risk:

$$R_{\boldsymbol{X}_1,\boldsymbol{Y}_1,\boldsymbol{X}_2,\boldsymbol{Y}_2}(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{2,m})) = (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{2,m}) - \beta_0)^\top \Sigma(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0) + \sigma^2. \tag{C.68}$$

Under (ℓ_2A1), from Lemma C.5.1, we have

$$\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0 = (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) + \widehat{\boldsymbol{\Sigma}}_2^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2/k_{2,m}.$$

Thus, the first term in the squared prediction risk (C.68) of $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ can be split into:

$$(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0)^\top \Sigma(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \beta_0) = \boldsymbol{B}_1 + \boldsymbol{C}_1 + \boldsymbol{V}_1,$$

where the terms $\boldsymbol{B}_1$, $\boldsymbol{C}_1$, and $\boldsymbol{V}_1$ are given as follows:

$$\begin{aligned}
\boldsymbol{B}_1 &= (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\Sigma(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0), \\
\boldsymbol{C}_1 &= (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\widehat{\boldsymbol{\Sigma}}_2^\dagger \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2/k_{2,m}, \\
\boldsymbol{V}_1 &= \boldsymbol{\varepsilon}_2(\boldsymbol{X}_2 \widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma \widehat{\boldsymbol{\Sigma}}_2^\dagger \boldsymbol{X}_2^\top/k_{2,m})\boldsymbol{\varepsilon}_2/k_{2,m}.
\end{aligned}$$

The rest of the proof shows concentration for the terms $\boldsymbol{C}_1$ and $\boldsymbol{V}_1$.

As argued in the proof of Proposition C.3.1, appealing to Lemma C.8.5 we have that $\boldsymbol{C}_1 \xrightarrow{\text{a.s.}} 0$ as $p_m, k_m \to \infty$ such that $p_m/k_{2,m} \to \phi \in (0,1) \cup (1,\infty)$, assuming $\limsup \|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0\|_2 < \infty$. This is because, from a bounding similar to (C.54), we have

$$\limsup \|\boldsymbol{X}_2 \widehat{\boldsymbol{\Sigma}}_2^\dagger (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)\|_2^2/k_{2,m} \le C \limsup \|\widetilde{\beta}(\mathcal{D}_{k_{1,m}} - \beta_0)\|_2^2 \le C,$$

almost surely for a constant $C < \infty$. Similarly, for the term $\boldsymbol{V}_1$, using Lemma C.8.6 along with the bound from (C.55), we have $\boldsymbol{V}_1 - \sigma^2 \operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma]/k_{2,m} \xrightarrow{\text{a.s.}} 0$. This finishes the proof. □

Lemma C.5.3 (Conditional deterministic approximation of squared risk of ingredient one-step predictor)**.** *Assume the setting of Lemma C.5.2. Let $k_{1,m}, k_{2,m}, p_m \to \infty$ such that $p_m/k_{2,m} \to \phi_2 \in (0,1) \cup (1,\infty]$. Then, we have*

$$R_{\boldsymbol{X}_1,\boldsymbol{Y}_1,\boldsymbol{X}_2,\boldsymbol{Y}_2}(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}})) - R^{\mathrm{g}}_{\boldsymbol{X}_1,\boldsymbol{Y}_1}(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}})) \xrightarrow{\text{a.s.}} 0,$$

where $R^{\mathrm{g}}_{\boldsymbol{X}_1,\boldsymbol{Y}_1}(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}}))$ is a certain generalized squared prediction risk of the predictor $\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}})$, fit on the first split data $\mathcal{D}_{k_{1,m}}$, given by

$$R^{\mathrm{g}}_{\boldsymbol{X}_1,\boldsymbol{Y}_1}(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}})) = \begin{cases} (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top \Sigma (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) + \sigma^2 & \text{if } \phi_2 = \infty \\ (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top g(\Sigma) (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) + \sigma^2 \operatorname{tr}[h(\Sigma)]/k_{2,m} + \sigma^2 & \text{if } \phi \in (1,\infty) \\ \sigma^2 \dfrac{1}{1-\phi_2} & \text{if } \phi \in (0,1), \end{cases} \tag{C.69}$$

where $g(\Sigma)$ and $h(\Sigma)$ are matrix functions of Σ given explicitly as follows:

$$g(\Sigma) = (1 + \widetilde{v}_g(0;\phi_2))(v(0;\phi_2)\Sigma + I_{p_m})^{-1}\Sigma(v(0;\phi_2)\Sigma + I_{p_m})^{-1}, \quad h(\Sigma) = \widetilde{v}(0;\phi_2)(v(0;\phi_2)\Sigma + I)^{-2}\Sigma^2,$$

and $v(0;\phi_2)$, $\widetilde{v}(0;\phi_2)$, and $\widetilde{v}_g(0;\phi_2)$ are as defined in (3.55), (3.56), and (3.57), respectively.

Proof. We will start with the functionals derived in Lemma C.5.2 and obtain corresponding asymptotic deterministic equivalents conditioned on $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ as $k_{1,m}, k_{2,m}, p_m \to \infty$, and $p_m/k_{2,m} \to \phi \in (0,1) \cup (1,\infty]$. We will split into three cases depending on where ϕ falls.

- $\underline{\phi_2 \in (0,1)}$: When $k_{1,m}, k_{2,m}, p_m \to \infty$ such that $p_m/k_{2,m} \to \phi_2 \in (0,1)$, $(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}) = 0$ almost surely and $\operatorname{tr}[\widehat{\boldsymbol{\Sigma}}^\dagger \Sigma]/k_{2,m} - \phi_2/(1-\phi_2) \xrightarrow{\text{a.s.}} 0$, as argued in the proof of Proposition C.3.2.

- $\underline{\phi \in (1,\infty)}$: Next we consider the case when $k_{1,m}, k_{2,m}, p_m \to \infty$, such that $p_m/k_{2,m} \to \phi \in (1,\infty)$. Consider the bias functional $(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\Sigma(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)$. Invoking Part 1 of Corollary C.6.12 with $f(\Sigma) = \Sigma$, as $k_{2,m}, p_m \to \infty$ such that $p_m/k_m \to \phi_2 \in (1,\infty)$, we have

  $$(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\Sigma(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2) \simeq (1 + \widetilde{v}_g(0;\phi_2))(v(0;\phi_2)\Sigma + I_{p_m})^{-1}\Sigma(v(0;\phi_2)\Sigma + I_{p_m})^{-1},$$

  where $v(0;\phi_2)$ and $\widetilde{v}_g(0;\phi_2)$ are as defined in (3.55) and (3.57), respectively. Now, note that the vector $(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)$ is independent of $\widehat{\boldsymbol{\Sigma}}_2^\dagger$. Thus, from the definition of asymptotic equivalence, we have

  $$(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top (I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)\Sigma(I_p - \widehat{\boldsymbol{\Sigma}}_2^\dagger \widehat{\boldsymbol{\Sigma}}_2)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) - (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top g(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) \xrightarrow{\text{a.s.}} 0.$$

  Consider now the variance resolvent $\widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma$. From Part 2 of Corollary C.6.12 with $f(\Sigma) = \Sigma$, as $k_{2,m}, p_m \to \infty$ such that $p_m/k_{2,m} \to \phi_2 \in (1,\infty)$, we have

  $$\widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma \simeq \widetilde{v}(0;\phi_2)(v(0;\phi_2)\Sigma + I_{p_m})^{-2}\Sigma^2.$$

  Hence, using Lemma C.7.2 (4), we have

  $$\sigma^2 \operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_2^\dagger \Sigma]/k_{2,m} - \sigma^2 \operatorname{tr}[\widetilde{v}(0;\phi_2)(v(0;\phi_2)\Sigma + I_{p_m})^{-2}\Sigma^2]/k_{2,m} \xrightarrow{\text{a.s.}} 0.$$

- $\underline{\phi_2 = \infty}$: Finally, consider the case when $k_{1,m}, k_{2,m}, p_m \to \infty$ and $p_m/k_{2,m} \to \infty$. We start by expressing the ingredient one-step estimator (3.51) as

  $$\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) = \widetilde{\beta}(\mathcal{D}_{k_{1,m}}) + (\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2^\top (\boldsymbol{Y}_2 - \boldsymbol{X}_2 \widetilde{\beta}(\mathcal{D}_{k_{1,m}}))/k_{2,m}.$$

  Using triangle inequality, note that

  $$\|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \widetilde{\beta}(\mathcal{D}_{k_{1,m}})\|_2 = \|(\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2^\top (\boldsymbol{Y}_2 - \boldsymbol{X}_2 \widetilde{\beta}(\mathcal{D}_{k_{1,m}}))/k_{2,m}\|_2$$

$$\le \|(\boldsymbol{X}_2^\top \boldsymbol{X}_2/k_{2,m})^\dagger \boldsymbol{X}_2/\sqrt{k_{2,m}}\|_{\mathrm{op}} \|\boldsymbol{Y}_2 - \boldsymbol{X}_2\widetilde{\beta}(\mathcal{D}_{k_{1,m}})/\sqrt{k_{2,m}}\|_2.$$

Under the setting of Lemma C.5.2, the second term in the display above is almost surely bounded. Hence, following the proof of Proposition C.3.3, it follows that $\|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}) - \widetilde{\beta}(\mathcal{D}_{k_{1,m}})\|_2 \xrightarrow{\text{a.s.}} 0$. From the analogous reasoning in the proof of Proposition C.3.3, this in turn implies that

$$R_{\boldsymbol{X}_1,\boldsymbol{Y}_1,\boldsymbol{X}_2,\boldsymbol{Y}_2}(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})) - (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top \Sigma (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) - \sigma^2 \xrightarrow{\text{a.s.}} 0.$$

This completes all three cases and finishes the proof. □

C.5.2 Proof of Lemma 3.4.8

The idea of the proof is to use the conditional deterministic risk approximation derived in Lemma C.5.3 and obtain a limiting expression for the deterministic approximation in terms of the assumed limiting distribution (3.52).

We start by noting that

$$\|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0\|_2^2 \le r_{\min}^{-1} \|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0\|_\Sigma^2.$$

Thus, under the assumption that there exists a deterministic approximation $R^{\mathrm{det}}(\phi_1; \widetilde{f})$ to the conditional risk of $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})$ such that $R(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}})) \xrightarrow{\text{p}} R^{\mathrm{det}}(\phi_1; \widetilde{f})$ as $k_{1,m}, p_m \to \infty$ and $p_m/k_{1,m} \to \phi_1$, for ϕ_1 satisfying $R^{\mathrm{det}}(\phi_1; \widetilde{f}) < \infty$, it follows that $\limsup \|\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0\|_2 < \infty$. We can now invoke Lemma C.5.3. Let $k_{2,m} \to \infty$ such that $p_m/k_{2,m} \to \phi_2 \in (0,1) \cup (1,\infty]$. We will split into various cases depending on ϕ_2.

1. The limit for $\phi_2 = \infty$ is clear from the $\phi_2 = \infty$ case in (C.69).

2. When $\phi_2 \in (1,\infty)$, we need to obtain limiting expressions for the quantities $(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top g(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)$ and $\mathrm{tr}[h(\Sigma)]/k_{2,m} = \mathrm{tr}[\widetilde{v}(0;\phi_2)\Sigma^2(v(0;\phi_2)\Sigma + I)^{-2}]/k_{2,m}$ in terms of the limiting distributions Q and H.

   For the former, we start by expanding the quadratic form:

$$\begin{aligned}
&(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top g(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) \\
&= (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top W g(R) W^\top (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) \\
&= \sum_{i=i}^{p_m} ((\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top w_i)^2 g(r_i) \\
&= \sum_{i=1}^{p_m} ((\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top w_i)^2 r_i \sum_{i=1}^{p_m} \frac{((\widetilde{\beta}(\mathcal{D}_{k_{1.m}}) - \beta_0)^\top w_i)^2 r_i \cdot g(r_i)/r_i}{\sum_{i=1}^{p_m}((\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top w_i)^2 r_i} \\
&= (R(\widetilde{f}(\cdot;\mathcal{D}_{1,m})) - \sigma^2) \int \widetilde{g}(r)\,\mathrm{d}\widehat{Q}_n(r),
\end{aligned} \tag{C.70}$$

   where $\widetilde{g}(r)$ is given by

$$\widetilde{g}(r) = \frac{g(r)}{r} = (1 + \widetilde{v}_g(0;\phi_2)) \frac{1}{(v(0;\phi_2)r + 1)^2}.$$

   Under the assumption that $\widehat{Q}_n \xrightarrow{\text{d}} Q$ in probability, we have

$$\int \widetilde{g}(r)\,\mathrm{d}\widehat{Q}_n(r) \xrightarrow{\text{p}} \int \widetilde{g}(r)\,\mathrm{d}Q(r) = \int \frac{(1 + \widetilde{v}_g(0;\phi_2))}{(v(0;\phi_2)r + 1)^2}\,\mathrm{d}Q(r). \tag{C.71}$$

   Observe that $\widetilde{g}$ is continuous. Since $R(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}})) \xrightarrow{\text{a.s.}} R^{\mathrm{det}}(\psi_1; \widetilde{f})$, from (C.70) and (C.71), we have

$$(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top g(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) \xrightarrow{\text{p}} (R^{\mathrm{det}}(\phi_1;\widetilde{f}) - \sigma^2)(1 + \widetilde{v}_g(0;\phi_2)) \int \frac{1}{(v(0;\phi_2)r + 1)^2}\,\mathrm{d}Q(r)$$

$$= R^{\det}(\phi_1; \widetilde{f})\Upsilon_b(\phi_1, \phi_2) - \sigma^2 \Upsilon_b(\phi_1, \phi_2), \tag{C.72}$$

where $\Upsilon_b(\phi_1, \phi_2)$ is as defined in (3.58).

For the latter, using Lemma C.7.2 (4) and noting that the integrand is continuous, we have

$$\begin{aligned} \operatorname{tr}[h(\Sigma)]/k_{2,m} = \frac{p_m}{k_{2,m}} \widetilde{v}(0; \phi_2) \int \frac{r^2}{(1 + v(0; \phi_2) r)^2} \, \mathrm{d}H_{p_m}(r) &\xrightarrow{\text{a.s.}} \phi_2 \widetilde{v}(0; \phi_2) \int \frac{\rho^2}{(v(0; \phi_2) r + 1)^2} \, \mathrm{d}H(r) \\ &= \widetilde{v}_g(0; \phi_2), \end{aligned} \tag{C.73}$$

where $\widetilde{v}_g(0; \phi_2)$ is as defined in (3.57).

Putting (C.69), (C.72), and (C.73) together, the result follows for $\phi_2 \in (1, \infty)$.

3. The final case of $\phi_2 \in (0, 1)$ follows analogous argument as in the proof of Proposition C.3.2.

This completes the proof.

C.5.3 Proof of Corollary 3.4.9

We will show that there exists a deterministic risk approximation $R^{\det}(\cdot, \cdot; \widetilde{f}) : (0, \infty] \times (0, \infty] \to [0, \infty]$ to the conditional prediction risk $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}}))$ of the one-step ingredient predictor $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ that satisfies the three-point program (PRG-1-C1)–(PRG-1-C3). In particular, we will show that the following $R^{\det}(\cdot, \cdot; \widetilde{f})$, that is a continuation of (3.54), satisfies the required conditions:

$$R^{\det}(\phi_1, \phi_2; \widetilde{f}) = \begin{cases} R^{\det}(\phi_1; \widetilde{f}) & \text{if } \phi_2 = \infty \\ (R^{\det}(\phi_1; \widetilde{f}) - \sigma^2)\Upsilon_b(\phi_1, \phi_2) + \sigma^2(1 - \Upsilon_b(\phi_1, \phi_2)) + \sigma^2 \widetilde{v}_g(0; \phi_2) & \text{if } \phi_2 \in (1, \infty) \\ \infty & \text{if } \phi_2 = 1 \\ \sigma^2 \dfrac{\phi_2}{1 - \phi_2} & \text{if } \phi_2 \in (0, 1), \end{cases}$$

where $R^{\det}(\cdot; \widetilde{f})$ is the assumed deterministic risk approximation to the conditional prediction risk $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}))$ of the base predictor $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})$, and $\Upsilon_b(\cdot; \cdot)$ and $\widetilde{v}_g(0; \cdot)$ are as defined in (3.58). Below we split the three verifications:

1. Let $\Phi_1^\infty := \{\phi_1 \in (0, \infty] : R^{\det}(\phi_1; \widetilde{f}) = \infty\}$ denote the set of limiting aspect ratios greater than one, where the deterministic risk approximation to the base procedure is ∞. By the hypothesis of Lemma 3.4.8, we have $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}})) \xrightarrow{\text{p}} R^{\det}(\phi_1; \widetilde{f})$ as $k_{1,m}, p_m \to \infty$ and $p_m/k_{1,m} \to \phi_1 \in (0, \infty] \setminus \Phi_1^\infty$. Now observe that $R^{\det}(\phi_1, \phi_2; \widetilde{f}) = \infty$ only at $\Phi^\infty := \{(\phi_1, \phi_2) : \phi_1 \in \Phi_1^\infty \text{ or } \phi_2 = 1\}$. This is because $\Upsilon_b(\phi_1, \phi_2), \widetilde{v}_g(0; \phi_2) < \infty$ for $\phi_2 \in (1, \infty)$ from Lemma C.6.13 (5). Note from the conclusion of Lemma 3.4.8 that $R(\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})) \xrightarrow{\text{p}} R^{\det}(\phi_1, \phi_2; \widetilde{f})$ as $k_{1,m}, k_{2,m}, p_m \to \infty$ and $(p_m/k_{1,m}, p_m/k_{2,m}) \to (\phi_1, \phi_2) \in (0, \infty] \times (0, \infty] \setminus \Phi^\infty$, or in other words, continuous convergence of the risk to the deterministic approximation holds for all limiting (ϕ_1, ϕ_2) for which $R^{\det}(\phi_1, \phi_2; \widetilde{f}) < \infty$. This verifies (PRG-1-C1).

2. From the argument above, we have $R^{\det}(\phi_1, \phi_2; \widetilde{f}) = \infty$ over Φ^∞. Pick any $(\phi_1, \phi_2) \in \Phi^\infty$. We will show that $R^{\det}(\phi_1', \phi_2'; \widetilde{f}) \to \infty$ as $(\phi_1', \phi_2') \to (\phi_1, \phi_2)$. From the definition of Φ^∞, the point (ϕ_1, ϕ_2) falls into either of the following two cases:

   - $\phi_2 = 1$: In this case, observe that $R^{\det}(\phi_1', \phi_2') \to \infty$ as $(\phi_1', \phi_2') \to (\phi_1, 1^+)$ because $\lim_{\phi_2' \to 1^-} \phi_2'/(1 - \phi_2') = \infty$, and $R^{\det}(\phi_1', \phi_2') \to \infty$ as $(\phi_1', \phi_2') \to (\phi_1, 1^+)$ because, from Lemma C.6.13 (5), $\lim_{\phi_2' \to 1^+} \widetilde{v}_g(0; \phi_2') = \infty$. Thus, $R^{\det}(\phi_1', \phi_2') \to \infty$ as $(\phi_1', \phi_2') \to (\phi_1, \phi_2)$.

- $\phi_1 \in \Phi_1^\infty$: In this case, $R^{\det}(\phi_1') \to \infty$ as $\phi_1' \to \phi_1$ from the assumption that $R^{\det}(\cdot; \widetilde{f})$ satisfies (PRG-0-C2). Because $\Upsilon_b(\phi_1', \phi_2'), \widetilde{v}_g(0; \phi_2') > 0$ over $(\phi_1', \phi_2') \in (0, \infty] \times (1, \infty]$ from arguments in Lemma C.6.13 (4) and Lemma C.6.13 (5), it follows that

$$\lim_{(\phi_1', \phi_2') \to (\phi_1, \phi_2)} R^{\det}(\phi_1', \phi_2'; \widetilde{f}) = \lim_{\phi_1' \to \phi_1} R^{\det}(\phi_1'; \widetilde{f}) = \infty.$$

Thus, $R^{\det}(\phi_1', \phi_2') \to \infty$ as $(\phi_1', \phi_2') \to (\phi_1, \phi_2)$.

Therefore, whenever $(\phi_1', \phi_2') \to (\phi_1, \phi_2)$, we have $R^{\det}(\phi_1', \phi_2'; \widetilde{f}) \to \infty$, and thus $R^{\det}(\cdot, \cdot; \widetilde{f})$ satisfies (PRG-1-C2).

3. Finally, the set of (ϕ_1, ϕ_2) such that $R^{\det}(\phi_1, \phi_2; \widetilde{f}) = \infty$ is Φ^∞. Because Φ^∞ is product of two sets each of which is closed in $\mathbb{R}$, this set is closed in $\mathbb{R}^2$. Therefore, $R^{\det}(\cdot, \cdot; \widetilde{f})$ satisfies (PRG-1-C3).

Put together, all of (PRG-1-C1)–(PRG-1-C3) hold, and this in turn implies that $\widetilde{f}(\cdot; \mathcal{D}_{k_{1,m}}, \mathcal{D}_{k_{2,m}})$ satisfies (DETPAR-1). This finishes the proof.

C.5.4 Proof of Proposition 3.4.10

It suffices to verify the hypothesis of Lemma 3.4.8 and then appeal to Corollary 3.4.9. We will use Corollary C.6.12 along with the Portmanteau theorem to certify existence of a limiting distribution Q assumed in Lemma 3.4.8. The form of Q is defined through limiting formulas for the generalized prediction risks of the base predictor.

Let f be any continuous and bounded function. We will show that $\int f(r) \, \mathrm{d}\widehat{Q}_n(r)$ converges to a deterministic limit that is a function of H and G, and show existence of Q through this limit. We start by noting that

$$\int f(r) \, \mathrm{d}\widehat{Q}_n(r) = (\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top f(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0), \tag{C.74}$$

where $f(\Sigma) = W f(R) W^\top$, and $f(R)$ is a matrix obtained by applying f component-wise to the diagonal entries of R. We will now obtain a limiting expression for the term on the right hand side of (C.74), which has the form of a generalized prediction risk of $\widetilde{\beta}(\mathcal{D}_{k_{1,m}})$. Similar to the proof of Proposition 3.3.14, we will first obtain a deterministic equivalent for the generalized prediction risk. Following similar steps as in the proof of Proposition C.3.1, we have that

$$(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0)^\top f(\Sigma)(\widetilde{\beta}(\mathcal{D}_{k_{1,m}}) - \beta_0) - \beta_0^\top (I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1) f(\Sigma)(I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1)\beta_0 + \mathrm{tr}[\widehat{\mathbf{\Sigma}}_1^\dagger f(\Sigma)]/k_{1,m} \xrightarrow{\text{a.s.}} 0. \tag{C.75}$$

Now, using first part of Corollary C.6.12, we can write

$$(I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1) f(\Sigma)(I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1) \simeq (1 + \widetilde{v}_g(0; \phi_1))(v(0; \phi_1)\Sigma + I_{p_m})^{-1} \Sigma (v(0; \phi_1)\Sigma + I_{p_m})^{-1}.$$

Using Property 4 of Appendix C.7, this then yields

$$\beta_0^\top (I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1) f(\Sigma)(I_p - \widehat{\mathbf{\Sigma}}_1^\dagger \widehat{\mathbf{\Sigma}}_1)\beta_0 \xrightarrow{\text{a.s.}} (1 + \widetilde{v}_g(0; \phi_1)) \int \frac{f(r)}{(v(0; \phi_1) r + 1)^2} \, \mathrm{d}G(r). \tag{C.76}$$

Similarly, using second part of Corollary C.6.12, we have

$$\widehat{\mathbf{\Sigma}}_1^\dagger f(\Sigma) \simeq \widetilde{v}(0; \phi_1)(v(0; \phi_1)\Sigma + I_{p_m})^{-2} \Sigma f(\Sigma).$$

Hence, appealing to Property 4 of Appendix C.7 again, we have

$$\mathrm{tr}[\widehat{\mathbf{\Sigma}}_1^\dagger f(\Sigma)]/k_{1,m} \xrightarrow{\text{a.s.}} \phi_1 \widetilde{v}(0; \phi_1) \int \frac{r f(r)}{(v(0; \phi_1) r + 1)^2} \, \mathrm{d}H(r). \tag{C.77}$$

Therefore, from (C.74)–(C.77), it follows that

$$\int f(r)\,\mathrm{d}\widehat{Q}_n(r) \xrightarrow{\text{a.s.}} (1+\widetilde{v}_g(0;\phi_1))\int \frac{f(r)}{(v(0;\phi_1)r+1)^2}\,\mathrm{d}G(r) + \phi_1\widetilde{v}(0;\phi_1)\int \frac{rf(r)}{(v(0;\phi_1)r+1)^2}\,\mathrm{d}H(r).$$

Observe that this defines a distribution Q because one can take $f(r) = e^{itr} = \cos(tr) + i\sin(tr)$, which then implies convergence of the characteristic function at all points. This finishes the proof. To get more insight into the risk behaviour of the ingredient one-step predictor, we can also write out an explicit formula for the deterministic approximation $R^{\mathrm{det}}(\cdot,\cdot;\widetilde{f}_{\mathrm{mn2}})$. We will do so below.

For the particular functional $R(\widetilde{f}(\cdot;\mathcal{D}_{k_{1,m}},\mathcal{D}_{k_{2,m}}))$, we have a specific f given by

$$f(r) = (1+\widetilde{v}_g(0;\phi_2))\frac{r}{(v(0;\phi_2)r+1)^2}.$$

Thus, the final expression for $R^{\mathrm{det}}(\phi_1,\phi_2)$ can be written explicitly as follows:

$$R^{\mathrm{det}}(\phi_1,\phi_2)$$

$$= \begin{cases} R^{\mathrm{det}}(\min\{\phi_1,\phi_2\}) & \text{if } \phi_1 = \infty \text{ or } \phi_2 = \infty \\ \rho^2(1+\widetilde{v}_g(0;\phi_1,\phi_2))(1+\widetilde{v}_g(0;\phi_2))\displaystyle\int \frac{r}{(1+v(0;\phi_1)r)^2(1+v(0;\phi_2)r)^2}\,\mathrm{d}G(r) & \\ \quad + \sigma^2(1+\widetilde{v}_g(0;\phi_2))\phi_1\widetilde{v}(0;\phi_1)\displaystyle\int \frac{r}{(v(0;\phi_1)r+1)^2(v(0;\phi_2)r+1)^2}\,\mathrm{d}H(r) & \\ \quad + \sigma^2\left(\phi_2\widetilde{v}(0;\phi_2)\displaystyle\int \frac{r}{(1+v(0;\phi_2)r)^2}\,\mathrm{d}H(r) + 1\right) & \text{if } (\phi_1,\phi_2) \in (1,\infty)\times(1,\infty) \\ \sigma^2\left(\phi_2\widetilde{v}(0;\phi_2)\displaystyle\int \frac{r}{(1+v(0;\phi_2)r)^2}\,\mathrm{d}H(r) + 1\right) & \text{if } (\phi_1,\phi_2) \in (0,1)\times(1,\infty) \\ \sigma^2\dfrac{1}{1-\phi_2} & \text{if } (\phi_1,\phi_2) \in (0,\infty)\times(0,1), \end{cases}$$

where $v(0;\phi)$ is as defined in (C.46), $\widetilde{v}(0;\phi)$ is as defined in (C.47), $\widetilde{v}_g(0;\phi)$ is as defined in (C.48), and $\widetilde{v}_g(0;\phi_1,\phi_2)$ is as defined below:

$$\widetilde{v}_g(0;\phi_1,\phi_2) = \frac{(1+\widetilde{v}_g(0;\phi_2))\phi_1\displaystyle\int \frac{r^2}{(1+v(0;\phi_2)r)^2(1+v(0;\phi_1)r)^2}\,\mathrm{d}H(r)}{\dfrac{1}{v(0;\phi_1)^2} - \phi_1\displaystyle\int \frac{r^2}{(1+v(0;\phi_1)r)^2}\,\mathrm{d}H(r)}.$$

Here, $R^{\mathrm{det}}(\cdot)$ is $R^{\mathrm{det}}(\cdot;\widetilde{f}_{\mathrm{mn2}})$ as defined in (C.45).

C.5.5 Proof of Proposition 3.4.11

Verification of the hypothesis of Lemma 3.4.8 is easy in this case because $\Sigma = I_p$. Observe that under (ℓ_1A2), the distribution $\widehat{Q}_n$ is simply a point mass at 1. Thus, the hypothesis of Lemma 3.4.8 is trivially satisfied. Moreover, we can explicitly write expressions for the functions $\widetilde{v}_g(0;\cdot)$ and $\Upsilon_b(\cdot;\cdot)$. Towards that end, we will first obtain expressions for the ingredient functions $v(0;\cdot)$ and $\widetilde{v}(0;\cdot)$.

- $\underline{v(0;\phi_2)}$: The fixed-point equation (3.55) can be solved explicitly since H is a point mass at 1. The fixed-point equation in this case simplifies to

$$\frac{1}{v(0;\phi_2)} = \phi_2\frac{1}{v(0;\phi_2)+1}. \tag{C.78}$$

  Solving (C.78) for $v(0;\phi_2)$, we get

$$v(0;\phi_2) = \frac{1}{\phi_2 - 1}, \quad \text{and} \quad 1 + v(0;\phi_2) = \frac{\phi_2}{\phi_2 - 1}. \tag{C.79}$$

- $\underline{\widetilde{v}(0;\phi_2)}$: Using (C.79), we can compute the inverse of $\widetilde{v}(0;\phi_2)$ per (3.56) as

$$\widetilde{v}(0;\phi_2)^{-1} = (\phi_2 - 1)^2 - \phi_2 \frac{(\phi_2 - 1)^2}{\phi_2^2} = (\phi_2 - 1)^2 - \frac{(\phi_2 - 1)^2}{\phi_2} = (\phi_2 - 1)^2 \frac{\phi_2 - 1}{\phi_2} = \frac{(\phi_2 - 1)^3}{\phi_2}.$$

  Thus, we have

$$\widetilde{v}(0;\phi_2) = \frac{\phi_2}{(\phi_2 - 1)^3}, \quad \text{and} \quad \widetilde{v}(0;\phi_2)\phi_2 = \frac{\phi_2^2}{(\phi_2 - 1)^3}. \tag{C.80}$$

Using (C.79) and (C.80), we can explicitly write out expressions for $\Upsilon_b(\phi_1, \phi_2)$ and $\widetilde{v}_g(0;\phi_2)$.

- $\underline{\widetilde{v}_g(0;\phi_2)}$: Substituting (C.79) and (C.80) into (3.57), we obtain

$$\widetilde{v}_g(0;\phi_2) = \frac{\phi_2^2}{(\phi_2 - 1)^3} \frac{(\phi_2 - 1)^2}{\phi_2^2} = \frac{1}{\phi_2 - 1}, \quad \text{and} \quad (1 + \widetilde{v}_g(0;\phi_2)) = \frac{\phi_2}{\phi_2 - 1}. \tag{C.81}$$

- $\underline{\Upsilon_b(\phi_1, \phi_2)}$: Substituting (C.79) and (C.80) into (3.58), we get

$$\Upsilon_b(\phi_1, \phi_2) = \frac{\phi_2}{\phi_2 - 1} \frac{(\phi_2 - 1)^2}{\phi_2^2} = \frac{\phi_2 - 1}{\phi_2}, \quad \text{and} \quad 1 - \Upsilon_b(\phi_1, \phi_2) = \frac{1}{\phi_2}. \tag{C.82}$$

  Observe that since the distribution Q does not depend on ϕ_1 in this case, $\Upsilon_b(\phi_1, \phi_2)$ in turn also does not depend on ϕ_1.

Therefore, using (C.81) and (C.82), the deterministic risk approximation from (3.54) simplifies in this case as follows:

$$R^{\det}(\phi_1, \phi_2; \widetilde{f}) \to \begin{cases} \rho^2 + \sigma^2 & \text{if } \phi_1 = \phi_2 = \infty \\ R^{\det}(\phi_1) & \text{if } \phi_2 = \infty \\ \rho^2 \left(1 - \dfrac{1}{\phi_2}\right) + \sigma^2 \left(\dfrac{1}{\phi_2 - 1}\right) + \sigma^2 & \text{if } \phi_1 = \infty \\ R^{\det}(\phi_1) \left(1 - \dfrac{1}{\phi_2}\right) + \sigma^2 \left(\dfrac{1}{\phi_2 - 1}\right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (1, \infty) \times (1, \infty) \\ \sigma^2 \left(\dfrac{\phi_1}{1 - \phi_1}\right) \left(1 - \dfrac{1}{\phi_2}\right) + \sigma^2 \left(\dfrac{1}{\phi_2 - 1}\right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (0, 1) \times (1, \infty) \\ \sigma^2 \left(\dfrac{\phi_2}{1 - \phi_2}\right) + \sigma^2 & \text{if } (\phi_1, \phi_2) \in (0, \infty) \times (0, 1). \end{cases}$$

Here, $R^{\det}(\cdot)$ is $R^{\det}(\cdot; \widetilde{f}_{\mathrm{mn1}})$ as defined in (C.60).

C.6 Technical lemmas and miscellaneous details

In this section, we gather various technical lemmas along with their proofs, and other miscellaneous details. Specific pointers to which lemmas are used in which proofs are provided at the start of each section.

C.6.1 Lemmas for verifying space-filling properties of discrete optimization grids

In this section, we collect supplementary lemmas that are used in the proofs of Theorems 3.3.11 and 3.4.4 in Appendices C.2 and C.4, respectively.

Lemma C.6.1 (Verifying space-filling property of the discrete grid used in the zero-step procedure)**.** *Let $\{p_n\}$, $\{m_{1,n}\}$, $\{m_{2,n}\}$ are three sequences of positive integers such that $m_{2,n} \le m_{1,n}$ for $n \ge 1$. Suppose*

$$\frac{p_n}{m_{1,n}} \to \gamma \in (0, \infty) \quad \textit{and} \quad \frac{m_{2,n}}{m_{1,n}} \to 0$$

as $n \to \infty$. Define a sequence of grids $\mathcal{G}_n$ as follows:

$$\mathcal{G}_n := \left\{ \frac{p_n}{m_{1,n} - k m_{2,n}} : 1 \le k \le \left\lceil \frac{m_{1,n}}{m_{2,n}} - 2 \right\rceil \right\}.$$

Then, for any $\zeta^\star \in [\gamma, \infty]$, $\Pi_{\mathcal{G}_n}(\zeta^\star) \to \zeta^\star$ as $n \to \infty$, where $\Pi_{\mathcal{G}_n}(y) = \arg\min_{x \in \mathcal{G}_n} |y - x|$ is the point in the grid $\mathcal{G}_n$ closest to y. In particular, in the context of Algorithm 2, taking $m_{1,n} = n_{\mathrm{tr}}$ and $m_{2,n} = \lfloor n^\nu \rfloor$ for $\nu \in (0, 1)$, we get the aspect ratios used in Algorithm 2 "converge" to $[\gamma, \infty]$ when $n_{\mathrm{tr}}/n \to 1$ under (PA(γ)).

Proof. We will consider different cases depending on where $\zeta^\star \in [\gamma, \infty]$ lands. See Figure C.3.

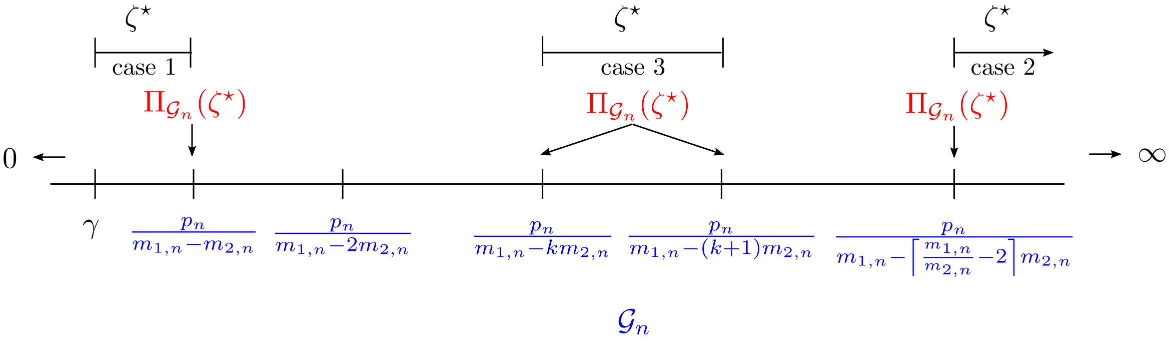

$\zeta^\star$
case 1
$\zeta^\star$
case 3
$\zeta^\star$
case 2
$\Pi_{\mathcal{G}_n}(\zeta^\star)$
$\Pi_{\mathcal{G}_n}(\zeta^\star)$
$\Pi_{\mathcal{G}_n}(\zeta^\star)$
0
∞
γ
$\frac{p_n}{m_{1,n}-m_{2,n}}$
$\frac{p_n}{m_{1,n}-2m_{2,n}}$
$\frac{p_n}{m_{1,n}-km_{2,n}}$
$\frac{p_n}{m_{1,n}-(k+1)m_{2,n}}$
$\frac{p_n}{m_{1,n}-\lceil \frac{m_{1,n}}{m_{2,n}}-2 \rceil m_{2,n}}$
$\mathcal{G}_n$

Figure C.3: Illustration of different cases of $\zeta \in [\gamma, \infty]$ and the corresponding projection $\Pi_{\mathcal{G}_n}(\zeta^\star)$.

1. Consider the first case when
$$\gamma \le \zeta^\star \le \frac{p_n}{m_{1,n} - m_{2,n}}.$$
In this case, $\Pi_{\mathcal{G}_n}(\zeta^\star)$ is simply the first point in the grid. Observe that in this case
$$\Pi_{\mathcal{G}_n}(\zeta^\star) - \zeta^\star \le \frac{p_n}{m_{1,n} - m_{2,n}} - \gamma = \frac{\dfrac{p_n}{m_{1,n}}}{1 - \dfrac{m_{2,n}}{m_{1,n}}} - \gamma \to \gamma - \gamma = 0$$
as $n \to \infty$ under the assumptions that $p_n/m_{1,n} \to \gamma$ and $m_{2,n}/m_{1,n} \to 0$.

2. Consider the second case when
$$\frac{p_n}{m_{1,n} - \left\lceil \dfrac{m_{1,n}}{m_{2,n}} - 2 \right\rceil} \le \zeta^\star \le \infty.$$
In this case, $\Pi_{\mathcal{G}_n}(\zeta^\star)$ is simply the last point in the grid. We will show eventually the only $\zeta^\star$ in this case is $\zeta^\star = \infty$. Note that $p_n/(m_{1,n} - k m_{2,n})$ increases with $k \ge 0$. If $\zeta^\star = \infty$, then $\Pi_{\mathcal{G}_n}(\zeta^\star) = p_n/(m_{1,n} - k^\star m_{2,n})$ for $k^* = \lceil m_{1,n}/m_{2,n} - 2 \rceil$. Hence, it suffices to prove that $p_n/(m_{1,n} - k^\star m_{2,n}) \to \infty$ as $n \to \infty$. This follows from the fact that
$$\frac{m_{1,n}}{m_{2,n}} - \left\lceil \frac{m_{1,n}}{m_{2,n}} - 2 \right\rceil \le 2,$$

and thus

$$\frac{p_n}{m_{1,n} - k^\star m_{2,n}} = \frac{p_n}{m_{2,n}(m_{1,n}/m_{2,n} - \lceil m_{1,n}/m_{2,n} - 2 \rceil)} \geq \frac{p_n}{2m_{2,n}} \to \infty = \zeta^*,$$

as $n \to \infty$ and $p_n/m_{1,n} \to \gamma \in (0, \infty)$.

3. Consider the third case when

$$\frac{p_n}{m_{1,n} - km_{2,n}} \leq \zeta^\star \leq \frac{p_n}{m_{1,n} - (k+1)m_{2,n}} \quad \text{for some } 1 \leq k \leq \left\lceil \frac{m_{1,n}}{m_{2,n}} - 2 \right\rceil. \tag{C.83}$$

From the first inequality in (C.83), we have

$$\frac{p_n}{m_{1,n} - km_{2,n}} \leq \zeta^\star \implies \frac{p_n}{m_{1,n}\zeta^\star} \leq 1 - k\frac{m_{2,n}}{m_{1,n}} \implies k\frac{m_{2,n}}{m_{1,n}} \leq 1 - \frac{p_n}{m_{1,n}\zeta^\star}. \tag{C.84}$$

Similarly, from the second inequality of (C.83), we have

$$\frac{p_n}{m_{1,n}\zeta^\star} \geq 1 - \frac{(k+1)m_{2,n}}{m_{1,n}} \implies k\frac{m_{2,n}}{m_{1,n}} \geq 1 - \frac{p_n}{m_{1,n}\zeta^\star} - \frac{m_{2,n}}{m_{1,n}}. \tag{C.85}$$

The upper and lower bounds from (C.85) and (C.84) together imply that

$$1 - \frac{p_n}{m_{1,n}\zeta^\star} - \frac{m_{2,n}}{m_{1,n}} \leq \frac{km_{2,n}}{m_{1,n}} \leq 1 - \frac{p_n}{m_{1,n}\zeta^\star}.$$

Because $\lim_{n\to\infty} m_{2,n}/m_{1,n} = 0$, we conclude that

$$\lim_{n\to\infty} \frac{km_{2,n}}{m_{1,n}} = 1 - \frac{\gamma}{\zeta^\star} \in (0, 1). \tag{C.86}$$

Now, note that since $\Pi_{\mathcal{G}_n}(\zeta^\star)$ is either of the two points of the grid partition, we have

$$\begin{aligned}
|\Pi_{\mathcal{G}_n}(\zeta^\star) - \zeta^\star| &\leq \frac{p_n}{m_{1,n} - (k+1)m_{2,n}} - \frac{p_n}{m_{1,n} - km_{2,n}} \\
&= \frac{p_n}{m_{1,n} - (k+1)m_{2,n}} \frac{m_{2,n}}{m_{1,n} - km_{2,n}} \\
&= \frac{\dfrac{p_n}{m_{1,n}}}{1 - \dfrac{(k+1)m_{2,n}}{m_{1,n}}} \frac{\dfrac{m_{2,n}}{m_{1,n}}}{1 - \dfrac{km_{2,n}}{m_{1,n}}} \\
&\to \frac{\gamma}{1 - \left(1 - \dfrac{\gamma}{\zeta^\star}\right)} \frac{0}{\left(1 - \left(1 - \dfrac{\gamma}{\zeta^\star}\right)\right)} = 0,
\end{aligned}$$

as $n \to \infty$ and $p_n/m_{1,n} \to \gamma$ and $m_{2,n}/m_{1,n} \to 0$, where the limiting in the convergences on the last line follow from (C.86).

This completes all the cases.

Finally, observe that for Algorithm 2, when $m_{2,n} = \lfloor n^\nu \rfloor$ for some $\nu \in (0, 1)$ and $m_{1,n} = n_{\text{tr}}$ such that $n_{\text{tr}}/n \to 1$ as $n \to \infty$, $p_n/m_{1,n} \to \gamma \in (0, \infty)$, and $m_{2,n}/m_{1,n} \to 0$, and hence the statement follows.

□

Lemma C.6.2 (Verifying space-filling property of the discrete grid used in the one-step procedure)**.** *Let* $\{p_n\}$*,* $\{m_{1,n}\}$*,* $\{m_{2,n}\}$ *are three sequences of positive integers such that* $m_{2,n} \leq m_{1,n}$ *for* $n \geq 1$*, and* $n \to \infty$*,*

$$\frac{p_n}{m_{1,n}} \to \gamma \in (0, \infty) \quad \textit{and} \quad \frac{m_{2,n}}{m_{1,n}} \to 0.$$

Define a sequence of grids $\mathcal{G}_n$ as follows:

$$\mathcal{G}_n := \left\{ \left(\frac{p_n}{m_{1,n} - k_1 m_{2,n}}, \frac{p_n}{k_2 m_{2,n}} \right) : k_1 \in \left\{ 2, \dots, \left\lceil \frac{m_{1,n}}{m_{2,n}} - 2 \right\rceil \right\}, k_2 \in \{0, \dots, k_1 - 1\} \right\}.$$

Let $\zeta_1^\star$ and $\zeta_2^\star$ be two non-negative real numbers such that

$$\frac{1}{\zeta_1^\star} + \frac{1}{\zeta_2^\star} \le \frac{1}{\gamma}.$$

Let $\Pi_{\mathcal{G}_n}(\zeta_1^\star, \zeta_2^\star) = (\pi_{1,n}, \pi_{2,n})$ denote the projection of the point $(\zeta_1^\star, \zeta_2^\star)$ on the grid $\mathcal{G}_n$ with respect to the ℓ_1 distance. Then, $\pi_{1,n} \to \zeta_1^\star$ and $\pi_{2,n} \to \zeta_2^\star$ as $n \to \infty$. In particular, in the context of Algorithm 3, taking $m_{1,n} = n_{\mathrm{tr}}$, $m_{2,n} = \lfloor n^\nu \rfloor$ for some $\nu \in (0,1)$, we get the aspect ratios used in Algorithm 3 "converge" to the set $\{(\zeta_1, \zeta_2) : \zeta_1^{-1} + \zeta_2^{-1} \le \gamma^{-1}\}$ when $n_{\mathrm{tr}}/n \to 1$ under (PA(γ)).

Proof. The proof follows the general strategy employed in the proof Lemma C.6.1 and uses the result as ingredient.

Fix any point $(\zeta_1^\star, \zeta_2^\star)$ that satisfies the constraint

$$\frac{1}{\zeta_1^\star} + \frac{1}{\zeta_2^\star} \le \frac{1}{\gamma}.$$

We will construct a pair $(g_1^\star, g_2^\star)$ in the grid $\mathcal{G}_n$ such that $(g_1^\star, g_2^\star) \to (\zeta_1^\star, \zeta_2^\star)$. Because

$$\|\Pi_{\mathcal{G}_n}(\zeta_1^\star, \zeta_2^\star) - (\zeta_1^\star, \zeta_2^\star)\|_{\ell_1} \le \|(g_1^\star, g_2^\star) - (\zeta_1^\star, \zeta_2^\star)\|_{\ell_1},$$

such a choice shows the desired result.

Define

$$(k_1^\star, k_2^\star) = \left(\left\lceil \frac{m_{1,n} - p_n/\zeta_1^\star}{m_{2,n}} \right\rceil, \left\lfloor \frac{p_n/\zeta_2^\star}{m_{2,n}} \right\rfloor \right), \quad \text{and} \quad (g_1^\star, g_2^\star) = \left(\frac{p}{m_{1,n} - k_1^\star m_{2,n}}, \frac{p}{k_2^\star m_{2,n}} \right).$$

By appealing to Lemma C.6.1, it follows that $\pi_{1,n} \to \zeta_1^\star$ as $n \to \infty$. Note that the value of $k_1^\star$ is exactly the right point of the grid interval in Figure C.3 in the proof of Lemma C.6.1. Since $\zeta_1^\star \in [\gamma, \infty]$ and the first coordinate of the grid $\mathcal{G}_n$ is the same as that in Lemma C.6.1, we have that $g_1^\star$ is a feasible choice and $g_1^\star \to \zeta_1^\star$. It remains to verify the conditions for $g_2^\star$.

Note that when $\zeta_2^\star = \infty$, $k_2^\star = 0$, which satisfies the desired condition. Assume that $\zeta_2^\star < \infty$. We verify below that $k_2^\star < k_1^\star$ so that $k_2^\star$ is a feasible choice and that

$$\frac{k_2^\star m_{2,n}}{p_n} \to \frac{1}{\zeta_2^\star},$$

which implies the desired convergence of the reciprocal.

Observe that

$$k_2^\star \le \frac{p_n}{\zeta_2^\star m_{2,n}} \le \frac{p_n}{m_{2,n}} \left(\frac{m_{1,n}}{p_n} - \frac{1}{\zeta_1^\star} \right) \le \frac{m_{1,n} - p_n/\zeta_1^\star}{m_{2,n}} = k_1^\star.$$

This verifies the first condition. For the second part, consider

$$0 \le \left| \frac{k_2^\star m_{2,n}}{p_n} - \frac{1}{\zeta_2^\star} \right| = \left| \left\lfloor \frac{p_n/\zeta_2^\star}{m_{2,n}} \right\rfloor \frac{m_{2,n}}{p_n} - \frac{1}{\zeta_2^\star} \right| \le \frac{m_{2,n}}{p_n} \to 0$$

under (PA(γ)) as $n \to \infty$.

Finally, note that for Algorithm 3, when $m_{2,n} = \lfloor n^\nu \rfloor$ for some $\nu \in (0,1)$ and $m_{1,n} = n_{\mathrm{tr}}$ such that $n_{\mathrm{tr}}/n \to 1$ as $n \to \infty$, $p_n/m_{1,n} \to \gamma \in (0,\infty)$, and $m_{2,n}/m_{1,n} \to 0$, and therefore the statement follows. □

C.6.2 Lemmas for restricting arbitrary sequences to specific convergent sequences

In this section, we collect supplementary lemmas that are used in the proofs of Lemmas 3.3.8 and 3.4.1 in Appendices C.2 and C.4, respectively.

Lemma C.6.3 (From subsequence convergence to sequence convergence)**.** *Let $\{a_m\}_{m\geq 1}$ be a sequence in $\mathbb{R}$. Suppose for any subsequence $\{a_{m_k}\}_{k\geq 1}$, there is a further subsequence $\{a_{m_{k_l}}\}_{l\geq 1}$ such that $\lim_{m\to\infty} a_{m_{k_l}} = 0$. Then $\lim_{m\to\infty} a_m = 0$.*

Proof. Let $\alpha := \limsup_{m\to\infty} a_m$ and $\beta := \liminf_{m\to\infty} a_m$. This means that there is subsequence $\{a_{m_k}\}_{k\geq 1}$ such that $\lim_{m\to\infty} a_{m_k} = \alpha$. Similarly, there is a (different) subsequence $\{a_{m_l}\}_{l\geq 1}$ such that $\lim_{m\to\infty} a_{m_l} = \beta$. But since every converging sequence has a further subsequence that converges to the same limit, the lemma follows. □

Lemma C.6.4 (Limit of minimization over finite grids in a metric space)**.** *Let (M, d) be a metric space, and C be a subset of M. Suppose $h : M \to \mathbb{R}$ is a function that attains its infimum over C at $\zeta^\star$. Let $\mathcal{G}$ be a finite set of points in C. Then, the following inequalities hold:*

$$0 \leq \min_{x\in\mathcal{G}} h(x) - \inf_{x\in C} h(x) \leq h(\Pi_{\mathcal{G}}(\zeta^\star)) - h(\zeta^\star), \tag{C.87}$$

where $\Pi_{\mathcal{G}}(y) = \arg\min_{x\in\mathcal{G}} d(x, y)$ is the point in the grid closest to y. Consequently, if $\mathcal{G}_n$ is a sequence of grids such that $\Pi_{\mathcal{G}_n}(\zeta^\star) \to \zeta^\star$, and $h(\cdot)$ is continuous at $\zeta^\star$, then

$$\min_{x\in\mathcal{G}_n} h(x) - \inf_{x\in C} h(x) \to 0. \tag{C.88}$$

Proof. Since $\mathcal{G} \subseteq C$ and $\Pi_{\mathcal{G}}(\zeta^\star) \in \mathcal{G}$, we have the following chain of inequalities:

$$h(\zeta^\star) = \inf_{x\in C} h(x) \leq \min_{x\in\mathcal{G}} h(x) \leq h(\Pi_{\mathcal{G}}(\zeta^\star)).$$

Subtracting $h(\zeta^\star)$ throughout, we get the desired result (C.87). In addition, if $\mathcal{G}_n$ is a sequence of grids such that $\Pi_{\mathcal{G}}(\zeta^\star) \to \zeta^\star$, then continuity of $h(\cdot)$ at $\zeta^\star$ implies $h(\Pi_{\mathcal{G}}(\zeta^\star)) \to h(\zeta^\star)$ leading to (C.88). □

Lemma C.6.5 (Limit points of argmin sequence over space-filling grids)**.** *Let (M, d) be a metric space and C be a compact subset of M. Let $\mathcal{G}_n$ be a sequence of grids such that for any $\zeta \in C$, $\Pi_{\mathcal{G}_n}(\zeta) \to \zeta$ as $n \to \infty$ where $\Pi_{\mathcal{G}_n}(y) = \arg\min_{x\in\mathcal{G}_n} d(x, y)$ is the point in the grid $\mathcal{G}_n$ closest to y. Let $h : C \to [0, \infty]$ be a lower semicontinuous function, and let $x_n \in \arg\min_{x\in\mathcal{G}_n} h(x)$. Then, for any arbitrary subsequence $\{x_{n_k}\}_{k\geq 1}$ of $\{x_n\}_{n\geq 1}$, there exists a further subsequence $\{x_{n_{k_l}}\}_{l\geq 1}$ such that $x_{n_{k_l}}$ converges to a point in $\arg\min_{\zeta\in C} h(\zeta)$ as $l \to \infty$.*

Proof. Because h is lower semicontinuous and C is compact, h attains its minimum on C (see, e.g., Section 1.6 of Pedersen (2012) and also see Theorem 1.9 of Rockafellar and Wets (2009) with the domain $\mathbb{R}^n$ replaced with any metric space.). Let $\mathcal{M} = \arg\min_{\zeta\in C} h(\zeta)$, which is non-empty. Because C is compact, for any arbitrary subsequence $\{x_{n_k}\}_{k\geq 1}$, there is a further subsequence $\{x_{n_{k_l}}\}_{l\geq 1}$ that converges to some point $p \in C$. Lower semicontinuity of h now implies that

$$\liminf_{l\to\infty} h(x_{n_{k_l}}) \geq h(p). \tag{C.89}$$

See, e.g., Section 1.5 of Pedersen (2012). By definition, $h(x_{n_{k_l}}) = \min_{x\in\mathcal{G}_{n_{k_l}}} h(x)$ and because $\Pi_{\mathcal{G}_{n_{k_l}}}(\zeta) \to \zeta$ for any $\zeta \in C$, Lemma C.6.4 implies that

$$\lim_{l\to\infty} h(x_{n_{k_l}}) \;=\; \min_{\zeta\in C} h(\zeta).$$

Combined with (C.89), we conclude that $h(p) = \min_{\zeta\in C} h(\zeta)$, and hence $p \in \mathcal{M} = \arg\min_{\zeta\in C} h(\zeta)$. □

C.6.3 Lemmas for certifying continuity from continuous convergence

In this section, we collect supplementary lemmas that are used in the proofs of Propositions 3.3.10 and 3.4.3 in Appendix C.2 and Appendix C.4, respectively.

Lemma C.6.6 (Deterministic functions; see, e.g., Problem 57, Chapter 4 of Pugh (2002), converse of Theorem 21.3 in Munkres (2000))**.** *Suppose f_n and f are (deterministic) functions from $I \subseteq \mathbb{R}$ to $\mathbb{R}$. For any $x \in I$ and any arbitrary sequence $\{x_n\}_{n\geq 1}$ in I for which $x_n \to x$, assume that $f_n(x_n) \to f(x)$ as $n \to \infty$. Then, f is continuous on I.*

Proof. The following is a standard proof by contradiction. Assume f is discontinuous at $a \in I$. Then, there exists a sequence $x_n \to a$ such that

$$f(x_n) \notin [f(a) - 2\epsilon, f(a) + 2\epsilon]$$

for some $\epsilon > 0$. Note that $f_n(x) \to f(x)$ for all $x \in I$. Now, consider another sequence y_n such that

$$\begin{aligned} y_1 = y_2 = \cdots = y_{N_1} = x_1, &\quad \text{where} \quad |f_{N_1}(x_1) - f(a)| > \epsilon \\ y_{N_1+1} = y_{N_1+2} = \cdots = y_{N_2} = x_2, &\quad \text{where} \quad |f_{N_2}(x_2) - f(a)| > \epsilon, N_2 > N_1 \\ &\vdots \end{aligned}$$

Observe that $y_n \to a$, however $f_n(y_n) \not\to f(a)$. Hence, a contradiction. □

Lemma C.6.7 (Extension of Lemma C.6.6 to random functions)**.** *Suppose f_n is a sequence of random real-valued functions from $I \subseteq \mathbb{R}$ such that, for every deterministic sequence $\{x_n\}_{n\geq 1}$ in I such that $x_n \to x \in I$, $f_n(x_n) \to f(x)$ in probability, for a deterministic function f on I. Then, f is continuous on I.*

Proof. The idea of the proof is similar to that of an analogous statement for fixed functions; see Lemma C.6.6. We will use proof by contradiction. Assume that f is discontinuous at $a \in I$. Then, as in the proof of Lemma C.6.6 for deterministic functions, there exists a $\epsilon > 0$ and a sequence $\{x_n\} \subset I$ such that $x_n \to a$ and

$$f(x_n) \notin [f(a) - 2\epsilon, f(a) + 2\epsilon]. \tag{C.90}$$

From the hypothesis, we have that, for each $x \in I$, $f_n(x) \to f(x)$ in probability. Let $p \in (0, 1)$ be a fixed number. Then, there exists an integer $N_1 \geq 1$ such that the event

$$\Omega_{N_1} = \{|f_{N_1}(x_1) - f(x_1)| < \epsilon\}$$

holds with probability at least p. Thus, on Ω_{N_1}, by the triangle inequality,

$$|f_{N_1}(x_1) - f(a)| \geq |f(x_1) - f(a)| - |f_{N_1}(x_1) - f(x_1)| > \epsilon, \tag{C.91}$$

where last inequality stems from (C.90). Next, for $i = 2, 3, \ldots$, let $N_i \geq N_{i-1} + 1$ be an integer such that the event

$$\Omega_{N_i} = \{|f_{N_i}(x_i) - f(x_i)| < \epsilon\}$$

has probability at least p. These sequences of numbers $\{N_i\}$ and events $\{\Omega_{N_i}\}$ exist because, by hypothesis, $f_n(x_i) \to f(x_i)$ in probability for each i. Furthermore $N_i \to \infty$ and, on each Ω_{N_i}, $|f_{N_i}(x_i) - f(a)| > \epsilon$ by the same argument used in (C.91).

Consider the sequence $\{y_n\}$ given by

$$\begin{aligned} y_1 = y_2 = \cdots = y_{N_1} &= x_1 \\ y_{N_1+1} = y_{N_1+2} = \cdots = y_{N_2} &= x_2 \\ &\vdots \end{aligned}$$

such that, by construction, $y_n \to a$. We will derive a contradiction by showing that it cannot be the case that $f_n(y_n) \to a$ in probability, thus violating the hypothesis. Indeed, the sequence of probability values $\{\mathbb{P}(|f_n(y_n) - f(a)| > \epsilon)\}$ does not converge to zero since, for each n, there exist infinitely many $N_i > n$ such that

$$\mathbb{P}(|f_{N_i}(y_{N_i}) - f(a)| > \epsilon) \geq \mathbb{P}(\Omega_{N_i}) > p > 0.$$

Thus, it must be the case that f is continuous at a. Continuity of f over I readily follows.

□

C.6.4 A lemma for lifting $\mathbb{Q}$-continuity to $\mathbb{R}$-continuity

The following lemma is used in the proofs of Propositions 3.3.10 and 3.4.3 in Appendices C.2 and C.4, respectively.

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point $x_\infty \in \mathbb{R}$, if for all sequences $\{x_n\}_{n\geq 1}$ in $\mathbb{R}$ for which $x_n \to x_\infty$ as $n \to \infty$, we have $f(x_n) \to f(x_\infty)$ as $n \to \infty$. Call this $\mathbb{R}$-continuity of f at the point x_∞, and call a function is $\mathbb{R}$-continuous if it is $\mathbb{R}$-continuous on its domain. Define a variant of continuity with respect to rational sequences, dubbed $\mathbb{Q}$-continuity, as follows.

Definition C.6.8 ($\mathbb{Q}$-continuity)**.** A function $f : \mathbb{R} \to \mathbb{R}$ is $\mathbb{Q}$-continuous at a point $x_\infty \in \mathbb{R}$, if for all sequences $\{x_n\}_{n\geq 1}$ in $\mathbb{Q}$ for which $x_n \to x_\infty$ as $n \to \infty$, we have $f(x_n) \to f(x_\infty)$ as $n \to \infty$. A function is $\mathbb{Q}$-continuous if it is $\mathbb{Q}$-continuous over its domain.

The following lemma shows that $\mathbb{Q}$-continuity implies $\mathbb{R}$-continuity.

Lemma C.6.9 ($\mathbb{Q}$-continuity implies $\mathbb{R}$-continuity)**.** *Suppose $f : \mathbb{R} \to \mathbb{R}$ is a $\mathbb{Q}$ continuous function. Then f is $\mathbb{R}$-continuous.*

Proof. To prove $\mathbb{R}$-continuity of f, fix any $y_\infty \in \mathbb{R}$, and consider any arbitrary sequence $\{y_n\}_{n\geq 1}$ in $\mathbb{R}$ such that $y_n \to y_\infty$ as $n \to \infty$. For any $\epsilon > 0$, if we can produce n_ϵ such that $|f(y_n) - f(y_\infty)| \leq \epsilon$ for all $n \geq n_\epsilon$, then $\mathbb{R}$-continuity of f follows. We will produce such n_ϵ below.

For every $m \geq 1$, construct a sequence $\{x_{k,m}\}_{k\geq 1}$ in $\mathbb{Q}$ such that $x_{k,m} \to y_m$ as $k \to \infty$; see Figure C.4. (Note this is possible because $\mathbb{Q}$ is dense in $\mathbb{R}$.) Now, for every $m \geq 1$, using $\mathbb{Q}$-continuity of f at y_m, we have $f(x_{k,m}) \to f(y_m)$ as $k \to \infty$. Fix $\epsilon > 0$. Let $k_0(\epsilon) = 1$ and for $m \geq 1$, define a positive integer $k_m(\epsilon)$ by

$$k_m(\epsilon) = \min\{k > k_{m-1}(\epsilon) : |f(x_{k,m}) - f(y_m)| \leq \epsilon/2\}.$$

Such a $k_m(\epsilon)$ always exists because $x_{k,m} \to y_m$ as $k \to \infty$ and f is $\mathbb{Q}$-continuous at y_m. Note that $k_m(\epsilon) > k_{m-1}(\epsilon)$, which in turn implies that $k_m(\epsilon) \geq m$ and thus $k_m(\epsilon) \to \infty$ as $m \to \infty$. Hence, as $m \to \infty$, $x_{k_m(\epsilon),m} \to y_\infty$. Using the $\mathbb{Q}$-continuity of f at y_∞, there exists a positive integer m_ϵ such that for all $m \geq m_\epsilon$, we have $|f(x_{k_m(\epsilon),m}) - f(y_\infty)| \leq \epsilon/2$. For all $m \geq m_\epsilon$, by the triangle inequality, observe that

$$|f(y_m) - f(y_\infty)| \leq |f(y_m) - f(k_m(\epsilon))| + |f(k_m(\epsilon)) - f(y_\infty)| \leq \epsilon.$$

Therefore, choosing $n_\epsilon = m_\epsilon$ completes the proof.

□

C.6.5 Lemmas on asymptotic deterministic equivalents for generalized bias and variance resolvents

In this section, we collect lemmas on asymptotic deterministic equivalents for generalized bias and variance resolvents associated with ridge and ridgeless regression that are used in the proof of Proposition 3.3.14 in Appendix C.3, and Proposition 3.4.10 and Lemma 3.4.8 in Appendix C.5.

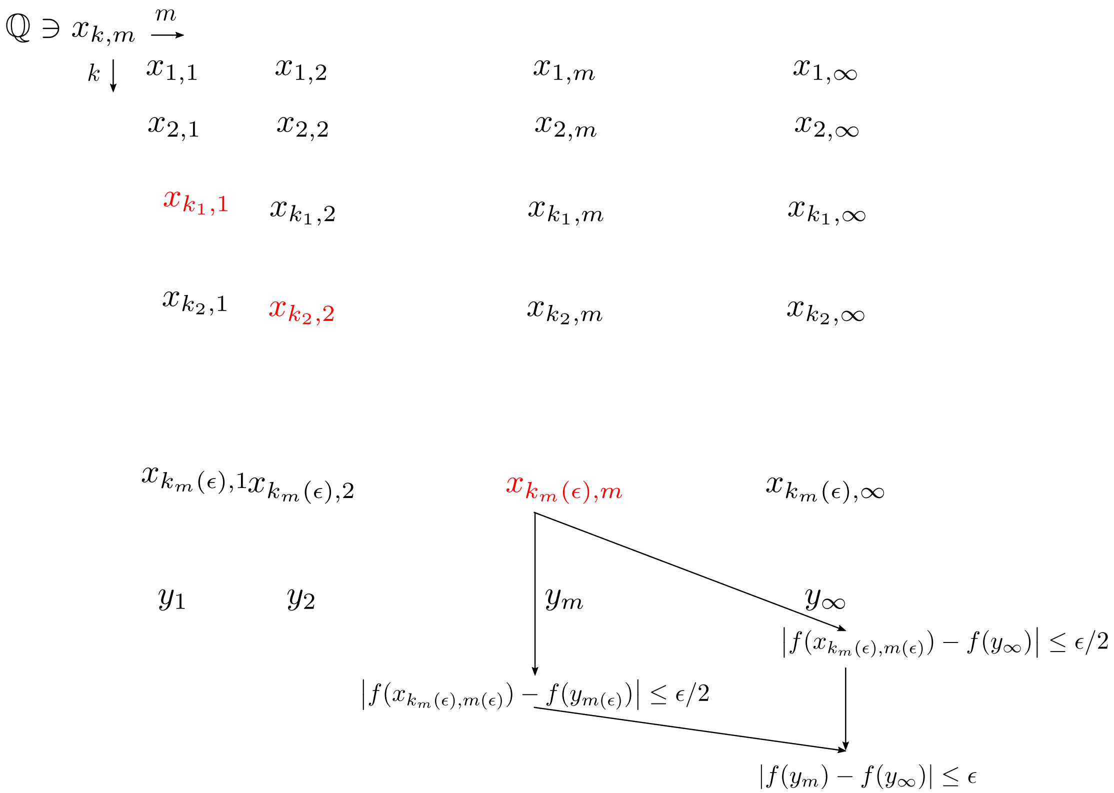

$\mathbb{Q} \ni x_{k,m} \xrightarrow{m}$
$k \downarrow$
$x_{1,1}$ $x_{1,2}$ $x_{1,m}$ $x_{1,\infty}$
$x_{2,1}$ $x_{2,2}$ $x_{2,m}$ $x_{2,\infty}$
$x_{k_1,1}$ $x_{k_1,2}$ $x_{k_1,m}$ $x_{k_1,\infty}$
$x_{k_2,1}$ $x_{k_2,2}$ $x_{k_2,m}$ $x_{k_2,\infty}$
$x_{k_m(\epsilon),1}x_{k_m(\epsilon),2}$ $x_{k_m(\epsilon),m}$ $x_{k_m(\epsilon),\infty}$
y_1 y_2 y_m y_∞
$\left|f(x_{k_m(\epsilon),m(\epsilon)}) - f(y_\infty)\right| \le \epsilon/2$
$\left|f(x_{k_m(\epsilon),m(\epsilon)}) - f(y_{m(\epsilon)})\right| \le \epsilon/2$
$|f(y_m) - f(y_\infty)| \le \epsilon$

Figure C.4: Illustration of the grid of rational sequences used in the proof of Lemma C.6.9.

Lemma C.6.10 (Deterministic equivalents for generalized bias and variance ridge resolvents)**.** *Suppose $X_i \in \mathbb{R}^p$, $1 \le i \le n$, are i.i.d. random vectors with each $X_i = Z_i \Sigma^{1/2}$, where $Z_i \in \mathbb{R}^p$ contains i.i.d. random variables Z_{ij}, $1 \le j \le p$, each with $\mathbb{E}[Z_{ij}] = 0$, $\mathbb{E}[Z_{ij}^2] = 1$, and $\mathbb{E}[|Z_{ij}|^{8+\alpha}] \le M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\Sigma \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix such that $r_{\min} I_p \preceq \Sigma \preceq r_{\max} I_p$ for some constants $r_{\min} > 0$ and $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the random matrix with X_i, $1 \le i \le n$, as its rows and let $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p \times p}$ denote the $p \times p$ random matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$. Let $A \in \mathbb{R}^{p \times p}$ be any deterministic positive semidefinite matrix that commutes with Σ such that $a_{\min} I_p \preceq A \preceq a_{\max} I_p$ for some constants $a_{\min} > 0$ and $a_{\max} < \infty$ (independent of p). Let $\gamma_n := p/n$. Then, for $\lambda > 0$, as $n, p \to \infty$ with $0 < \liminf \gamma_n \le \limsup \gamma_n < \infty$, the following asymptotic deterministic equivalences hold:*

1. *Generalized variance of ridge regression:*
$$(\widehat{\boldsymbol{\Sigma}} + \lambda I_p)^{-2} \widehat{\boldsymbol{\Sigma}} A \simeq \widetilde{v}(-\lambda; \gamma_n)(v(-\lambda; \gamma_n)\Sigma + I_p)^{-2} \Sigma A, \tag{C.92}$$
where $v(-\lambda; \gamma_n) \ge 0$ is the unique solution to the fixed-point equation
$$v(-\lambda; \gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma (v(-\lambda; \gamma_n)\Sigma + I_p)^{-1}]/p, \tag{C.93}$$
and $\widetilde{v}(-\lambda; \gamma_n)$ is defined via $v(-\lambda; \gamma_n)$ by the equation
$$\widetilde{v}(-\lambda; \gamma_n)^{-1} = v(-\lambda; \gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2 (v(-\lambda; \gamma_n)\Sigma + I_p)^{-2}]/p. \tag{C.94}$$

2. *Generalized bias of ridge regression:*
$$\lambda^2 (\widehat{\boldsymbol{\Sigma}} + \lambda I_p)^{-1} A (\widehat{\boldsymbol{\Sigma}} + \lambda I_p)^{-1} \simeq (v(-\lambda; \gamma_n)\Sigma + I_p)^{-1} (\widetilde{v}_g(-\lambda; \gamma_n)\Sigma + A)(v(-\lambda; \gamma_n)\Sigma + I_p)^{-1}, \tag{C.95}$$
where $v(-\lambda; \gamma_n)$ as defined in (C.98), and $\widetilde{v}_g(-\lambda; \gamma_n)$ is defined via $v(-\lambda; \gamma_n)$ by the equation
$$\widetilde{v}_g(-\lambda; \gamma_n) = \frac{\gamma_n \operatorname{tr}[A \Sigma (v(-\lambda; \gamma_n)\Sigma + I_p)^{-2}]/p}{v(-\lambda; \gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2 (v(-\lambda; \gamma_n)\Sigma + I_p)^{-2}]/p}. \tag{C.96}$$

Proof. The main idea for both the first and second parts is to use Corollary C.7.4 as the starting point, and apply the calculus rules for asymptotic deterministic equivalents listed in Appendix C.7 to manipulate into the desired equivalents.

Part 1. For the first part, observe that we can express the resolvent of interest (associated with the generalized variance of ridge regression) as a derivative (with respect to λ) of a certain resolvent:

$$(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}A = (\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A - \lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}A = \frac{\partial}{\partial\lambda}[\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A)]. \tag{C.97}$$

To find a deterministic equivalent for $(\widehat{\mathbf{\Sigma}}+\lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}A$, it thus suffices to obtain a deterministic equivalent for the resolvent $\lambda(\widehat{\mathbf{\Sigma}}+\lambda I_p)^{-1}A$ and take its derivative, thanks to the differentiation rule from Lemma C.7.2 (5). Similar derivative trick is used in the proof of Theorem 2.1 in Liu and Dobriban (2019) and Theorem 2.1 in Dobriban and Wager (2018) to compute the standard variance of ridge regression, by Dobriban and Sheng (2020) in the context of distributed ridge regression, and in the earlier works by Karoui and Kösters (2011); Rubio and Mestre (2011); Ledoit and Péché (2011), among others, to compute certain limiting trace functionals.

Starting with Corollary C.7.4, we have

$$\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is the unique solution to the fixed point equation

$$v(-\lambda;\gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}]/p. \tag{C.98}$$

Since A has bounded operator norm (uniformly in p), from Lemma C.7.2 (3), we have

$$\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}A, \tag{C.99}$$

where $v(-\lambda;\gamma_n)$ is as defined by (C.98). It now remains to take the derivative of the right hand side of (C.99) with respect to λ. Before doing so, we will briefly argue that the differentiation rule indeed applies in this case. Let $T \in \mathbb{R}^{p\times p}$ be a matrix with trace norm uniformly bounded in p. Note that

$$\begin{aligned}
\operatorname{tr}[T\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A] &= \operatorname{tr}[T(I_p - \widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1})A] \\
&\le \|(I_p - \widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1})A\|_{\mathrm{op}} \operatorname{tr}[T] \\
&\le \|I_p - \widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}\|_{\mathrm{op}} \|A\|_{\mathrm{op}} \operatorname{tr}[T] \\
&\le \|A\|_{\mathrm{op}} \operatorname{tr}[T] \le C,
\end{aligned}$$

for some constant $C < \infty$. Here, the first inequality follows from Proposition 3.4.10 of Pedersen (2012) (see also, Problem III.6.2 of Bhatia (1997)), and the second inequality follows from the submultiplicativity of the operator norm. Similarly, note that

$$\operatorname{tr}[T(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}A] \le \|(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}\|_{\mathrm{op}}\|A\|_{\mathrm{op}} \operatorname{tr}[T] \le C,$$

for some constant $C < \infty$. Thus, we can safely apply the differentiation rule from Lemma C.7.2 (5) to get

$$(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}A \simeq \frac{\partial}{\partial\lambda}[(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}A].$$

Taking derivative, we have

$$\frac{\partial}{\partial\lambda}[(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}A] = -\frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)](v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma A. \tag{C.100}$$

We can write - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$ in terms of $v(-\lambda;\gamma_n)$ by taking derivative of (C.98) with respect to λ and solving for - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$. Taking the derivative of (C.98) yields the following equation:

$$-\frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)]v(-\lambda;\gamma_n)^{-2} = 1 + \gamma_n - \frac{\partial}{\partial\lambda}[v(-\lambda;\gamma_n)] \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}]/p. \tag{C.101}$$

Denoting - $\partial/\partial\lambda[v(-\lambda;\gamma_n)]$ by $\widetilde{v}(-\lambda;\gamma_n)$ and solving for $\widetilde{v}(-\lambda;\gamma_n)$ in (C.101), we get

$$\widetilde{v}(-\lambda;\gamma_n)^{-1} = v(-\lambda;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}]/p. \tag{C.102}$$

Combining (C.97), (C.100), and (C.102), the statement follows. This completes the proof of the first part.

Part 2. For the second part, observe that we can express the resolvent of interest (appearing in the generalized bias of ridge regression) as a derivative of a certain parameterized resolvent at a fixed value of the parameter:

$$\begin{aligned}\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1} &= \lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1}|_{\rho=0} \\ &= -\frac{\partial}{\partial\rho}[\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1}]\,\Big|_{\rho=0}\,.\end{aligned} \tag{C.103}$$

It is worth remarking that in contrast to Part 1, we needed to introduce another parameter ρ for this part to appropriately pull out the matrix A in the middle. This trick has been used in the proof of Theorem 5 in Hastie et al. (2022) in the context of standard bias calculation for ridge regression. Our strategy henceforth will be to obtain a deterministic equivalent for the resolvent $\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1}$, take its derivative with respect to ρ, and set $\rho = 0$. Towards that end, we first massage it to make it amenable for application of Lemma C.7.3 as follows:

$$\begin{aligned}\lambda\big(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A\big)^{-1} &= \lambda\big(\widehat{\mathbf{\Sigma}} + \lambda(I_p + \rho A)\big)^{-1} \\ &= (I_p + \rho A)^{-1/2}\lambda\big((I_p + \rho A)^{-1/2}\widehat{\mathbf{\Sigma}}(I_p + \rho\Sigma)^{-1/2} + \lambda I_p\big)^{-1}(I_p + \rho A)^{-1/2} \\ &= (I_p + \rho A)^{-1/2}\lambda\big(\widehat{\mathbf{\Sigma}}_{\rho,A} + \lambda I_p\big)^{-1}(I_p + \rho A)^{-1/2},\end{aligned} \tag{C.104}$$

where $\widehat{\mathbf{\Sigma}}_{\rho,A} := \Sigma_{\rho,A}^{1/2}(\boldsymbol{Z}^\top\boldsymbol{Z}/n)\Sigma_{\rho,A}^{1/2}$ and $\Sigma_{\rho,A} := (I_p + \rho A)^{-1/2}\Sigma(I_p + \rho A)^{-1/2}$. We will now obtain a deterministic equivalent for $\lambda(\widehat{\mathbf{\Sigma}}_{\rho,A} + \lambda I_p)^{-1}$, and use the product rule to arrive at the deterministic equivalent for $\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1}$.

Using Corollary C.7.4, we have

$$\lambda(\widehat{\mathbf{\Sigma}}_{\rho,A} + \lambda I_p)^{-1} \simeq (v_g(-\lambda,\rho;\gamma_n)\Sigma_{\rho,A} + I_p)^{-1}, \tag{C.105}$$

where $v_g(-\lambda,\rho;\gamma_n)$ is the unique solution to the fixed-point equation

$$v_g(-\lambda,\rho;\gamma_n)^{-1} = \lambda + \gamma_n \operatorname{tr}[\Sigma_{\rho,A}(v_g(-\lambda,\rho;\gamma_n)\Sigma_{\rho,A} + I_p)^{-1}]/p. \tag{C.106}$$

Combining (C.104) with (C.105), and using the product rule from Lemma C.7.2 (3) (which is applicable since $(I_p + \rho A)^{-1/2}$ is a deterministic matrix), we get

$$\begin{aligned}\lambda(\widehat{\mathbf{\Sigma}} + \lambda I_p + \lambda\rho A)^{-1} &= (I_p + \rho A)^{-1/2}\lambda(\widehat{\mathbf{\Sigma}}_{\rho,A} + \lambda I_p)^{-1}(I_p + \rho A)^{-1/2} \\ &\simeq (I_p + \rho A)^{-1/2}(v_g(-\lambda,\rho;\gamma_n)\Sigma_{\rho,A} + I_p)^{-1}(I_p + \rho A)^{-1/2} \\ &= (I_p + \rho A)^{-1/2}(v_g(-\lambda,\rho;\gamma_n)(I_p + \rho A)^{-1/2}\Sigma(I_p + \rho A)^{-1/2} + I_p)^{-1}(I_p + \rho A)^{-1/2} \\ &= (v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}.\end{aligned}$$

Similarly, the right hand side of the fixed-point equation (C.106) can be simplified by substituting back for $\Sigma_{\rho,A}$ to yield

$$\begin{aligned}v_g(-\lambda,\rho;\gamma_n)^{-1} &= \lambda + \gamma_n \operatorname{tr}[(I_p + \rho A)^{-1/2}\Sigma(I_p + \rho A)^{-1/2}(v_g(-\lambda,\rho;\gamma_n)\Sigma_{\rho,A} + I_p)^{-1}]/p \\ &= \lambda + \gamma_n \operatorname{tr}[\Sigma(v_g(-\lambda,\rho;\gamma_n)(I_p + \rho A)^{1/2}\Sigma_{\rho,A}(I_p + \rho A)^{1/2} + (I_p + \rho A))^{-1}]/p \\ &= \lambda + \gamma_n \operatorname{tr}[\Sigma(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}]/p.\end{aligned} \tag{C.107}$$

Finally, we will now use the differentiation rule from Lemma C.7.2 (5) (with respect to ρ this time). The applicability of the differentiation rule follows analogously to first part for $\rho > -1/a_{\min}$. Additionally, it is easy to verify that both sides of (C.107) are analytic in ρ. Taking derivative with respect to ρ, we get

$$\begin{aligned} &-\frac{\partial}{\partial\rho}[(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}] \\ &\quad = (v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}\left(\frac{\partial}{\partial\rho}[v_g(-\lambda,\rho;\gamma_n)]\Sigma + A\right)(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}. \end{aligned} \tag{C.108}$$

Setting $\rho = 0$ and observing that $v_g(-\lambda, 0; \gamma_n) = v(-\lambda;\gamma_n)$, where $v(-\lambda;\gamma_n)$ is as defined in (C.98), we have

$$\begin{aligned} &\frac{\partial}{\partial\rho}[(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-1}]\ \Big|_{\rho=0} \\ &\quad = (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}\left(\frac{\partial}{\partial\rho}[v_g(-\lambda,\rho;\gamma_n)]\ \Big|_{\rho=0}\ \Sigma + A\right)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}. \end{aligned} \tag{C.109}$$

To obtain an equation for $\partial/\partial\rho[v_g(-\lambda,\rho;\gamma_n)]|_{\rho=0}$, we can differentiate the fixed-point equation (C.107) with respect to ρ to yield

$$\begin{aligned} &-\frac{\partial}{\partial\rho}[v_g(-\lambda,\rho;\gamma_n)]v_g(-\lambda,\rho;\gamma_n)^{-2} \\ &\quad = -\gamma_n\frac{\partial}{\partial\rho}[v_g(-\lambda,\rho;\gamma_n)]\operatorname{tr}[\Sigma^2(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-2}]/p \\ &\qquad - \gamma_n \operatorname{tr}[A\Sigma(v_g(-\lambda,\rho;\gamma_n)\Sigma + I_p + \rho A)^{-2}]/p. \end{aligned}$$

Setting $\rho = 0$ in the equation above, and using the fact that $v_g(-\lambda, 0; \gamma_n) = v(-\lambda;\gamma_n)$, and denoting $\partial/\partial\rho[v_g(-\lambda,\rho;\gamma_n)]|_{\rho=0}$ by $\widetilde{v}_g(-\lambda;\gamma_n)$, we get that

$$\widetilde{v}_g(-\lambda;\gamma_n) = \frac{\gamma_n \operatorname{tr}[A\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}]/p}{v(-\lambda;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}]/p}. \tag{C.110}$$

Therefore, from (C.103) and (C.109), we finally have

$$\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(-\lambda;\gamma_n)\Sigma + A)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is as defined in (C.98), and $\widetilde{v}_g(-\lambda;\gamma_n)$ is as defined in (C.110). This completes the proof of the second part.

□

Lemma C.6.11 (Deterministic equivalents for generalized bias and variance ridgeless resolvents)**.** *Assume the setting of Lemma C.6.10 with $\gamma_n \in (1,\infty)$. Then, the following deterministic equivalences hold:*

1. *Generalized variance of ridgeless regression:*

$$\widehat{\mathbf{\Sigma}}^{+}A \simeq \widetilde{v}(0;\gamma_n)(v(0;\gamma_n)\Sigma + I_p)^{-2}\Sigma A, \tag{C.111}$$

where $v(0;\gamma_n)$ is the unique solution to the fixed-point equation

$$\gamma_n^{-1} = \operatorname{tr}[v(0;\gamma_n)\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-1}]/p, \tag{C.112}$$

and $\widetilde{v}(0;\gamma_n)$ is defined through $v(0;\gamma_n)$ via

$$\widetilde{v}(0;\gamma_n) = \left(v(0;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(0;\gamma_n)\Sigma + I_p)^{-2}]/p\right)^{-1}. \tag{C.113}$$

2. *Generalized bias of ridgeless regression:*

$$(I_p - \widehat{\mathbf{\Sigma}}^{+}\widehat{\mathbf{\Sigma}})A(I_p - \widehat{\mathbf{\Sigma}}^{+}\widehat{\mathbf{\Sigma}}) \simeq (v(0;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(0;\gamma_n)\Sigma + A)(v(0;\gamma_n)\Sigma + I_p)^{-1}, \tag{C.114}$$

where $v(0;\gamma_n)$ is as defined in (C.112), and $\widetilde{v}_g(0;\gamma_n)$ is defined via $v(0;\gamma_n)$ by

$$\widetilde{v}_g(0;\gamma_n) = \gamma_n \operatorname{tr}[A\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-2}]/p \cdot \big(v(0;\gamma_n)^{-2} - \gamma_n \operatorname{tr}[\Sigma^2(v(0;\gamma_n)\Sigma + I_p)^{-2}]/p\big)^{-1}. \tag{C.115}$$

Proof. The proofs for both the parts use the results of Lemma C.6.10 and a limiting argument as $\lambda \to 0^+$. The results of Lemma C.6.10 are pointwise in λ, but can be strengthened to be uniform in λ over a range that includes $\lambda = 0$ allowing one to take the limits of the deterministic equivalents obtained in Lemma C.6.10 as $\lambda \to 0^+$.

Part 1. We will use the result in Part 1 of Lemma C.6.10 as our starting point. Let $\Lambda := [0, \lambda_{\max}]$ where $\lambda_{\max} < \infty$, and let T be a matrix with bounded trace norm. Note that

$$|\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT]| \le \|(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}A\|_{\mathrm{op}} \operatorname{tr}[T] \le C\|(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}}\|A\|_{\mathrm{op}} \le C \tag{C.116}$$

for some constant $C < \infty$. Here, the last inequality follows because $s_i^2/(s_i^2+\lambda)^2 \le 1$ where s_i^2, $1 \le i \le p$, are the eigenvalues of $\widehat{\mathbf{\Sigma}}$, and the operator norm A is assumed to be bounded. Consider the magnitude of the derivative (in λ) of the map $\lambda \mapsto \operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT]$ given by

$$\left|\frac{\partial}{\partial\lambda}\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT]\right| = 2|\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-3}\widehat{\mathbf{\Sigma}}AT]|.$$

Following the argument in (C.116), for $\lambda \in \Lambda$, observe that

$$|\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-3}\widehat{\mathbf{\Sigma}}AT]| \le \|(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-3}\widehat{\mathbf{\Sigma}}\|_{\mathrm{op}}\|A\|_{\mathrm{op}} \operatorname{tr}[T] \le C$$

for some constant $C < \infty$. Similarly, in the same interval $\operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma AT] \le C$. In addition, from Lemma C.6.14, we have the map $\lambda \mapsto \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}AT]$ is differentiable in λ and the derivative for $\lambda \in \Lambda$ is bounded. Therefore, the family of functions $\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma AT]$ forms an equicontinuous family in λ over $\lambda \in \Lambda$. Thus, the convergence in Part 1 of Lemma C.6.10 is uniform in λ. We can now use the Moore-Osgood theorem to interchange the limits to obtain

$$\begin{aligned}
&\lim_{p\to\infty}\Big\{\operatorname{tr}[\widehat{\mathbf{\Sigma}}^+ AT] - \operatorname{tr}[\widetilde{v}(0;\gamma_n)(v(0;\gamma_n)\Sigma + I_p)^{-2}\Sigma AT]\Big\}\\
&= \lim_{p\to\infty}\lim_{\lambda\to 0^+}\Big\{\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma AT)]\Big\}\\
&= \lim_{\lambda\to 0^+}\lim_{p\to\infty}\Big\{\operatorname{tr}[(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-2}\widehat{\mathbf{\Sigma}}AT] - \operatorname{tr}[\widetilde{v}(-\lambda;\gamma_n)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-2}\Sigma AT)]\Big\}\\
&= 0.
\end{aligned}$$

In the first equality above, we used the fact that $\widehat{\mathbf{\Sigma}}^+ = \widehat{\mathbf{\Sigma}}^+\widehat{\mathbf{\Sigma}}\widehat{\mathbf{\Sigma}}^+ = \lim_{\lambda\to 0^+}(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}\widehat{\mathbf{\Sigma}}(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}$, and that the functions $v(\cdot;\gamma_n)$ and $\widetilde{v}(\cdot;\gamma_n)$ are continuous (which follows, from say Lemma C.6.15 (1)). This provides the right hand side of (C.111). Similarly, the fixed-point equation (C.98) as $\lambda \to 0^+$ becomes

$$v(0;\gamma_n)^{-1} = \gamma_n \operatorname{tr}[\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-1}]/p.$$

Moving $v(0;\gamma_n)$ to the other side (from Lemma C.6.13 (1), it follows that $v(0;\gamma_n) > 0$ for $\gamma_n \in (1,\infty)$), we arrive at the desired result.

Part 2. As done in Part 1, it is not difficult to show that over $\lambda \in \Lambda$ the family of functions $\operatorname{tr}[\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}T] - \operatorname{tr}[(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(-\lambda;\gamma_n)\Sigma + A)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}T]$ form an equicontinuous family. Therefore, the convergence in Part 2 of Lemma C.6.10 is uniform in λ over Λ (that includes 0). Using the Moore-Osgood theorem to the interchange the limits, one has

$$\lim_{p\to\infty}\Big\{\operatorname{tr}[(I_p - \widehat{\mathbf{\Sigma}}^+\widehat{\mathbf{\Sigma}})A(I_p - \widehat{\mathbf{\Sigma}}^+\widehat{\mathbf{\Sigma}})T]$$

$$
\begin{aligned}
&\qquad - \operatorname{tr}[(v(0;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(0;\gamma_n)\Sigma + A)(v(0;\gamma_n)\Sigma + I_p)^{-1}T]\Big\} \\
&= \lim_{p\to\infty}\lim_{\lambda\to 0^+}\Big\{\operatorname{tr}[\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}T] \\
&\qquad\qquad - \operatorname{tr}[(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(-\lambda;\gamma_n)\Sigma + A)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}T]\Big\} \\
&= \lim_{\lambda\to 0^+}\lim_{p\to\infty}\Big\{\operatorname{tr}[\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}A(\widehat{\mathbf{\Sigma}} + \lambda I_p)^{-1}T] \\
&\qquad\qquad - \operatorname{tr}[(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}(\widetilde{v}_g(-\lambda;\gamma_n)\Sigma + A)(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}T]\Big\} \\
&= 0.
\end{aligned}
$$

Now both (C.113) and (C.115) follow by taking $\lambda \to 0^+$ in (C.95) and (C.96), respectively.

This concludes the proof.

□

Corollary C.6.12 (Limiting deterministic equivalents for generalized bias and variance ridgeless resolvents)**.** *Assume the setting of Lemma C.6.10. Let $f : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a function. Then, as $n, p \to \infty$ and $p/n \to \gamma \in (1, \infty)$, the following equivalences hold:*

1. *Limiting generalized variance of ridgeless regression:*
$$\widehat{\mathbf{\Sigma}}^{+} f(\Sigma) \simeq \widetilde{v}(0;\gamma)(v(0;\gamma)\Sigma + I_p)^{-2}\Sigma f(\Sigma), \tag{C.117}$$
where $v(0;\gamma)$ and $\widetilde{v}(0;\gamma)$ are defined by (C.112) and (C.113), respectively.

2. *Limiting generalized bias of ridgeless regression:*
$$(I_p - \widehat{\mathbf{\Sigma}}^{+}\widehat{\mathbf{\Sigma}}) f(\Sigma)(I_p - \widehat{\mathbf{\Sigma}}^{+}\widehat{\mathbf{\Sigma}}) \simeq (1 + \widetilde{v}_g(0;\gamma))(v(0;\gamma)\Sigma + I_p)^{-1} f(\Sigma)(v(0;\gamma)\Sigma + I_p)^{-1}, \tag{C.118}$$
where $v(0;\gamma)$ is as defined in (C.112) and $\widetilde{v}_g(0;\gamma)$ is as defined in (C.115) with A replaced by $f(\Sigma)$.

Proof. The proof follows from Lemma C.6.11, in conjunction with Lemma C.6.13 ((1), (3), (4)) to provide continuity of the functions $v(0;\cdot)$, $\widetilde{v}(0;\cdot)$, and $\widetilde{v}_g(0;\cdot)$ (in the aspect ratio) over $(1,\infty)$. □

C.6.6 Lemmas on properties of solutions of certain fixed-point equations

In this section, we collect helper lemmas that are used in the proofs of Proposition 3.3.14 in Appendix C.3, Corollary 3.4.9 in Appendix C.5, and Lemma C.6.11 and Corollary C.6.12 in Appendix C.6.

Lemma C.6.13 (Continuity and limiting behavior of functions of the solution of a fixed-point equation in the aspect ratio)**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(0;\cdot) : \phi \mapsto v(0;\phi)$, over $(1,\infty)$, where $v(0;\phi) \ge 0$ is the unique solution to the fixed-point equation*
$$\frac{1}{\phi} = \int \frac{v(0;\phi) r}{1 + v(0;\phi) r}\,\mathrm{d}P(r). \tag{C.119}$$
Then, the following properties hold:

1. *The function $v(0;\cdot)$ is continuous and strictly decreasing over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+} v(0;\phi) = \infty$, and $\lim_{\phi\to\infty} v(0;\phi) = 0$.*

2. *The function $\phi \mapsto (\phi v(0;\phi))^{-1}$ is strictly increasing over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+}(\phi v(0;\phi))^{-1} = 0$ and $\lim_{\phi\to\infty}(\phi v(0;\phi))^{-1} = 1$.*

3. *The function $\widetilde{v}(0;\cdot) : \phi \mapsto \widetilde{v}(0;\phi)$, where*
$$\widetilde{v}(0;\phi) = \left(\frac{1}{v(0;\phi)^2} - \phi\int \frac{r^2}{(1 + r v(0;\phi))^2}\,\mathrm{d}P(r)\right)^{-1},$$
is continuous over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+}\widetilde{v}(0;\phi) = \infty$, and $\lim_{\phi\to\infty}\widetilde{v}(0;\phi) = 0$.

4. *The function* $\widetilde{v}_g(0;\cdot) : \phi \mapsto \widetilde{v}_g(0;\phi)$*, where*

$$\widetilde{v}_g(0;\phi) = \widetilde{v}(0;\phi)\phi \int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}P(r),$$

is continuous over $(1,\infty)$*. Furthermore,* $\lim_{\phi \to 1^+} \widetilde{v}_g(0;\phi) = \infty$*, and* $\lim_{\phi \to \infty} \widetilde{v}_g(0;\phi) = 0$.

5. *Let* Q *be a (fixed) probability distribution supported on* $[a,b]$ *that depends on a scalar* ϕ_1*. Then, the function* $\Upsilon_b(\phi_1;\cdot) : \phi \mapsto \Upsilon_b(\phi_1,\phi)$*, where*

$$\Upsilon_b(\phi_1,\phi) = (1 + \widetilde{v}_g(0;\phi)) \int \frac{1}{(1+v(0;\phi)r)^2} \, \mathrm{d}Q(r),$$

is continuous over $(1,\infty)$*. Furthermore,* $\Upsilon_b(\phi_1,\phi) < \infty$ *for* $\phi \in (1,\infty)$*, and* $\lim_{\phi \to \infty} \Upsilon_b(\phi_1,\phi) = 1$.

Proof. We consider the five parts separately below. Before doing so though, it is worth mentioning that for $\phi \in (1,\infty)$, there is a unique non-negative solution $v(0;\phi)$ to the fixed-point equation (C.119) as stated in the statement. This follows from Lemma C.6.15 (1). The following properties refer to the function $v(0;\cdot) : \phi \mapsto v(0;\phi)$ defined via this unique solution.

Part 1. We begin with the first part. Observe that the function

$$t \mapsto \int \frac{1}{1+tr} \, \mathrm{d}P(r)$$

is strictly decreasing and strictly convex over $(0,\infty)$. Thus, the function

$$T : t \mapsto 1 - \int \frac{1}{1+tr} \, \mathrm{d}P(r) = \int \frac{t}{1+tr} \, \mathrm{d}P(r)$$

is strictly increasing and strictly concave over $(0,\infty)$, with $\lim_{t \to 0} T(t) = 0$ and $\lim_{t \to \infty} T(t) = 1$. Since the inverse image of a strictly increasing and strictly concave real function is strictly increasing and strictly convex (see, e.g. Proposition 3 of Hiriart-Urruty and Martınez-Legaz (2003)), we have that T^{-1} is strictly convex and strictly increasing. This also implies that T^{-1} is continuous. Note that $v(0;\phi) = T^{-1}(\phi^{-1})$. Since ϕ^{-1} is continuous, it follows that $v(0;\cdot)$ is continuous. In addition, since $\phi \mapsto \phi^{-1}$ is strictly decreasing, we have that $v(0;\cdot)$ is strictly decreasing. Moreover, $\lim_{\phi \to 1^+} T^{-1}(\phi^{-1}) = \infty$, and $\lim_{\phi \to \infty} T^{-1}(\phi^{-1}) = 0$.

Part 2. From (C.119), we have

$$\frac{1}{\phi v(0;\phi)} = \int \frac{r}{1+v(0;\phi)r} \, \mathrm{d}P(r).$$

Because $v(0;\phi)$ is strictly decreasing over $(1,\infty)$, the right side of the display above is strictly increasing. Furthermore, because $\lim_{\phi \to 1^+} v(0;\phi) = \infty$, we have $\lim_{\phi \to 1^+} (\phi v(0;\phi))^{-1} = 0$, and because $\lim_{\phi \to \infty} v(0;\phi) = 0$, we have $\lim_{\phi \to \infty} (\phi v(0;\phi))^{-1} = 1$.

Part 3. From Part 1, the function $1/v(0;\cdot)^2$ is continuous. In addition, observe that the function

$$\phi \mapsto \int \frac{r^2}{(1+v(0;\phi)r)^2} \, \mathrm{d}P(r)$$

is also continuous. Finally, note that

$$\frac{1}{v(0;\phi)^2} - \phi \int \frac{r^2}{(1+rv(0;\phi))^2} \, \mathrm{d}P(r) = \frac{1}{v(0;\phi)^2} \left(1 - \phi \int \left(\frac{rv(0;\phi)}{1+rv(0;\phi)} \right)^2 \mathrm{d}P(r) \right) > 0,$$

where the last inequality holds for all $\phi \in (1, \infty)$ because $v(0; \phi) > 0$ over $\phi \in (1, \infty)$ from Part 1, and the term in the parenthesis is strictly positive over $\phi \in (1, \infty)$ because

$$\phi \int \left(\frac{rv(0; \phi)}{1 + rv(0; \phi)} \right)^2 \, \mathrm{d}P(r) < \phi \int \frac{rv(0; \phi)}{1 + rv(0; \phi)} \, \mathrm{d}P(r) = 1,$$

where the last equality follows from (C.119). Thus, $\widetilde{v}(0; \cdot)$ is continuous.

Furthermore, since $\lim_{\phi \to 1^+} v(0; \phi) = \infty$, it follows that $\lim_{\phi \to 1^+} \widetilde{v}(0; \phi) = \infty$. Similarly, from $\lim_{\phi \to \infty} v(0; \phi) = 0$ and the fact that

$$\lim_{\phi \to \infty} \int \frac{r^2}{(1 + rv(0; \phi))^2} \, \mathrm{d}P(r) \geq a^2 > 0,$$

it follows that $\lim_{\phi \to \infty} \widetilde{v}(0; \phi) = 0$.

Part 4. Similar to Part 3, continuity of $\widetilde{v}_g(0; \cdot)$ follows from the continuity of $\widetilde{v}(0; \cdot)$ and $v(0; \phi)$. To compute the desired limits, observe that

$$1 + \widetilde{v}_g(0; \phi) = \frac{\dfrac{1}{v(0; \phi)^2}}{\dfrac{1}{v(0; \phi)^2} - \phi \displaystyle\int \frac{r^2}{(1 + v(0; \phi) r)^2} \, \mathrm{d}P(r)}.$$

We thus have

$$(1 + \widetilde{v}_g(0; \phi))^{-1} = 1 - v(0; \phi)^2 \phi \int \frac{r^2}{(1 + rv(0; \phi))^2} \, \mathrm{d}P(r) \tag{C.120}$$

$$= 1 - \phi \int \frac{r^2}{(v(0; \phi)^{-1} + r)^2} \, \mathrm{d}P(r). \tag{C.121}$$

Because $\lim_{\phi \to 1^+} v(0; \phi) = \infty$, from (C.121), we have

$$\lim_{\phi \to 1^+} (1 + \widetilde{v}_g(0; \phi))^{-1} = 1 - \lim_{\phi \to 1^+} \phi \int \frac{r^2}{(v(0; \phi)^{-1} + r)^2} \mathrm{d}P(r) = 1 - 1 = 0.$$

It follows then that $\lim_{\phi \to 1^+} \widetilde{v}_g(0; \phi) = \infty$.

On the other hand, observe from (C.120) that

$$(1 + \widetilde{v}_g(0; \phi))^{-1} = 1 - \phi v(0; \phi) v(0; \phi) \int \frac{r^2}{(1 + rv(0; \phi))^2} \, \mathrm{d}P(r). \tag{C.122}$$

From Part 2, we have $\lim_{\phi \to \infty} \phi v(0; \phi) = 1$, and from Part 1, we have $\lim_{\phi \to \infty} v(0; \phi) = 0$. Moreover, since P is supported on $[a, b]$, and $v(0; \phi) > 0$ for $\phi \in (1, \infty)$ from Part 1, for $\phi \in (1, \infty)$, note that

$$0 < \int \frac{r^2}{(1 + rv(0; \phi))^2} < b^2.$$

Thus, from (C.122), we obtain

$$\lim_{\phi \to \infty} (1 + \widetilde{v}_g(0; \phi))^{-1} = 1 - 0 = 1.$$

We hence conclude that $\lim_{\phi \to \infty} \widetilde{v}_g(0; \phi) = 0$.

Part 5. The continuity claim follows from the continuity of $v(0;\cdot)$ and $\widetilde{v}_g(0;\cdot)$ from Parts 1 and 4, respectively. From calculation similar to that in Part 4, it follows that $(1+\widetilde{v}_g(0;\phi)) < \infty$ for $\phi \in (1,\infty)$. Now, since $v(0;\phi) > 0$ for $\phi \in (1,\infty)$ from Part 1, and Q is supported on $[a,b]$, observe that

$$\int \frac{1}{(1+v(0;\phi)r)^2} \, \mathrm{d}Q(r) \le 1 < \infty.$$

Hence, $\Upsilon_b(\phi_1,\phi) < \infty$ for $\phi \in (1,\infty)$. Moreover, because $\lim_{\phi\to\infty}(1+\widetilde{v}_g(0;\phi)) = 1$, and $\lim_{\phi\to\infty} v(0;\phi) = 0$, we obtain

$$\lim_{\phi\to\infty} \Upsilon_b(\phi_1,\phi) = \lim_{\phi\to\infty}(1+\widetilde{v}_g(0;\phi)) \cdot \lim_{\phi\to\infty} \int \frac{1}{(1+v(0;\phi)r)^2} \, \mathrm{d}Q(r) = 1.$$

Therefore, $\lim_{\phi\to\infty} \Upsilon_b(\phi_1,\phi) = 1$, as desired.

This completes all the five parts, and finishes the proof. □

Lemma C.6.14 (Bounding derivatives of the solution of a fixed-point equation in the regularization parameter)**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a,b]$. Let $\gamma \in (1,\infty)$ be a real number. Let $\Lambda = [0,\lambda_{\max}]$ for some constant $\lambda_{\max} < \infty$. For $\lambda \in \Lambda$, let $v(-\lambda;\gamma) \ge 0$ denote the solution to the fixed-point equation*

$$\frac{1}{v(-\lambda;\gamma)} = \lambda + \gamma \int \frac{r}{v(-\lambda;\gamma)r+1} \, \mathrm{d}P(r).$$

Then, the function $\lambda \mapsto v(-\lambda;\gamma)$ is twice differentiable over Λ. Furthermore, over Λ, $v(-\lambda;\gamma)$, $\partial/\partial\lambda[v(-\lambda;\gamma)]$, and $\partial^2/\partial\lambda^2[v(-\lambda;\gamma)]$ are bounded above. Furthermore, over Λ, absolute values of $v(-\lambda;\gamma)$, $\partial/\partial\lambda[v(-\lambda;\gamma)]$, and $\partial^2/\partial\lambda^2[v(-\lambda;\gamma)]$ are bounded above.

Proof. Start by re-writing the fixed-point equation as

$$\lambda = \frac{1}{v(-\lambda;\gamma)} - \gamma \int \frac{r}{v(-\lambda;\gamma)r+1} \, \mathrm{d}P(r).$$

Define a function f by

$$f(x) = \frac{1}{x} - \gamma \int \frac{r}{xr+1} \, \mathrm{d}P(r).$$

Observe that $v(-\lambda;\gamma) = f^{-1}(\lambda)$. The claim of twice differentiability of the function $\lambda \mapsto v(-\lambda;\gamma_n)$ follows from Lemma C.6.15 (4). The claim of boundedness of the function and its first derivatives (with respect to λ) follows from Lemma C.6.15 ((4), (5), (6)).

□

Lemma C.6.15 (Bounding derivatives of the solution of a fixed-point equation)**.** *Let $a > 0$ and $b < \infty$ be two real numbers. Let P be a probability distribution supported on $[a,b]$. Let $\gamma \in (1,\infty)$ be a real number. Define a function f by*

$$f(x) = \frac{1}{x} - \gamma \int \frac{r}{xr+1} \, \mathrm{d}P(r). \tag{C.123}$$

Then, the following properties hold:

1. *There is a unique $0 < x_0 < \infty$ such that $f(x_0) = 0$. The function f is twice differentiable and strictly decreasing over $(0,x_0)$, with $\lim_{x\to 0^+} f(x) = \infty$ and $f(x_0) = 0$.*

2. *The derivative f' is strictly increasing over $(0,x_0)$, with $\lim_{x\to 0^+} f'(x) = -\infty$ and $f'(x_0) < 0$.*

3. *The second derivative f'' is strictly decreasing over $(0,x_0)$, with $\lim_{x\to 0^+} f''(x) = \infty$ and $f''(x_0) > 0$.*

4. *The inverse function f^{-1} is twice differentiable, bounded over $[0,\infty)$ by $x_0 < \infty$, and strictly decreasing over $(0,\infty)$, with $f^{-1}(0) = x_0$ and $\lim_{y\to\infty} f^{-1}(y) = 0$.*

5. *The derivative of the inverse function* $(f^{-1})'$ *is bounded over* $[0, \infty)$ *by*

$$\frac{x_0^2}{1 - \gamma \displaystyle\int \left(\frac{x_0 r}{x_0 r + 1} \right)^2 \, \mathrm{d}P(r)} < \infty.$$

6. *The second derivative of the inverse function* $(f^{-1})''$ *is bounded over* $[0, \infty)$ *by*

$$\frac{2x_0^3}{\left(1 - \gamma \displaystyle\int \left(\frac{x_0 r}{x_0 r + 1} \right)^2 \, \mathrm{d}P(r) \right)^3} < \infty.$$

Proof. We consider different parts separately below.

Part 1. Observe that

$$f(x) = \frac{1}{x} - \gamma \int \frac{r}{xr + 1} \, \mathrm{d}P(r) = \frac{1}{x} \left(1 - \gamma \int \frac{xr}{xr + 1} \, \mathrm{d}P(r) \right).$$

The function $g : x \mapsto 1/x$ is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g(x) = \infty$ and $\lim_{x \to \infty} g(x) = 0$, while the function

$$h : x \mapsto 1 - \gamma \int \frac{xr}{xr + 1} \, \mathrm{d}P(r)$$

is strictly decreasing over $(0, \infty)$ with $h(0) = 1$ and $\lim_{x \to \infty} h(x) = 1 - \gamma < 0$. Thus, there is a unique $0 < x_0 < \infty$ such that $h(x_0) = 0$, and consequently $f(x_0) = 0$. Because h is positive over $[0, x_0]$, f, a product of two positive strictly decreasing functions, is strictly decreasing over $(0, x_0)$, with $\lim_{x \to 0^+} f(x) = \infty$ and $f(x_0) = 0$.

Part 2. The derivative f' at x is given by

$$f'(x) = -\frac{1}{x^2} + \gamma \int \frac{r^2}{(xr + 1)^2} \, \mathrm{d}P(r) = -\frac{1}{x^2} \left(1 - \gamma \int \left(\frac{xr}{xr + 1} \right)^2 \, \mathrm{d}P(r) \right).$$

The function $g : x \mapsto 1/x^2$ is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g(x) = \infty$ and $\lim_{x \to \infty} g(x) = 0$. On the other hand, the function

$$h : x \mapsto 1 - \gamma \int \left(\frac{xr}{xr + 1} \right)^2 \, \mathrm{d}P(r)$$

strictly decreasing over $(0, \infty)$ with $h(0) = 1$ and $h(x_0) > 0$. This follows because for $x \in [0, x_0]$,

$$\begin{aligned} \gamma \int \left(\frac{xr}{xr + 1} \right)^2 \, \mathrm{d}P(r) &\le \left(\frac{x_0 b}{x_0 b + 1} \right) \gamma \int \left(\frac{xr}{xr + 1} \right) \, \mathrm{d}P(r) \\ &< \gamma \int \frac{xr}{xr + 1} \, \mathrm{d}P(r) \le \gamma \int \frac{x_0 r}{x_0 r + 1} \, \mathrm{d}P(r) = 1, \end{aligned} \tag{C.124}$$

where the first inequality in the chain above follows as the support of P is $[a, b]$, and the last inequality follows since $f(x_0) = 0$ and $x_0 > 0$, which implies that

$$\frac{1}{x_0} = \gamma \int \frac{r}{x_0 r + 1} \, \mathrm{d}P(r), \quad \text{or equivalently that} \quad 1 = \gamma \int \frac{x_0 r}{x_0 r + 1} \, \mathrm{d}P(r).$$

Thus, $-f'$, a product of two positive strictly decreasing functions, is strictly decreasing, and in turn, f' is strictly increasing. Moreover, $\lim_{x \to 0^+} f'(x) = -\infty$ and $f'(x_0) < 0$.

Part 3. The second derivative f'' at x is given by

$$f''(x) = \frac{2}{x^3} - 2\gamma \int \frac{r^3}{(xr+1)^3} \, \mathrm{d}P(r) = \frac{2}{x^3} \left(1 - \gamma \int \left(\frac{xr}{xr+1} \right)^3 \mathrm{d}P(r) \right).$$

The rest of the arguments are similar to those in Part 2. The function $g : x \mapsto 1/x^3$ is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g(x) = \infty$ and $\lim_{x \to \infty} g(x) = 0$, while the function

$$h : x \mapsto 1 - \gamma \int \left(\frac{xr}{xr+1} \right)^3 \mathrm{d}P(r)$$

is strictly decreasing over $(0, \infty)$ with $h(0) = 1$ and $h(x_0) > 0$ as

$$\begin{aligned} \gamma \int \left(\frac{xr}{xr+1} \right)^3 \mathrm{d}P(r) &\le \left(\frac{x_0 b}{x_0 b + 1} \right)^2 \gamma \int \left(\frac{xr}{xr+1} \right) \mathrm{d}P(r) \\ &< \gamma \int \frac{xr}{xr+1} \, \mathrm{d}P(r) \le \gamma \int \frac{x_0 r}{x_0 r + 1} \, \mathrm{d}P(r) = 1. \end{aligned} \tag{C.125}$$

It then follows that f'' is strictly decreasing, with $\lim_{x \to 0^+} f''(x) = \infty$ and $f''(x_0) > 0$.

Part 4. Because f is twice differentiable and strictly monotonic over $(0, x_0)$, f^{-1} is twice differentiable and strictly monotonic (see, e.g., Problem 2, Chapter 5 of Rudin (1976)). Since $f(x_0) = 0$, $f^{-1}(0) = x_0$, and since $\lim_{x \to 0^+} f(x) = \infty$, $\lim_{y \to \infty} f^{-1}(y) = 0$. Hence, f^{-1} is bounded above over $[0, \infty)$ by $x_0 < \infty$.

Part 5. Because $f'(x) \neq 0$ over $(0, x_0)$, by the inverse function theorem, we have

$$\left| (f^{-1})'(f(x)) \right| = \left| \frac{1}{f'(x)} \right| < \left| \frac{1}{f'(x_0)} \right| = \frac{1}{\frac{1}{x_0^2} \left(1 - \gamma \int \left(\frac{xr}{xr+1} \right)^2 \mathrm{d}P(r) \right)} < \infty,$$

where the first inequality uses the fact that $|f'(x_0)| < |f'(x)|$ for $x \in (0, x_0]$ from Part 2, and the last inequality uses the bound from (C.124).

Part 6. Similar to Part 5, by inverse function theorem, we have

$$\left| (f^{-1})''(f(x)) \right| = \left| \frac{f''(x)}{f'(x)^3} \right| = \frac{\frac{2}{x^3} \left(1 - \gamma \int \left(\frac{xr}{xr+1} \right)^3 \mathrm{d}P(r) \right)}{\frac{1}{x^6} \left(1 - \gamma \int \left(\frac{xr}{xr+1} \right)^2 \mathrm{d}P(r) \right)^3} \le \frac{2x_0^3}{\left(1 - \gamma \int \left(\frac{xr}{xr+1} \right)^2 \mathrm{d}P(r) \right)^3} < \infty,$$

where the first inequality uses the bound from (C.125), and the second inequality uses the bound from (C.124).

This finishes all the six parts, and concludes the proof.

□

We remark that the technique of Lemma A.2 of Hastie et al. (2022) can be applied to obtain similar conclusions as those in Lemmas C.6.14 and C.6.15. However, since our parameterization is slightly different, we make use of the inverse function theorem instead of the implicit function theorem employed in Hastie et al. (2022).

C.6.7 Proof of Theorem C.6.16 (Risk characterization of one-step procedure with ridgeless regression)

The following theorem characterizes the risk of the one-step procedure starting with MN2LS base procedure for isotropic features under square error. Let $R^{\mathrm{det}}(\gamma; \widetilde{f}^{\mathrm{os}})$ denote the risk of the one-step predictor starting with the MN2LS base predictor on i.i.d. data with limiting aspect ratio γ.

Theorem C.6.16 (Limiting risk of one-step procedure with ridgeless regression)**.** *Suppose assumptions* (ℓ_2A1), (ℓ_2A2) *with* $\Sigma = I$, (ℓ_2A3) *hold true. Let* $\mathrm{SNR} := \rho^2/\sigma^2$. *Then, the limiting risk of the one-step predictor starting with the MN2LS base predictor under* (PA(γ)) *is given as follows:*

- *When* $\mathrm{SNR} \le 1$*:*

$$\frac{R^{\mathrm{det}}(\gamma; \widehat{f}^{\mathrm{os}})}{\sigma^2} - 1 = \begin{cases} \dfrac{\gamma}{1-\gamma} & \text{if } \gamma \le \dfrac{\mathrm{SNR}}{\mathrm{SNR}+1} < 1 \\ \mathrm{SNR} & \text{otherwise.} \end{cases}$$

- *When* $1 < \mathrm{SNR} \le \mathrm{SNR}^{\star} (\approx 10.7041)$*:*

$$\frac{R^{\mathrm{det}}(\gamma; \widehat{f}^{\mathrm{os}})}{\sigma^2} - 1 =$$

$$\begin{cases} \dfrac{\gamma}{1-\gamma} & \text{if } \gamma \le 1 - \dfrac{1}{2\sqrt{2\sqrt{\mathrm{SNR}}-1}} < 1 \\ 2\sqrt{2\sqrt{\mathrm{SNR}}-1} - 1 & \text{if } 1 - \dfrac{1}{2\sqrt{2\sqrt{\mathrm{SNR}}-1}} < \gamma \le \left(2 - \dfrac{1}{\sqrt{\mathrm{SNR}}} - \dfrac{1}{\sqrt{2\sqrt{\mathrm{SNR}}-1}}\right)^{-1} \\ \left\{\mathrm{SNR}\left(1 - \dfrac{1}{\zeta_1}\right) + \dfrac{1}{\zeta_1 - 1}\right\}\left(1 - \dfrac{1}{\zeta_2}\right) & \\ \quad + \dfrac{1}{\zeta_2 - 1} & \text{otherwise,} \end{cases}$$

  where $\mathrm{SNR}^{\star}$ *(which is approximately 10.7041) is value of* $x > 1$ *that solves*

$$1 - \frac{1}{2\sqrt{2\sqrt{x}-1}} = \left(2 - \frac{1}{x} - \frac{1}{\sqrt{2\sqrt{x}-1}}\right)^{-1}, \tag{C.126}$$

  and $\zeta_1, \zeta_2 \ge 1$ *are solutions to the equations*

$$\mathrm{SNR}\left(\frac{1}{\zeta_1} - \frac{1}{\zeta_2}\right) = \frac{\zeta_1^2}{(\zeta_1 - 1)^2} - \frac{\zeta_2^2}{(\zeta_2 - 1)^2} + \frac{1}{\zeta_1 - 1}\left(1 - \frac{\zeta_1}{\zeta_2}\frac{\zeta_1}{(\zeta_1 - 1)}\right) \tag{C.127}$$

$$\frac{1}{\zeta_1} + \frac{1}{\zeta_2} = \frac{1}{\gamma}. \tag{C.128}$$

- *When* $\mathrm{SNR} > \mathrm{SNR}^{\star}$*:*

$$\frac{R^{\mathrm{det}}(\gamma; \widehat{f}^{\mathrm{os}})}{\sigma^2} - 1 = \begin{cases} \dfrac{\gamma}{1-\gamma} & \text{if } \gamma \le \gamma^{\star} < 1 \\ \left\{\mathrm{SNR}\left(1 - \dfrac{1}{\zeta_1}\right) + \dfrac{1}{\zeta_1 - 1}\right\}\left(1 - \dfrac{1}{\zeta_2}\right) + \dfrac{1}{\zeta_2 - 1} & \text{otherwise,} \end{cases}$$

  where $\mathrm{SNR}^{\star}$ *is as defined in* (C.126), $\gamma^{\star}$ *is given by*

$$1 - \left(1 + \min_{\gamma \le 1}\left\{\mathrm{SNR}\left(1 - \frac{1}{\zeta_1}\right) + \frac{1}{\zeta_1 - 1}\right\}\left(1 - \frac{1}{\zeta_2}\right) + \frac{1}{\zeta_2 - 1}\right)^{-1},$$

  and $\zeta_1, \zeta_2 \ge 1$ *are solutions to the set of equations* (C.127) *and* (C.128).

Furthermore, in each case, the limiting risk is a non-decreasing function of γ.

Proof. From Proposition 3.4.10, it follows that that the limiting risk of the ingredient one-step predictor for various limiting split proportions (ζ_1, ζ_2) under isotropic features is given by

$$R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f}) - 1 = \begin{cases} \left\{\rho^2\left(1 - \frac{1}{\zeta_1}\right) + \sigma^2\left(\frac{1}{\zeta_1 - 1}\right)\right\}\left(1 - \frac{1}{\zeta_2}\right) + \sigma^2\left(\frac{1}{\zeta_2 - 1}\right) & \text{when } \zeta_1 > 1, \zeta_2 > 1 \\ \left\{\sigma^2\left(\frac{\zeta_1}{1 - \zeta_1}\right)\right\}\left(1 - \frac{1}{\zeta_2}\right) + \sigma^2\left(\frac{1}{\zeta_2 - 1}\right) & \text{when } \zeta_1 < 1, \zeta_2 > 1 \\ \sigma^2\left(\frac{\zeta_2}{1 - \zeta_2}\right) & \text{when } \zeta_2 < 1. \end{cases}$$

Note that the last case covers both $\zeta_1 > 1$ and $\zeta_1 < 1$. Given a fixed γ, our goal is to minimize $R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f})$ with the constraint $\frac{1}{\zeta_1} + \frac{1}{\zeta_2} \le \frac{1}{\gamma}$.

To simplify the calculations below, we first scale out the factor of σ^2 and express the risk in terms of $\mathrm{SNR} := \frac{\rho^2}{\sigma^2}$ to write

$$\frac{R^{\mathrm{det}}(\zeta_1, \zeta_2; \widetilde{f})}{\sigma^2} - 1 = \begin{cases} \left\{\mathrm{SNR}\left(1 - \frac{1}{\zeta_1}\right) + \left(\frac{1}{\zeta_1 - 1}\right)\right\}\left(1 - \frac{1}{\zeta_2}\right) + \left(\frac{1}{\zeta_2 - 1}\right) & \text{when } \zeta_1 > 1, \zeta_2 > 1 \\ \left\{\frac{\zeta_1}{1 - \zeta_1}\right\}\left(1 - \frac{1}{\zeta_2}\right) + \left(\frac{1}{\zeta_2 - 1}\right) & \text{when } \zeta_1 < 1, \zeta_2 > 1 \\ \left(\frac{\zeta_2}{1 - \zeta_2}\right) & \text{when } \zeta_2 < 1. \end{cases}$$

The problem of minimizing $R(\widehat{\beta}^{\mathrm{os}})$ can now be broken into three separate minimization problems, one for each of the cases above. The final allocation is then the one that gives the minimum among the three cases.

We next notice a simple observation that lets us eliminate the third case. Any feasible allocation of ζ_1 and ζ_2 in the third case is also a feasible allocation for the second case. This can be seen by making ζ_1 for the second case equal to ζ_2 in the third case and letting ζ_2 for the second case tend to ∞. Moreover, this gives the same objective value for both the cases. Hence, the minimum of the second case is no larger than the minimum of the third case and we can ignore the minimization of the third case.

Overall we are thus left with two minimization problems:

$$\begin{array}{ll} \text{minimize} & \left\{\mathrm{SNR}\left(1 - \frac{1}{\zeta_1}\right) + \left(\frac{1}{\zeta_1 - 1}\right)\right\}\left(1 - \frac{1}{\zeta_2}\right) + \left(\frac{1}{\zeta_2 - 1}\right) \\ \text{subject to} & \frac{1}{\zeta_1} + \frac{1}{\zeta_2} \le \frac{1}{\gamma} \\ & \zeta_1 > 1 \\ & \zeta_2 > 1 \end{array} \tag{C.129}$$

from the first case, and

$$\begin{array}{ll} \text{minimize} & \left\{\frac{\zeta_1}{1 - \zeta_1}\right\}\left(1 - \frac{1}{\zeta_2}\right) + \left(\frac{1}{\zeta_2 - 1}\right) \\ \text{subject to} & \frac{1}{\zeta_1} + \frac{1}{\zeta_2} \le \frac{1}{\gamma} \\ & \zeta_1 < 1 \\ & \zeta_2 > 1 \end{array} \tag{C.130}$$

from the second case. We now in turn analyze both of these optimization problems.

Optimization problem (C.130)

Let's start with the problem (C.130). Note that the objective function of the optimization problem (C.130) does not depend on SNR. Hence the optimal value will only be a function of γ. In addition, the constraint $\zeta_1 < 1$ is only satisfied when $\gamma < 1$. Thus, when $\gamma > 1$, the problem is infeasible. We divide the remaining range of γ into two main cases of $0 < \gamma < 0.5$ and $0.5 < \gamma < 1$. In each of the cases, we show that the minimum value of the problem is $\frac{\gamma}{1-\gamma}$, which is achieved by setting $\zeta_1 = \gamma$ and $\zeta_2 = \infty$.

When $\gamma \leq 0.5$. We first note that any allocation $\zeta_1 > 0.5$ is suboptimal because when $\zeta_1 > 0.5$, we have $\frac{\zeta_1}{1-\zeta_1} > 1$ by Lemma C.6.17 (3). Thus using Lemma C.6.18 (3), the objective function in this case is always larger than 1 for such ζ_1. However, we can achieve 1 by setting $\zeta_1 = 0.5$ and $\zeta_2 \to \infty$. Therefore we only need to consider $\zeta_1 \leq 0.5$. For such ζ_1, we have $\frac{\zeta_1}{1-\zeta_1} \leq 1$ by Lemma C.6.17 (1). Now using Lemma C.6.18 (1), the optimal allocation is obtained by setting $\zeta_2 \to \infty$ and choosing the least ζ_1, which is γ, and the corresponding optimal value is $\frac{\gamma}{1-\gamma}$.

When $0.5 < \gamma < 1$. We claim that the optimum value is still $\frac{\gamma}{1-\gamma}$, which is achieved by setting $\zeta_1 = \gamma$ and $\zeta_2 \to \infty$. This is a slightly more involved argument than the previous case because now $\frac{\zeta_1}{1-\zeta_1}$ will be larger than 1 since $\zeta_1 > \gamma > 0.5$, and hence there is a possibility of optimal allocation other than $\zeta_1 = \gamma$ and $\zeta_2 = \infty$. We proceed as follows.

Consider any feasible $\zeta_1 < 1$. On one hand, using Lemma C.6.18 (2), we note that the unconstrained optimal $\zeta_2^\star$ for this ζ_1 is $\frac{\sqrt{\frac{\zeta_1}{1-\zeta_1}}}{\sqrt{\frac{\zeta_1}{1-\zeta_1}}-1}$. On the other hand, from the constraint $\frac{1}{\zeta_2} \leq \frac{1}{\gamma} - \frac{1}{\zeta_1}$, we know that we need to satisfy $\zeta_2 \geq \frac{1}{\frac{1}{\gamma}-\frac{1}{\zeta_1}}$. There are now two possible scenarios.

- When $\frac{4}{7} < \gamma < 1$.

  In this case, we verify that any feasible ζ_1 (such that $\gamma \leq \zeta_1 < 1$) satisfies

$$\frac{\sqrt{\frac{\zeta_1}{1-\zeta_1}}}{\sqrt{\frac{\zeta_1}{1-\zeta_1}} - 1} < \frac{1}{\frac{1}{\gamma} - \frac{1}{\zeta_1}}.$$

  To see this, the above inequality after separating components of γ and ζ_1 reads

$$\frac{1}{\gamma} < \frac{1}{\zeta_1} + 1 - \sqrt{\frac{1}{\zeta_1} - 1}.$$

  It is easy to check that the function $x \mapsto 1 + \frac{1}{x} - \sqrt{\frac{1}{x} - 1}$ attains minimum value of $\frac{7}{4}$ (at $x = \frac{4}{5}$) on the interval $0.5 < x < 1$. Thus whenever $\gamma > \frac{4}{7}$, this condition will be satisfied for all feasible ζ_1. In this case, from Lemma C.6.18 (2), the optimal ζ_2 that satisfy the constraint is $\frac{1}{\frac{1}{\gamma}-\frac{1}{\zeta_1}}$. Plugging this value into the objective function, we arrive at the objective function

$$\left\{\frac{\zeta_1}{1-\zeta_1}\right\}\left(1 - \frac{1}{\gamma} + \frac{1}{\zeta_1}\right) + \frac{\frac{1}{\gamma} - \frac{1}{\zeta_1}}{1 - \frac{1}{\gamma} + \frac{1}{\zeta_1}}$$

  and the overall optimization problem reduces to

$$\begin{aligned} \text{minimize} \quad & \left\{\tfrac{\zeta_1}{1-\zeta_1}\right\}\left(1 - \tfrac{1}{\gamma} + \tfrac{1}{\zeta_1}\right) + \tfrac{\frac{1}{\gamma}-\frac{1}{\zeta_1}}{1-\frac{1}{\gamma}+\frac{1}{\zeta_1}} \\ \text{subject to} \quad & \zeta_1 \geq \gamma \geq \tfrac{4}{7} \\ & \zeta_1 < 1. \end{aligned} \tag{C.131}$$

  We can verify that the objective function is increasing in the constraint set and achieves the minimum at $\zeta_1 = \gamma$. The corresponding ζ_2 then tends to ∞ as desired.

- When $0.5 < \gamma < \frac{4}{7}$, or equivalently $\frac{7}{4} < \frac{1}{\gamma} < 2$.

  In this case, we can check that when

$$\frac{\frac{2}{\gamma} - \sqrt{\frac{4}{\gamma} - 7} - 1}{2\left(\frac{1}{\gamma^2} - \frac{2}{\gamma} + 2\right)} \leq \zeta_1 \leq \frac{\frac{2}{\gamma} + \sqrt{\frac{4}{\gamma} - 7} - 1}{2\left(\frac{1}{\gamma^2} - \frac{2}{\gamma} + 2\right)}, \tag{C.132}$$

we have

$$\frac{1}{\gamma} > \frac{1}{\zeta_1} + 1 - \sqrt{\frac{1}{\zeta_1} - 1}$$

which leads to

$$\frac{1}{\frac{1}{\gamma} - \frac{1}{\zeta_1}} < \frac{\sqrt{\frac{\zeta_1}{1-\zeta_1}}}{\sqrt{\frac{\zeta_1}{1-\zeta_1}} - 1}$$

Thus $\zeta_2^\star = \frac{\sqrt{\frac{\zeta_1}{1-\zeta_1}}}{\sqrt{\frac{\zeta_1}{1-\zeta_1}}-1}$ is feasible. The objective at this ζ_2 is $2\sqrt{\frac{\zeta_1}{1-\zeta_1}} - 1$. Now note that the function $x \mapsto 2\sqrt{\frac{x}{1-x}} - 1$ is increasing for $0 < x < 1$ and thus the optimal ζ_1 in this case is the lower point of the above interval (C.132). The optimal value for this case is thus given by

$$2\sqrt{\frac{\frac{2}{\gamma} - \sqrt{\frac{4}{\gamma} - 7} - 1}{\frac{2}{\gamma^2} - \frac{4}{\gamma} + 4 - \frac{2}{\gamma} + \sqrt{\frac{4}{\gamma} - 7} + 1}} - 1.$$

While when

$$\gamma < \zeta_1 < \frac{\frac{2}{\gamma} - \sqrt{\frac{4}{\gamma} - 7} - 1}{2\left(\frac{1}{\gamma^2} - \frac{2}{\gamma} + 2\right)}, \quad \text{or} \quad \frac{\frac{2}{\gamma} + \sqrt{\frac{4}{\gamma} - 7} - 1}{2\left(\frac{1}{\gamma^2} - \frac{2}{\gamma} + 2\right)} < \zeta_1 < 1,$$

we have

$$\frac{1}{\gamma} < \frac{1}{\zeta_1} + 1 - \sqrt{\frac{1}{\zeta_1} - 1}.$$

As argued before, in this case, the optimal ζ_2 is $\frac{1}{\frac{1}{\gamma} - \frac{1}{\zeta_1}}$ and the objective function at this value is given by

$$\left\{\frac{\zeta_1}{1-\zeta_1}\right\}\left(1 - \frac{1}{\gamma} + \frac{1}{\zeta_1}\right) + \frac{\frac{1}{\gamma} - \frac{1}{\zeta_1}}{1 - \frac{1}{\gamma} + \frac{1}{\zeta_1}}.$$

This function is again increasing in ζ_1 in the constrained set and hence the optimal value of ζ_1 is the lower point when $\zeta_1 = \gamma$ leading to the optimal value $\frac{\gamma}{1-\gamma}$. Now, we have

$$\frac{\gamma}{1-\gamma} < 2\sqrt{\frac{\frac{2}{\gamma} - \sqrt{\frac{4}{\gamma} - 7} - 1}{\frac{2}{\gamma^2} - \frac{4}{\gamma} + 4 - \frac{2}{\gamma} + \sqrt{\frac{4}{\gamma} - 7} + 1}} - 1$$

for $0.5 < \gamma < \frac{4}{7}$. Thus overall, even in this case, the optimal allocation is $\zeta_1 = \gamma$ and $\zeta_2 \to \infty$.

Optimization problem (C.129)

We now turn to problem (C.129). In this case, the solution depends on both SNR and γ. Note that the objective function can be written more compactly as $h(\zeta_2; h(\zeta_1; \mathrm{SNR}))$ where $h(\gamma; \mathrm{SNR})$ is defined as

$$h(\gamma; \mathrm{SNR}) = \mathrm{SNR}\left(1 - \frac{1}{\gamma}\right) + \frac{1}{\gamma - 1}.$$

We first consider the case when $\mathrm{SNR} \leq 1$. We argue that the optimum value in this case is SNR itself and it is achieved by setting both $\zeta_1 \to \infty$ and $\zeta_2 \to \infty$. This can be seen as follows. For any feasible $\zeta_1 > 1$, the minimum value of $h(\gamma; \mathrm{SNR})$ is SNR and it is achieved as $\zeta_1 \to \infty$ from Lemma C.6.18 (1). Since this minimum value is less than 1, $h(\zeta_2; \mathrm{SNR})$ is again minimized as $\zeta_2 \to \infty$ and overall minimum is SNR.

Let us consider the case when $\mathrm{SNR} > 1$. For ease of notation, we denote SNR by s.

We first claim that we can restrict to $\zeta_1 \geq \frac{\sqrt{s}}{\sqrt{s}-1}$ without loss of generality. This is because for any $1 < \zeta_1 < \frac{\sqrt{s}}{\sqrt{s}-1}$, there is a corresponding $\zeta_1 \geq \frac{\sqrt{s}}{\sqrt{s}-1}$ that gives either the same or smaller objective value while enlarging the constraint set for ζ_2. This claim follows from Lemma C.6.19 (1).

Next observe that the minimum without the constraint $\frac{1}{\zeta_1} + \frac{1}{\zeta_2} \leq \frac{1}{\gamma}$ is

$$2\sqrt{2\sqrt{s}-1}-1,$$

which is achieved by setting $\zeta_1 = \frac{\sqrt{s}}{\sqrt{s}-1}$ and $\zeta_2 = \frac{\sqrt{2\sqrt{s}-1}}{\sqrt{2\sqrt{s}-1}-1}$. The values of γ for which this value is achievable are:

$$\gamma \leq \left(1 - \frac{1}{\sqrt{s}} + 1 - \frac{1}{\sqrt{2\sqrt{s}-1}}\right)^{-1}. \tag{C.133}$$

In other words, the optimum value of problem (C.129) is $2\sqrt{2\sqrt{s}-1}-1$ for γ satisfying (C.133) achieved by setting $\zeta_1 = \frac{\sqrt{s}}{\sqrt{s}-1}$ and $\zeta_2 = \frac{\sqrt{2\sqrt{s}-1}}{\sqrt{2\sqrt{s}-1}-1}$.

Now we consider γ bigger than (C.133). For such γ, we need to move either (or both) of ζ_1 and ζ_2 from their unconstrained optimum values above. We claim that the constraint $\frac{1}{\zeta_1} + \frac{1}{\zeta_2} \leq \frac{1}{\gamma}$ need to be satisfied with equality in this case. This can be seen as follows. By way of contradiction, suppose the optimal allocation is $(\zeta_1^\star, \zeta_2^\star)$, and $\frac{1}{\zeta_1^\star} + \frac{1}{\zeta_2^\star} < \frac{1}{\gamma}$. We now argue that we can strictly decrease the objective function while satisfying the constraint by producing a feasible allocation $(\zeta_1^{\star\star}, \zeta_2^{\star\star})$ that strictly dominates the assumed allocation. We have two cases to consider.

1. $\zeta_1^\star \geq \frac{\sqrt{s}}{\sqrt{s}-1}$ and $\zeta_2^\star > \frac{\sqrt{2\sqrt{s}-1}}{\sqrt{2\sqrt{s}-1}-1}$. In this case, observe that we can keep $\zeta_1^{\star\star} = \zeta_1^\star$ and decrease $\zeta_2^\star$ so that $\zeta_2^{\star\star} = \frac{1}{\gamma} - \frac{1}{\zeta_1^\star}$. This is feasible. Now note that

$$h(\zeta_2^{\star\star}; h(\zeta_1^{\star\star}; s)) = h(\zeta_2^{\star\star}; h(\zeta_1^\star; s)) < h(\zeta_2^\star; h(\zeta_1^\star; s))$$

where the inequality follows from Lemma C.6.19 (2). Thus, the new allocation strictly decreases the objective value.

2. $\zeta_1^\star > \frac{\sqrt{s}}{\sqrt{s}-1}$ and $\zeta_2^\star = \frac{\sqrt{2\sqrt{s}-1}}{\sqrt{2\sqrt{s}-1}-1}$. In this case, we can decrease $\zeta_1^\star$ first so that $\zeta_1^{\star\star} = \frac{1}{\gamma} - \frac{1}{\zeta_2^\star}$, and keep $\zeta_2^{\star\star} = \zeta_2^\star$. Observe that this modification keeps us in the feasible region. Now note that

$$h(\zeta_2^{\star\star}; h(\zeta_1^{\star\star}; s)) = h(\zeta_2^\star; h(\zeta_1^{\star\star}; s)) < h(\zeta_2^\star; h(\zeta_1^\star; s))$$

where the inequality follows from Lemma C.6.19 (1). Thus, the objective value is again strictly smaller.

Hence, in both the cases, the objective value can be strictly improved while staying within the feasible constraint. Therefore, we must hit the constraint with equality.

With the equality constraint, we can now use the method of Lagrange multipliers. The Lagrangian is given by

$$\mathcal{L}(\zeta_1, \zeta_2, \mu) = h(\zeta_2; h(\zeta_1; s)) + \mu\left(\frac{1}{\zeta_1} + \frac{1}{\zeta_2} - \frac{1}{\gamma}\right).$$

The optimality conditions are given by the following system of equations in (ζ_1, ζ_2, μ)

$$\left\{s\left(1 - \frac{1}{\zeta_1}\right) + \frac{1}{\zeta_1 - 1}\right\}\frac{1}{\zeta_2^2} - \frac{1}{(\zeta_2 - 1)^2} - \frac{\mu}{\zeta_2^2} = 0$$

$$\left(1 - \frac{1}{\zeta_2}\right)\left\{\frac{s}{\zeta_1^2} - \frac{1}{(\zeta_1 - 1)^2}\right\} - \frac{\mu}{\zeta_1^2} = 0$$

$$\frac{1}{\zeta_1} + \frac{1}{\zeta_2} = \frac{1}{\gamma}.$$

After minor simplifications, these lead to

$$\begin{aligned}
s\left(1 - \frac{1}{\zeta_1}\right) - \mu &= \frac{\zeta_1^2}{(\zeta_1 - 1)^2} - \frac{1}{\zeta_1 - 1} \\
s\left(1 - \frac{1}{\zeta_2}\right) - \mu &= \frac{\zeta_1^2}{(\zeta_1 - 1)^2}\left(1 - \frac{1}{\zeta_2}\right) \\
\frac{1}{\zeta_1} + \frac{1}{\zeta_2} &= \frac{1}{\gamma}.
\end{aligned}$$

Eliminating μ, we get two equations in two unknowns (ζ_1, ζ_2):

$$\begin{aligned}
s\left(\frac{1}{\zeta_1} - \frac{1}{\zeta_2}\right) &= \frac{\zeta_1^2}{(\zeta_1 - 1)^2} - \frac{\zeta_2^2}{(\zeta_2 - 1)^2} + \frac{1}{\zeta_1 - 1}\left(1 - \frac{\zeta_1}{\zeta_2}\frac{\zeta_1}{(\zeta_1 - 1)}\right) \\
\frac{1}{\zeta_1} + \frac{1}{\zeta_2} &= \frac{1}{\gamma},
\end{aligned}$$

as claimed.

Finally, to obtain various boundary cutoff points for γ and SNR in each of the cases, note that:

- When $x = \frac{\mathrm{SNR}}{\mathrm{SNR}+1}$, we have $\frac{x}{1-x} = \mathrm{SNR}$.
- When $x = 1 - \frac{1}{2\sqrt{2\sqrt{\mathrm{SNR}}-1}}$, we have $\frac{x}{x-\gamma} = 2\sqrt{2\sqrt{\mathrm{SNR}} - 1} - 1$. In addition, from a short calculation it follows that, when $\mathrm{SNR} \approx 10.704$, we have $1 - \frac{1}{2\sqrt{2\sqrt{\mathrm{SNR}}-1}} = \left(2 - \frac{1}{\sqrt{\mathrm{SNR}}} - \frac{1}{\sqrt{2\sqrt{\mathrm{SNR}}-1}}\right)^{-1}$.
- When $x = \gamma^\star$, we have $\frac{x}{1-x} = \min_{\gamma \le 1} h(\gamma_2; h(\gamma_1; \mathrm{SNR}))$.

This finishes the proof. See Figure C.5 for an illustration of the optimal splitting of the aspect ratios $(\zeta_1^\star(\gamma), \zeta_2^\star(\gamma))$ for a given γ for two different SNR values. □

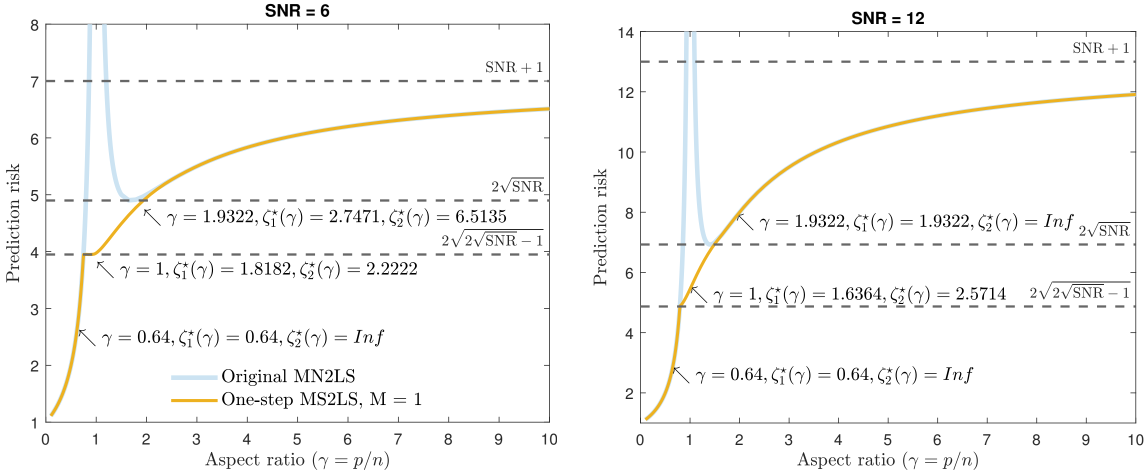


Figure C.5: Illustration of the optimal splitting of the aspect ratios for the one-step optimization with MN2LS base prediction procedure. Here, $(\zeta_1^\star(\gamma), \zeta_2^\star(\gamma))$ indicates the optimal splitting of the aspect ratio γ for the first and second splits.

C.6.8 Lemmas on properties of risk profile of ridgeless regression

In this section, we collect helper lemmas used in the proof of Theorem C.6.16. All the lemmas in this section are quite elementary, and only abstracted out for ease of repeated use in the proof of Theorem C.6.16.

Lemma C.6.17 (Properties of ridgeless risk profile in the underparameterized regime)**.** *The function $g : x \mapsto \frac{x}{1-x}$ over the domain $(0, 1)$ has the following properties:*

1. *The function g is increasing in x.*
2. *When $x \leq 0.5$, $g(x) \leq 1$.*
3. *When $x > 0.5$, $g(x) > 1$.*

Proof. The claims are easy to check. See Figure C.6 (the $x < 1$ segment) for illustration. □

Lemma C.6.18 (Properties of ridgeless risk profile in the overparameterized regime)**.** *Let $h(\cdot; s) : x \mapsto s\left(1 - \frac{1}{x}\right) + \frac{1}{x-1}$ be a function defined on the domain $x > 1$, parametrized by $s \geq 0$. The function h has the following properties:*

1. *When $s \leq 1$, the function is decreasing in x and approaches the minimum value of s as $x \to \infty$.*
2. *When $s > 1$, the function attains the minimum value of $2\sqrt{s} - 1$ at $x = \frac{\sqrt{s}}{\sqrt{s}-1}$.*
3. *When $s > 1$, $h(x; s) > 1$ for all $x > 1$.*
4. *For $x > \frac{\sqrt{s}}{\sqrt{s}-1}$, the function is increasing in x.*
5. *The function $s \mapsto h(x; s)$ is increasing in s for $s \geq 0$ for any fixed $x > 1$.*

Proof. The first property is easy to check. The second property follows elementary calculus. The third property follows from the second property. The fourth property follows by inspecting the derivative of $h(\cdot; s)$ for $x > \frac{\sqrt{s}}{\sqrt{s}-1}$. The fifth property is easy to check. See Figure C.6 (the $x > 1$ segment) for illustration.

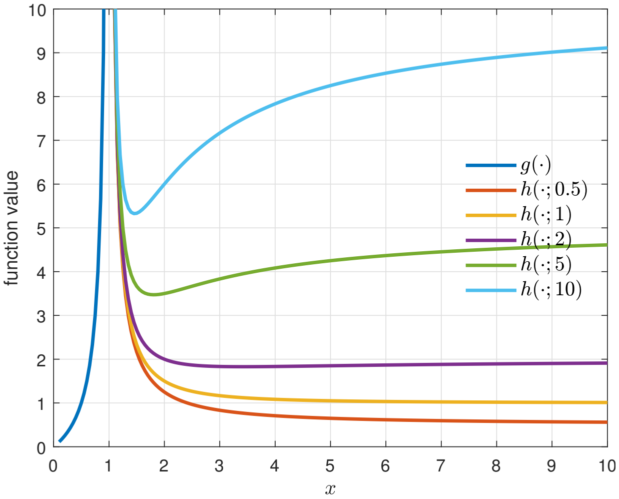


Figure C.6: Illustration of ridgeless risk profile with varying SNR.

□

Lemma C.6.19 (Properties of ridgeless one-step ingredient risk profile in the overparameterized regime)**.** *Let $h(x; s) : x \mapsto s\left(1 - \frac{1}{x}\right) + \frac{1}{x-1}$ be a function defined on the domain $x > 1$, parameterized by $s \geq 1$. Let $g : (x, y) \mapsto h(y; h(x; s))$ be a function defined on the domain $x > 1$ and $y > 1$, parameterized by $s \geq 1$. The function g has the following properties:*

1. *For any fixed $y > 1$, the function g is minimized at $x = \frac{\sqrt{s}}{\sqrt{s}-1}$ and increasing in x for $x \geq \frac{\sqrt{s}}{\sqrt{s}-1}$.*
2. *For any fixed $x > 1$, $g(x, y)$ is increasing over $y \geq \frac{\sqrt{h(x;s)}}{\sqrt{h(x;s)}-1}$.*

Proof. The first claim follows from Lemma C.6.18 (2), (4), (5). The second claim follows from Lemma C.6.18 (4). □

C.6.9 Control of additive error term in expectation

The following remark complements Remark 3.2.8 and specifies the growth allowed conditions on $\widehat{\sigma}_\Xi$ to ensure that $\mathbb{E}[\Delta_n^{\mathrm{add}}] = o(1)$.

Remark C.6.20 (Tolerable growth rates on $\widehat{\sigma}_\Xi$ for $\mathbb{E}\Delta_n^{\mathrm{add}} = o(1)$)**.** Suppose $|\Xi| \leq n^S$ for some $S < \infty$. Under the setting of Lemma 3.2.4, if for some $t \geq 1$,

$$\max_{\xi \in \Xi} \|\widehat{\sigma}_\xi\|_{L_t} = o\left(\frac{n_{\mathrm{te}}^{1/2}}{n^{-A+(A+S)/t}}\right),$$

then $\mathbb{E}[\Delta_n^{\mathrm{add}}] = o(1)$. On the other hand, under the setting of Lemma 3.2.5, if

$$\max_{\xi \in \Xi} \|\widehat{\sigma}_\xi\|_{L_2} = o\left(\frac{n_{\mathrm{te}}^{1/2}}{n^{(S-A)/2}}\right)$$

then $\mathbb{E}[\Delta_n^{\mathrm{add}}] = o(1)$. The remark follows simply by observing that the first term in the expectation bounds (3.11) and (3.13) for both Lemmas 3.2.4 and 3.2.5 are $o(1)$, while the second term in Lemma 3.2.4 is of order

$$O\left(\frac{n^{-A/r+S/t}}{n_{\mathrm{te}}^{1/2}}\right) \max_{\xi \in \Xi} \|\widehat{\sigma}_\xi\|_{L_t},$$

for $r, t \geq 1$ and $1/r + 1/t = 1$, and the second term in Lemma 3.2.5 is of order

$$O\left(\frac{n^{-A/2+S/2}}{n_{\mathrm{te}}^{1/2}}\right) \max_{\xi \in \Xi} \|\widehat{\sigma}_\xi\|_{L_2}.$$

It is worth mentioning that one can also derive suitable growth rates on $\widehat{\kappa}_\Xi$ that yield conditions for $\mathbb{E}[\Delta_n^{\mathrm{mul}}] = o(1)$. However, this does not directly lead to control of $\mathbb{E}[R(\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n))]$ in the multiplicative form (3.8). This is because of the denominator $(1 - \Delta_n^{\mathrm{mul}})_+$ appearing in (3.8). For every n, there is a non-zero probability that the denominator $(1 - \Delta_n^{\mathrm{mul}})_+$ is zero. Hence, the right hand side of (3.8) may not have a finite expectation in general. However, assuming $\mathbb{E}[R(\widehat{f}^\xi(\cdot; \mathcal{D}_n))] < C$ for some $C < \infty$ for all $\xi \in \Xi$, one can control $\mathbb{E}[R(\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n))]$ by explicitly analyzing $\mathbb{P}(\Delta_n^{\mathrm{mul}} > 1/2)$, and using the bound

$$R(\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n)) \leq \frac{1 + \Delta_n^{\mathrm{mul}}}{(1 - \Delta_n^{\mathrm{mul}})_+} \cdot \min_{\xi \in \Xi} R(\widehat{f}^\xi(\cdot; \mathcal{D}_{\mathrm{tr}}).\mathbb{I}_{\Delta_n^{\mathrm{mul}} \leq 1/2} + \sum_{\xi \in \Xi} R(\widehat{f}^\xi(\cdot; \mathcal{D}_n))\mathbb{I}_{\Delta_n^{\mathrm{mul}} > 1/2}.$$

C.6.10 A lemma on norm equivalence implications

The following lemma formalizes various norm equivalence implications mentioned in Remarks 3.2.19 and 3.2.20.

Proposition C.6.21 (Norm equivalence implications)**.** *The following statements hold.*

1. *Suppose a random X satisfies $L_4 - L_2$ equivalence, i.e., there exists a constant C such that $\mathbb{E}[X^4] \leq C\mathbb{E}[X^2]$, then the random variable satisfies $L_2 - L_1$ equivalence, i.e., there exists a constant C such that $\mathbb{E}[X^2] \leq C\mathbb{E}[|X|]$.*

2. *A random variable W satisfying $\psi_2 - L_2$ equivalence also satisfies $\psi_1 - L_1$ equivalence.*

Proof. We will use the fact that the map $p \mapsto \log \mathbb{E}[|X|^p]$ $(p \geq 1)$ is convex. In other words, for $\lambda \in (0,1)$, we have

$$\log \mathbb{E}[|X|^{\lambda r + (1-\lambda)s}] \leq \lambda \log \mathbb{E}[|X|^r] + (1-\lambda) \log \mathbb{E}[|X|^s]. \tag{C.134}$$

We now use $r = 4$ and $s = 1$, and $\lambda = 1/3$ so that $\lambda r + (1-\lambda)s = 2$. Plugging these choices in (C.134) yields

$$\log \mathbb{E}[X^2] \leq \frac{1}{3} \log \mathbb{E}[X^4] + \frac{2}{3} \log \mathbb{E}[|X|].$$

In terms of norms the inequality then becomes

$$2 \log \|X\|_{L_2} \leq \frac{4}{3} \log \|X\|_{L_4} + \frac{2}{3} \log \|X\|_{L_1}.$$

This yields

$$\frac{2}{3} \log \frac{\|X\|_{L_2}}{\|X\|_{L_1}} \leq \frac{4}{3} \log \frac{\|X\|_{L_4}}{\|X\|_{L_2}}.$$

Manipulating both sides, we end up with

$$\frac{\|X\|_{L_2}}{\|X\|_{L_1}} \leq \left(\frac{\|X\|_{L_4}}{\|X\|_{L_2}} \right)^2$$

as desired.

The second facts follows because $\psi_2 - L_2$ equivalence implies $L_p - L_2$ equivalence for each $p \geq 1$, i.e., for each $p \geq 1$, we have that

$$\|W\|_{L_p} \leq C\sqrt{p} \|W\|_{L_2},$$

for an universal constant C; see Vershynin (2018, Proposition 2.5.2), for example. This in particular implies, $L_4 - L_2$ equivalence, and by the first fact implies $L_2 - L_1$. Thus, there exists a universal constant C such that

$$\|W\|_{L_2} \leq \|W\|_{L_1}.$$

Combining with the inequality above, we then get for $p \geq 1$,

$$\|W\|_{L_p} \leq C\sqrt{p} \|W\|_{L_1} \leq Cp \|W\|_{L_1}.$$

Now, using Vershynin (2018, Proposition 2.7.1), this implies $\psi_1 - L_1$ equivalence.

Alternatively, assuming $\psi_2 - L_2$ equivalence, observe the following chain of inequalities:

$$C\|X\|_{L_4} \overset{(a)}{\leq} \|X\|_{\psi_1} \overset{(b)}{\leq} (\log 2)^{1/2} \|X\|_{\psi_2} \overset{(c)}{\leq} C\|X\|_{L_2}$$

where (a) follows from Vershynin (2018, Proposition 2.5.2), (b) follows from Wellner and van der Vaart (2013, Problem 2.2.5), (c) follows from the assumed $\psi_2 - L_2$ equivalence. Finally, since $\psi_2 - L_2$ equivalence implies $L_4 - L_2$ equivalence, and from the fact this implies $L_2 - L_1$ equivalence concludes the proof.

Figure C.7 visually summarizes the norm equivalence implications. □

C.6.11 Proof of (3.63)

Below we prove the risk decomposition (3.63) for the ingredient zero-step predictor under squared error loss. The proof follows from the following iterated bias-variance decomposition.

$$\begin{aligned}
&\mathbb{E}\big[(Y_0 - \widetilde{f}_M(X_0; \mathcal{D}_{\mathrm{tr}}))^2 \mid \mathcal{D}_{\mathrm{tr}}\big] \\
&= \mathbb{E}\Big[\mathbb{E}\big[(Y_0 - \widehat{f}_M(X_0; \mathcal{D}_{\mathrm{tr}}))^2 \mid \mathcal{D}_{\mathrm{tr}}, (X_0, Y_0)\big] \mid \mathcal{D}_{\mathrm{tr}}\Big]
\end{aligned}$$

$$\begin{tikzcd}
\boxed{\psi_2 - L_2} \arrow[r, Rightarrow] \arrow[dd, Rightarrow] & \boxed{\psi_1 - L_2} \arrow[dd, Rightarrow] \arrow[dr, Rightarrow] & \\
& & \boxed{\psi_1 - L_1} \\
\boxed{L_4 - L_2} \arrow[r, Rightarrow] & \boxed{L_2 - L_1} \arrow[ur, Rightarrow] &
\end{tikzcd}$$

Figure C.7: Visual illustration of norm equivalence implications discussed in Remarks 3.2.19 and 3.2.20, and in the proof of Proposition C.6.21. In the figure, $\boxed{A} \Rightarrow \boxed{B}$ indicates that equivalence A implies equivalence B.

$$\begin{aligned}
&= \mathbb{E}\Big[\Big(Y_0 - \mathbb{E}\big[\widetilde{f}_M(X_0;\mathcal{D}_{\mathrm{tr}}) \mid \mathcal{D}_{\mathrm{tr}}, (X_0, Y_0)\big]\Big)^2 \mid \mathcal{D}_{\mathrm{tr}}\Big] + \mathbb{E}\Big[\mathrm{Var}\big(\widetilde{f}_M(X_0;\mathcal{D}_{\mathrm{tr}}) \mid \mathcal{D}_{\mathrm{tr}}, (X_0, Y_0)\big) \mid \mathcal{D}_{\mathrm{tr}}\Big] \\
&= \mathbb{E}\left[\left(Y_0 - \frac{1}{\binom{n}{k_n}} \sum_{i_1,\dots,i_{k_n}} \widetilde{f}\big(X_0; \{(X_{i_j}, Y_{i_j}) : 1 \le j \le k_n\}\big)\right)^2 \,\middle|\, \mathcal{D}_{\mathrm{tr}}\right] \\
&\quad + \mathbb{E}\left[\frac{1}{M}\mathrm{Var}\left(\widetilde{f}(X_0;\mathcal{D}_{\mathrm{tr},1}) \mid \mathcal{D}_{\mathrm{tr}}, (X_0, Y_0)\right) \,\middle|\, \mathcal{D}_{\mathrm{tr}}\right] \\
&= R(\widetilde{f}_\infty(\cdot;\mathcal{D}_{\mathrm{tr}})) + \frac{1}{M}\mathbb{E}\left[\frac{1}{\binom{n}{k_n}} \sum_{i_1,\dots,i_{k_n}} \left(\widetilde{f}\big(X_0; \{(X_0, Y_0) : 1 \le j \le k_n\}\big) - \widetilde{f}_\infty(X_0;\mathcal{D}_{\mathrm{tr}})\right)^2 \,\middle|\, \mathcal{D}_{\mathrm{tr}}\right],
\end{aligned}$$

where in the last line $f_\infty(\cdot;\mathcal{D}_{\mathrm{tr}}) : \mathbb{R}^p \to \mathbb{R}$ is defined such that for any $x \in \mathbb{R}^p$

$$\widetilde{f}_\infty(x;\mathcal{D}_{\mathrm{tr}}) = \frac{1}{\binom{n}{k_n}} \sum_{1 \le i_1 < \dots < i_{k_n} \le n_{\mathrm{tr}}} \widetilde{f}(x; \{(X_{i_j}, Y_{i_j}) : 1 \le j \le k_n\}).$$

C.7 Calculus of deterministic equivalents

We use the language of deterministic equivalents in the proofs of Proposition 3.3.14 and Proposition 3.4.11 in Appendix C.3 and Appendix C.5, respectively. In this section, we provide a basic review of the definitions and useful calculus rules. For more details, see Dobriban and Sheng (2021).

Definition C.7.1. Consider sequences $\{A_p\}_{p\ge1}$ and $\{B_p\}_{p\ge1}$ of (random or deterministic) matrices of growing dimension. We say that A_p and B_p are equivalent and write $A_p \simeq B_p$ if $\lim_{p\to\infty} |\operatorname{tr}[C_p(A_p - B_p)]| = 0$ almost surely for any sequence C_p matrices with bounded trace norm such that $\limsup \|C_p\|_{\mathrm{tr}} < \infty$ as $p \to \infty$.

An observant reader will notice that Dobriban and Sheng (2021) use the notation $A_p \asymp B_p$ to denote deterministic asymptotic equivalence. In this work, we instead prefer to use the notation $A_p \simeq B_p$ for such equivalence to stress the fact that this equivalence is exact in the limit rather than up to constants as the "standard" use of the asymptotic notation $\asymp$ would hint at.

Lemma C.7.2 (Calculus of deterministic equivalents, Dobriban and Wager (2018), Dobriban and Sheng (2021))**.** *Let A_p, B_p, and C_p be sequences of (random or deterministic) matrices. The calculus of deterministic equivalents satisfy the following properties:*

1. *Equivalence: The relation $\simeq$ is an equivalence relation.*

2. *Sum: If $A_p \simeq B_p$ and $C_p \simeq D_p$, then $A_p + C_p \simeq B_p + D_p$.*

3. *Product: If A_p a sequence of matrices with bounded operator norms, i.e., $\|A_p\|_{\mathrm{op}} < \infty$, and $B_p \simeq C_p$, then $A_p B_p \simeq A_p C_p$.*

4. *Trace: If $A_p \simeq B_p$, then $\operatorname{tr}[A_p]/p - \operatorname{tr}[B_p]/p \to 0$ almost surely.*

5. *Differentiation: Suppose $f(z, A_p) \simeq g(z, B_p)$ where the entries of f and g are analytic functions in $z \in S$ and S is an open connected subset of $\mathbb{C}$. Suppose for any sequence C_p of deterministic matrices with bounded trace norm we have $|\operatorname{tr}[C_p(f(z, A_p) - g(z, B_p))]| \le M$ for every p and $z \in S$. Then we have $f'(z, A_p) \simeq g'(z, B_p)$ for every $z \in S$, where the derivatives are taken entry-wise with respect to z.*

We record deterministic equivalent for the standard ridge resolvent.

Lemma C.7.3 (Deterministic equivalent for basic ridge resolvent, adapted from Theorem 1 of Rubio and Mestre (2011); see also Theorem 3.1 of Dobriban and Sheng (2021))**.** *Suppose $X_i \in \mathbb{R}^p$, $1 \le i \le n$, are i.i.d. random vectors where each $X_i = Z_i \Sigma^{1/2}$, where Z_i contains i.i.d. entries Z_{ij}, $1 \le j \le p$, with $\mathbb{E}[Z_{ij}] = 0$, $\mathbb{E}[Z_{ij}^2] = 1$, and $\mathbb{E}[|Z_{ij}|^{8+\alpha}] \le M_\alpha$ for some $\alpha > 0$ and $M_\alpha < \infty$, and $\Sigma \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix such that $0 \preceq \Sigma \preceq r_{\max} I_p$ for some constant (independent of p) $r_{\max} < \infty$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the matrix with X_i, $1 \le i \le n$ as rows and $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p \times p}$ denote the random matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$. Define $\gamma_n = p/n$. Then, for $z \in \mathbb{C}^{>0}$, as $n, p \to \infty$ such that $0 < \liminf \gamma_n \le \limsup \gamma_n < \infty$, we have*

$$(\widehat{\boldsymbol{\Sigma}} - zI_p)^{-1} \simeq (c(e(z;\gamma_n))\Sigma - zI_p)^{-1}, \tag{C.135}$$

where $c(e(z;\gamma_n))$ is defined as

$$c(e(z;\gamma_n)) = \frac{1}{1 + \gamma_n e(z;\gamma_n)}, \tag{C.136}$$

and $e(z;\gamma_n)$ is the unique solution in $\mathbb{C}^{>0}$ to the fixed-point equation

$$e(z;\gamma_n) = \operatorname{tr}[\Sigma(c(e(z;\gamma_n))\Sigma - zI_p)^{-1}]/p. \tag{C.137}$$

Furthermore, $e(z;\gamma_n)$ is the Stieltjes transform of a certain positive measure on $\mathbb{R}_{\ge 0}$ with total mass $\operatorname{tr}[\Sigma]/p$.

We note that in defining $e(\lambda;\gamma_n)$, it is also implicitly a parameterized by Σ. We suppress this dependence for notational simplicity, and only explicitly indicate dependence on z and γ_n that will be useful for our purposes.

Corollary C.7.4. *Assume the setting of Lemma C.7.3. For $\lambda > 0$, we have*

$$\lambda(\widehat{\boldsymbol{\Sigma}} + \lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma_n)} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}]/p.$$

Proof. From Lemma C.7.3, for $z \in \mathbb{C}^{>0}$, we have the basic equivalence for ridge resolvent

$$(\widehat{\boldsymbol{\Sigma}} - zI_p)^{-1} \simeq (c(e(z;\gamma_n))\Sigma - zI_p)^{-1}, \tag{C.138}$$

where $c(e(z;\gamma_n))$ is defined by (C.136) and and $e(z;\gamma_n)$ is the unqiue solution in $\mathbb{C}^{>0}$ to the fixed-point equation (C.137). Substituting for $e(z;\gamma_n)$ from (C.136) into (C.137), we can write the fixed-point equation for $c(e(z;\gamma_n))$ as

$$\frac{1}{c(e(z;\gamma_n))\gamma_n} - \frac{1}{\gamma_n} = \operatorname{tr}[\Sigma(c(e(z;\gamma_n))\Sigma - zI_p)^{-1}]/p. \tag{C.139}$$

Manipulating (C.139), we can write

$$\frac{1}{c(e(z;\gamma_n))} - 1 = \gamma_n \operatorname{tr}[\Sigma(c(e(z;\gamma_n))\Sigma - zI_p)^{-1}]/p = \frac{\gamma_n}{(-z)} \operatorname{tr}[\Sigma(c(e(z;\gamma_n))/(-z)\Sigma + I_p)^{-1}]/p. \tag{C.140}$$

Moving $(-z)$ across in (C.140), we have equivalently the following equation for $c(e(z;\gamma_n))$:

$$\frac{(-z)}{c(e(z;\gamma_n))} + z = \gamma_n \operatorname{tr}[\Sigma(c(e(z;\gamma_n))/(-z)\Sigma + I_p)^{-1}]/p. \tag{C.141}$$

Now defining $c(e(z;\gamma_n))/(-z)$ by $v(z;\gamma_n)$, the fixed-point equation (C.141) becomes

$$\frac{1}{v(z;\gamma_n)} = -z + \gamma_n \operatorname{tr}[\Sigma(v(z;\gamma_n)\Sigma + I_p)^{-1}]/p. \tag{C.142}$$

Note that (C.142) is also known as the Silverstein equation (Silverstein, 1995), and $v(z;\gamma_n)$ as the companion Stieltjes transform. Along the same lines, from (C.138), we have

$$(-z)(\widehat{\boldsymbol{\Sigma}} - zI_p)^{-1} \simeq (-z)(c(e(z;\gamma_n))\Sigma - zI_p)^{-1} = (c(e(z;\gamma_n))/(-z)\Sigma + I_p)^{-1}. \tag{C.143}$$

Substituting for $v(z;\gamma_n)$, we can thus write

$$(-z)(\widehat{\boldsymbol{\Sigma}} - zI_p)^{-1} \simeq (v(z;\gamma_n)\Sigma + I_p)^{-1}. \tag{C.144}$$

Now, taking $z = -\lambda$ in (C.142) and (C.144) yields the equivalence

$$\lambda(\widehat{\boldsymbol{\Sigma}} + \lambda I_p)^{-1} \simeq (v(-\lambda;\gamma_n)\Sigma + I_p)^{-1},$$

where $v(-\lambda;\gamma_n)$ is the unique solution to the fixed point equation

$$\frac{1}{v(-\lambda;\gamma_n)} = \lambda + \gamma_n \operatorname{tr}[\Sigma(v(-\lambda;\gamma_n)\Sigma + I_p)^{-1}]/p.$$

Finally, since $v(-\lambda;\gamma_n)$ is a Stieltjes transform of a probability measure (with support on $\mathbb{R}_{\geq 0}$), we have that for $\operatorname{Re}(\lambda) > 0$, by taking $\operatorname{Im}(\lambda) \to 0$, we have that $\operatorname{Im}(v(-\lambda;\gamma_n)) \to 0$, and thus the statement follows. □

We remark that we will directly apply Corollary C.7.4 for a real $\lambda > 0$ (in particular, in Lemma C.6.10). The limiting argument to go from a complex λ to a real λ follow as done in the proof of Corollary C.7.4. See, for example, proof of Theorem 5 in Hastie et al. (2022) (that uses Lemma 2.2 of Knowles and Yin (2017)) for more details.

C.8 Useful results

In this section, we gather statements of concentration results available in the literature that are used in the proofs in Appendices C.1, C.3 and C.5.

Non-asymptotic statements

Tail bounds. The following two tail bounds are used in the proofs of Lemmas 3.2.4, 3.2.5, 3.2.9 and 3.2.10 in Appendix C.1.

Lemma C.8.1 (Bernstein's inequality, adapted from Theorem 2.8.1 of Vershynin (2018))**.** *Let $Z_1, \ldots, Z_n$ be independent mean-zero sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n Z_i\right| \geq t\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^n \|Z_i\|_{\psi_1}^2}, \frac{t}{\max_{1\leq i\leq n}\|Z_i\|_{\psi_1}}\right\}\right),$$

where $c > 0$ is an absolute constant. In other words, with probability at least $1-\eta$, we have

$$\left|\sum_{i=1}^n Z_i\right| \leq \max\left\{\sqrt{\frac{1}{c}\sum_{i=1}^n \|Z_i\|_{\psi_1}^2 \log\left(\frac{2}{\eta}\right)}, \frac{1}{c}\max_{1\leq i\leq n}\|Z_i\|_{\psi_1}\log\left(\frac{2}{\eta}\right)\right\}.$$

Lemma C.8.2 (Concentration for median-of-means (MOM) estimator, adapted from Theorem 2 of Lugosi and Mendelson (2019))**.** *Let $W_1, \dots, W_n$ be i.i.d. random variables with mean μ and variance bounded by σ^2. Suppose we split the data $\{W_1, \dots, W_n\}$ into B batches $\mathcal{T}_1, \dots, \mathcal{T}_B$. Let $\widehat{\mu}_b$ be sample mean computed on $\mathcal{T}_b$ for $b = 1, \dots, B$. Define*

$$\widehat{\mu}_B^{\mathit{MOM}} := \mathrm{median}(\widehat{\mu}_1, \dots, \widehat{\mu}_B).$$

Then, we have

$$\mathbb{P}\left\{ \left| \widehat{\mu}_B^{\mathit{MOM}} - \mu \right| > \sigma\sqrt{4B/n} \right\} \le \exp(-B/8).$$

Thus, letting $0 < \eta < 1$ be a real number, $B = \lceil 8\log(1/\eta) \rceil$, with probability at least $1 - \eta$,

$$\left| \widehat{\mu}_B^{\mathit{MOM}} - \mu \right| \le \sigma\sqrt{\frac{32\log(1/\eta)}{n}}.$$

With $B = \lceil 8\log(1/\eta) \rceil$, we use the notation $\mathtt{MOM}(\{W_1, \dots, W_n\}, \eta)$ for $\widehat{\mu}_B^{\mathtt{MOM}}$, that is,

$$\mathtt{MOM}(\{W_1, \dots, W_n\}, \eta) \ := \ \widehat{\mu}^{\mathtt{MOM}}_{\lceil 8\log(1/\eta) \rceil}. \tag{C.145}$$

Moment bounds. The following two moment bounds imply Lemmas C.8.5 and C.8.6 that are used in the proofs of Proposition 3.3.14 and Corollary 3.4.9 in Appendix C.3 and Appendix C.5, respectively.

Lemma C.8.3 (Moment bound on centered linear form, adapted from Lemma 7.8 of Erdos and Yau (2017))**.** *Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a random vector containing i.i.d. entries Z_i, $i = 1, \dots, n$, such that for each i, $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i^2] = 1$, and $\mathbb{E}[|Z_i|^k] \le M_k$. Let $a \in \mathbb{R}^p$ be a deterministic vector. Then,*

$$\mathbb{E}[|a^\top \boldsymbol{Z}|^q] \le C_q M_q \|a\|_2^q$$

for a constant C_q that only depends on q.

Lemma C.8.4 (Moment bound on centered quadratic form, adapted from Lemma B.26 of Bai and Silverstein (2010))**.** *Let $\boldsymbol{Z} \in \mathbb{R}^n$ be a random vector with i.i.d. entries Z_i, $i = 1, \dots, n$, such that for each i, $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i^2] = 1$, and $\mathbb{E}[|Z_i|^k] \le M_k$ for $k > 2$ and some constant M_k. Let $A \in \mathbb{R}^{p \times p}$ be a deterministic matrix. Then, for $q \ge 1$,*

$$\mathbb{E}\left[|\boldsymbol{Z}^\top A \boldsymbol{Z} - \mathrm{tr}[A]|^q \right] \le C_q \left\{ (M_4 \,\mathrm{tr}[AA^\top])^{q/2} + M_{2q} \,\mathrm{tr}[(AA^\top)^{q/2}] \right\}$$

for a constant C_q that only depends on q.

Asymptotic statements

As a consequence of Lemma C.8.3 and Lemma C.8.7, we have the following concentration of a linear form with independent components.

Lemma C.8.5 (Concentration of linear form with independent components)**.** *Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a random vector with i.i.d. entries Z_i, $i = 1, \dots, p$ such that for each i, $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[|Z_i|^{4+\alpha}] \le M_\alpha$ for some constant $M_\alpha < \infty$. Let $\boldsymbol{A} \in \mathbb{R}^p$ be a random vector independent of $\boldsymbol{Z}$ such that $\limsup_p \|\boldsymbol{A}_p\|^2/p \le M_n$ almost surely for a constant $M_n < \infty$. Then, $\boldsymbol{A}^\top \boldsymbol{Z}/p \to 0$ almost surely as $p \to \infty$.*

As a consequence of Lemma C.8.4 and Lemma C.8.7, we have the following concentration of a quadratic form with independent components.

Lemma C.8.6 (Concentration of quadratic form with independent components)**.** *Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a random vector with i.i.d. entries Z_i, $i = 1, \dots, p$ such that for each i, $\mathbb{E}[Z_i] = 0$, $\mathbb{E}[Z_i^2] = 1$, $\mathbb{E}[|Z_i|^{4+\alpha}] \le M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ be a random matrix such that $\limsup \|\boldsymbol{D}\|_{op} \le M_o$ almost surely as $p \to \infty$ for some constant $M_o < \infty$. Then, $\boldsymbol{Z}^\top \boldsymbol{D} \boldsymbol{Z}/p - \mathrm{tr}[\boldsymbol{D}]/p \to 0$ almost surely as $p \to \infty$.*

Lemma C.8.7 (Moment version of the Borel-Cantelli lemma)**.** *Let $\{Z_n\}_{n \ge 1}$ be a sequence of real-valued random variables such that the sequence $\{\mathbb{E}|Z_n|^q\}_{n \ge 1}$ is summable for some $q > 0$. Then, $Z_n \to 0$ almost surely as $n \to \infty$.*

C.9 Notation

Below we list general notation used in this work.

- We denote scalar random variables in regular upper case (e.g., X), and vector and matrix random variables in bold upper case (e.g., $\boldsymbol{X}$). We use calligraphic letters to denote sets (e.g., $\mathcal{D}$), and blackboard letters to denote some specials sets listed next.
- We use $\mathbb{N}$ to denote the set of natural numbers. We use $\mathbb{Q}$ to denote the set of rational numbers, $\mathbb{Q}_{>0}$ to denote the set of positive rational numbers; $\mathbb{R}$ to denote the set of real numbers, $\mathbb{R}_{\geq 0}$ to denote the set of non-negative real numbers, $\mathbb{R}_{>0}$ to denote the set of positive real numbers; $\mathbb{C}$ to denote the set of complex numbers, $\mathbb{C}^{>0}$ to denote the upper half of the complex plane, i.e., $\mathbb{C}^{>0} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$.
- For a real number a, $(a)_+$ denotes its positive part, $\lfloor a \rfloor$ denotes its floor, $\lceil a \rceil$ denotes its ceiling, $\mathrm{sgn}(a)$ denotes its sign. For a complex number z, $\mathrm{Re}(z)$ denotes its real part, $\mathrm{Im}(z)$ denotes its imaginary part, $\overline{z}$ denote its conjugate, $|z|$ denotes its absolute value.
- For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality, $\mathcal{A}^{\complement}$ denotes its complement, $\mathbb{I}_{\mathcal{A}}$ denotes its indicator function. For a function f, $\partial/\partial x[f]$ denotes its partial derivative with respect to variable x. We also use f' to denote derivative of f when it is clear from the context.
- For an event A, $\mathbb{P}(A)$ denotes its probability, and $\mathbb{I}_A$ its indicator random variable. For a random variable X, $\mathbb{E}[X]$ denotes its expectation, $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$ denotes its variance; $\mathbb{E}[X^r]$ denotes its r-th moment, $\mathbb{E}[|X|^r]$ denotes its r-th absolute moment, $\|X\|_{L_r} = (\mathbb{E}[|X|^r])^{1/r}$ denotes its L_r norm, for a real number $r \geq 1$; $\|X\|_\psi$ denotes its ψ norm for an Orlicz function ψ; see Section 3.2.2 for more details.
- For a vector $a \in \mathbb{R}^p$, $\|a\|_r$ denotes its ℓ_r norm for $r \geq 1$, $\|a\|_A = \sqrt{a^\top A a}$ denotes its norm with respect to a positive semidefinite matrix $A \in \mathbb{R}^{p \times p}$.
- For a matrix $A \in \mathbb{R}^{n \times p}$, $A^\top \in \mathbb{R}^{p \times n}$ denote its transpose, $A^\dagger \in \mathbb{R}^{p \times n}$ denotes the its Moore-Penrose inverse, $\|A\|_{\mathrm{op}}$ denotes its operator norm, $\|A\|_{\mathrm{tr}}$ denotes its trace norm or nuclear norm ($\|A\|_{\mathrm{tr}} = \mathrm{tr}[(A^\top A)^{1/2}] = \sum_i \sigma_i(A)$), where $\sigma_1(A) \geq \sigma_2(A) \geq \dots$ denote its singular values in non-increasing order. For a square matrix $A \in \mathbb{R}^{p \times p}$, $\mathrm{tr}[A] = \sum_{i=1}^p A_{ii}$ denotes its trace. A p-dimensional identity matrix is denoted as I_p or simply I when it is clear from the context.
- For a $p \times p$ positive semidefinite matrix A with eigenvalue decomposition $A = VRV^\top$ for an orthonormal matrix V and a diagonal matrix R, and a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we denote by $f(A)$ the $p \times p$ positive semidefinite matrix $Vf(R)V^\top$, where $f(R)$ is a $p \times p$ diagonal matrix obtained by applying the function f to each diagonal entry of R.
- For two sequences of matrices A_n and B_n, we use the notation $A_n \simeq B_n$ to denote a certain notion of asymptotic equivalence; see Appendix C.7 for more details. For symmetric matrices A and B, $A \preceq B$ denotes the Loewner ordering to mean that the matrix $B - A$ is positive semidefinite.
- We write $a \asymp b$ when there exist absolute constants C_l and C_u such that $C_l \leq a/b \leq C_u$. We write $a \lesssim b$ when there exists an absolute constant C such that $a \leq Cb$.
- We use O and o to denote the big-O and little-o asymptotic notation, respectively. We use O_p and o_p to denote the probabilistic big-O and little-o asymptotic notation, respectively. We denote convergence in probability by $\xrightarrow{\mathrm{p}}$, almost sure convergence by $\xrightarrow{\mathrm{a.s.}}$, weak convergence by $\xrightarrow{\mathrm{d}}$.
- Finally, we use generic letters $C, C_1, C_2, \dots$ to denote constants whose value may change from line to line.

Notation	Meaning (Location)
(X, Y)	feature vector $X \in \mathbb{R}^p$ and response variable $Y \in \mathbb{R}$ (Section 3.2.1)
$\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$	dataset with n observations (X_i, Y_i), $1 \le i \le n$ (Section 3.2.1)
$\widehat{f}(\cdot; \mathcal{D}_n) : \mathbb{R}^p \to \mathbb{R}$	predictor fitted on dataset $\mathcal{D}_n$ using prediction procedure $\widehat{f}$ (Section 3.2.1)
$\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\ge 0}$	non-negative loss function (Section 3.2.1)
$\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n))$	prediction loss of predictor $\widehat{f}(\cdot; \mathcal{D}_n)$ evaluated at test point (X_0, Y_0) (Section 3.2.1)
$R(\widehat{f}(\cdot; \mathcal{D}_n))$	prediction risk of predictor $\widehat{f}(\cdot; \mathcal{D}_n)$ (3.5)
$\widehat{R}(\widehat{f}(\cdot; \mathcal{D}_n))$	estimator of prediction risk of $\widehat{f}(\cdot; \mathcal{D}_n)$ (Section 3.2.1)
$\widehat{f}^{\mathrm{cv}}(\cdot; \mathcal{D}_n)$	cross-validated predictor fitted using dataset $\mathcal{D}_n$ (Algorithm 1)
$\widehat{f}^{\xi}$, $\xi \in \Xi$	collection of prediction procedures indexed by set Ξ (Algorithm 1)
n_{tr}, n_{te}	number of train and test observations (Algorithm 1)
$\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{te}}$	random split of $\mathcal{D}_n$ into train and test datasets with n_{tr} and n_{te} observations (Algorithm 1)
$\mathcal{I}_{\mathrm{tr}}, \mathcal{I}_{\mathrm{te}}$	disjoint subsets of $\mathcal{I}_n := \{1, \dots, n\}$ that are index sets for $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$ (Algorithm 1)
`CEN`, `AVG`, `MOM`	centering procedure, averaging, median-of-means (3.2, 3.3)
η	parameter in median-of-means (C.145)
Δ_n^{add}, Δ_n^{mul}	error terms in the additive and multiplicative oracle risk inequalities (3.6a, 3.6b)
$\widehat{\sigma}_{\xi}$, $\widehat{\sigma}_{\Xi}$	conditional second moment of loss and their max over Ξ (Lemmas 3.2.4 and 3.2.5)
$\widehat{\kappa}_{\xi}$, $\widehat{\kappa}_{\Xi}$	conditional kurtosis-like parameter of loss and their max over Ξ (Lemmas 3.2.9 and 3.2.10)
$\\|\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n))\\|_{\psi_1 \mid \mathcal{D}_n}$	conditional ψ_1 norm of prediction loss (3.9)
$\\|\ell(Y_0, \widehat{f}(X_0; \mathcal{D}_n))\\|_{L_r \mid \mathcal{D}_n}$	conditional L_r norm of prediction loss ($r \ge 1$) (3.10)
$\widetilde{\beta}_{\mathrm{ridge}}$, $\widetilde{\beta}_{\mathrm{lasso}}$, $\widetilde{\beta}_{\mathrm{mn2}}$, $\widetilde{\beta}_{\mathrm{mn1}}$	ridge, lasso, min ℓ_2, ℓ_1-norm least squares estimation procedures (3.20–3.24)
$\widetilde{f}_{\mathrm{mn2}}$, $\widetilde{f}_{\mathrm{mn1}}$	min ℓ_2, ℓ_1-norm least squares prediction procedures (3.22, 3.25)
$\widehat{f}^{\mathrm{zs}}(\cdot; \mathcal{D}_n)$	zero-step predictor fitted on dataset $\mathcal{D}_n$ (Algorithm 2)
$\nu \in (0, 1)$	exponent for block sizes $\lfloor n^{\nu} \rfloor$ in zero-step prediction procedure (Algorithm 2)
n_{ξ}	$n - \xi \lfloor n^{\nu} \rfloor$ (Algorithm 2)
M	number of sub-samples for averaging for zero-step ingredient predictor (3.26)
$\mathcal{D}_{\mathrm{tr}}^{\xi, j}$, $1 \le j \le M$	random subset of $\mathcal{D}_{\mathrm{tr}}$ of size n_{ξ} (Algorithm 2)
$\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi, j})$	zero-step ingredient predictor fitted on dataset $\mathcal{D}_{\mathrm{tr}}^{\xi, j}$ using base prediction procedure $\widetilde{f}$ (3.26)
$R^{\mathrm{det}}(m; \widetilde{f})$	deterministic approximation to $R(\widetilde{f}(\cdot; \mathcal{D}_m))$ (Definition 3.3.2)
$R^{\mathrm{det}}_{\nearrow}(n; \widetilde{f})$	monotonized deterministic approximation at sample size n under general asymptotics (3.30)
$\mathrm{PA}(\gamma)$	proportional asymptotics regime ($\mathrm{PA}(\gamma)$)
DETPA-0	assumption of deterministic risk approximation to conditional risk under PA (DETPA-0)
DETPAR-0	reduction of assumption DETPA-0 (Lemma 3.3.8, DETPAR-0)
$R^{\mathrm{det}}(p_m/m; \widetilde{f})$	deterministic risk approximation at aspect ratio p_m/m under PA (Section 3.3.3.1)
$\xi_n^{\star}$	optimal sequence of ξ for zero-step monotonized risk approximation (3.30, DETPA-0)
PRG-0-C1,C2	deterministic risk approximation program for zero-step (PRG-0-C1)–(PRG-0-C2)
k_m, p_m	sample size and feature size when verifying zero-step profile assumption (Lemma 3.3.8)
ρ^2, σ^2, SNR	signal energy, noise energy, signal-to-noise ratio (ρ^2/σ^2) (Section 3.3.4)
$R^{\mathrm{det}}_{\mathrm{mn2}}(\phi; \rho^2, \sigma^2)$	MN2LS risk approximation at aspect ratio ϕ, signal energy ρ^2, noise energy σ^2 (3.60)
$\widetilde{f}_{\infty}(\cdot; \mathcal{D}_{\mathrm{tr}})$	zero-step ingredient predictor fitted on $\mathcal{D}_n$ with $M = \infty$ (3.62)
$\widehat{f}^{\mathrm{os}}(\cdot; \mathcal{D}_n)$	one-step predictor fitted on dataset $\mathcal{D}_n$ (Algorithm 3)
$(n_{1,\xi_1}, n_{2,\xi_2})$	$(n - \xi_1 \lfloor n^{\nu} \rfloor, \xi_2 \lfloor n^{\nu} \rfloor)$ (Algorithm 3)
$(\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$, $1 \le j \le M$	random pairs of disjoint subsets of $\mathcal{D}_{\mathrm{tr}}$ of sizes $(n_{1,\xi_1}, n_{2,\xi_2})$ (Algorithm 3)
$\widetilde{f}(\cdot; \mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$	one-step ingredient predictor fitted on datasets $(\mathcal{D}_{\mathrm{tr}}^{\xi_1, j}, \mathcal{D}_{\mathrm{tr}}^{\xi_2, j})$ (3.43)
DETPA-1, DETPA-1*	assumption of deterministic risk approximation to conditional risk under PA (DETPA-1)
DETPAR-1	reduction of assumption DETPA-1 (Lemma 3.4.1, DETPAR-1)
$R^{\mathrm{det}}(p/n_1, p/n_2; \widetilde{f})$	risk approximation of ingredient one-step predictor at aspect ratios $(p/n_1, p/n_2)$ (Section 3.4.3.1)
$(\xi_{1,n}^{\star}, \xi_{2,n}^{\star})$	optimal pair of sequence of ξ for one-step monotonized risk approximation (3.45)
PRG-1-C1,C2,C3	deterministic risk approximation program for one-step (PRG-1-C1)–(PRG-1-C3)
$k_{1,m}, k_{2,m}, p_m$	sample size and feature sizes when verifying one-step profile assumption (Lemma 3.4.1)
w_i, r_i, $1 \le i \le p_m$	eigenvectors and eigenvalues of feature covariance matrix $\Sigma \in \mathbb{R}^{p_m \times p_m}$ (Section 3.4.3.2)
$\widehat{Q}_n$, Q	a certain random distribution and its weak limit (C.69)
H_{p_m}, H	empirical distribution of eigenvalues of Σ and limiting spectral distribution (3.53)
$v(0; \phi_2), \widetilde{v}(0; \phi_2), \widetilde{v}_g(0; \phi_2), \Upsilon_b(\phi_1, \phi_2)$	scalars in risk approximation of one-step procedure with linear base procedure (3.55–3.58)
$R^{\mathrm{det}}_{\mathrm{mn2}}(\phi_1, \phi_2; \rho^2, \sigma^2)$	MN2LS one-step risk approx at aspect ratios (ϕ_1, ϕ_2), signal energy ρ^2, noise energy σ^2 (3.60)

Table C.1: Summary of some of the main notation used in this work.

Appendix D

Supplement for Chapter 4

D.1 Notation and organization

Notation

Below we provide an overview of some general notation used in this work.

We denote scalars in non-bold lower or upper case (e.g., n, λ, C), vectors in bold lower case (e.g., $\boldsymbol{x}$, $\boldsymbol{\beta}$), and matrices in bold upper case (e.g., $\boldsymbol{X}$). We denote sets using calligraphic letters (e.g., $\mathcal{D}$), and use blackboard letters to denote some special sets: $\mathbb{N}$ denotes the set of positive integers, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers, $\mathbb{R}_{>0}$ denotes the set of positive real numbers, $\mathbb{C}$ denotes the set of complex numbers, $\mathbb{C}^{+}$ denotes the set of complex numbers with positive imaginary part, and $\mathbb{C}^{-}$ denotes the set of complex numbers with negative imaginary part. For a natural number n, we use $[n]$ to denote the set $\{1, \dots, n\}$.

For a real number x, $(x)_{+}$ denotes its positive part, $\lfloor x \rfloor$ its floor, and $\lceil x \rceil$ its ceiling. For a vector $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2$ denotes its ℓ_2 norm. For a pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ denotes their inner product. For an event A, $\mathbb{1}_A$ denotes the associated indicator random variable. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{X}^{\top} \in \mathbb{R}^{p \times n}$ denotes its transpose, and $\boldsymbol{X}^{+} \in \mathbb{R}^{p \times n}$ denote its Moore-Penrose inverse. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\operatorname{tr}[\boldsymbol{A}]$ denotes its trace, and $\boldsymbol{A}^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided it is invertible. For a positive semidefinite matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{1/2}$ denotes its principal square root. A $p \times p$ identity matrix is denoted $\boldsymbol{I}_p$, or simply by $\boldsymbol{I}$, when it is clear from the context.

For a real matrix $\boldsymbol{X}$, its operator norm (or spectral norm) with respect to ℓ_2 vector norm is denoted by $\|\boldsymbol{X}\|_{\mathrm{op}}$, and its trace norm (or nuclear norm) is denoted by $\|\boldsymbol{X}\|_{\mathrm{tr}}$ (recall that $\|\boldsymbol{X}\|_{\mathrm{tr}} = \operatorname{tr}[(\boldsymbol{X}^{\top}\boldsymbol{X})^{1/2}]$). For a positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ with eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{R}\boldsymbol{V}^{-1}$ for an orthonormal matrix $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ and a diagonal matrix $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ with non-negative entries, and a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we denote by $f(\boldsymbol{A})$ the $p \times p$ positive semidefinite matrix $\boldsymbol{V} f(\boldsymbol{R}) \boldsymbol{V}^{-1}$. Here, $f(\boldsymbol{R})$ is a $p \times p$ diagonal matrix obtained by applying the function f to each diagonal entry of $\boldsymbol{R}$.

For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ denotes the Loewner ordering. For sequences of matrices $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$, $\boldsymbol{A}_n \simeq \boldsymbol{B}_n$ denotes a certain notion of asymptotic equivalence (see Definitions D.8.1 and D.8.2). We use O_p and o_p to denote probabilistic big-O and little-o notation, respectively. We denote convergence in probability by "$\xrightarrow{\text{p}}$", almost sure convergence by "$\xrightarrow{\text{a.s.}}$", and convergence in distribution by "$\xrightarrow{\text{d}}$".

Organization

Below we outline the structure of the rest of the supplement.

- In Appendix D.2, we present proofs of results related to general subagged predictors from Section 4.3.
- In Appendices D.3 and D.4, we present proof of Theorem 4.4.1 related to subagging from Section 4.4.2 for ridge and ridgeless predictors, respectively. The proofs for the two cases are separated due to length. However, the proof architecture for the two is similar.

- In Appendix D.5, we present proof of Theorem 4.4.6 related to splagging from Section 4.4.3 for ridge and ridgeless predictors. Because some of this proof builds on that of Theorem 4.4.1, we can combine the two cases of ridge and ridgeless predictors, unlike the split cases for Theorem 4.4.1.
- In Appendix D.6, we present proofs of results related to the bias-variance component monotonicity properties in Propositions 4.4.5 and 4.4.10 for subagging and splagging, respectively. In this section, we also provide proofs of results related to cross-validation and profile monotonicity and those related to oracle properties of optimized bagging from Section 4.5.
- In Appendix D.7, we present proofs of specialized results related to subagging and splagging under isotopic features from Section 4.6.
- In Appendix D.8, we formalize several calculus rules for a certain notion of conditional asymptotic equivalence of sequences of matrices that are used in the proofs of constituent lemmas in Appendices D.3 to D.5.
- In Appendix D.9, we collect various technical helper lemmas related to concentrations and convergences along with their proofs that are used in proofs in Appendices D.2 to D.5.
- In Appendix D.10, we present additional numerical illustrations for Theorems 4.4.1, 4.4.6 and 4.5.5, and for specialized isotropic results from Section 4.6.

D.2 Proofs in Section 4.3

D.2.1 Asymptotic data conditional risk, squared loss

Proof of Proposition 4.3.2. The key idea in the proof is to use the conditional risk decomposition from Proposition 4.3.1. Below we present the proof for sampling from $\mathcal{I}_k$. The proof for sampling from $\mathcal{I}_k^\pi$ is analogous.

SRSWR. We will do the case of SRSWR from $\mathcal{I}_k$ first. From Proposition 4.3.1, we have

$$\begin{aligned} R(\widetilde{f}_M;\mathcal{D}_n) &= \mathbb{E}_{(\boldsymbol{x},y)}[\mathbb{E}[(\widetilde{f}_M - y)^2 \mid \mathcal{D}_n, (\boldsymbol{x},y)]] \\ &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y) \,|\, \mathcal{D}_n\right] + \frac{1}{M}\mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y) \,|\, \mathcal{D}_n\right] \\ &= R(\widetilde{f}_\infty;\mathcal{D}_n) + \frac{1}{M}C_n, \end{aligned} \tag{D.1}$$

where $C_n = \mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x};\mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2 \,\middle|\, \mathcal{D}_n\right]$.

Since for $M = 1$ and $M = 2$, we have

$$\begin{aligned} R(\widetilde{f}_1;\mathcal{D}_n) &= R(\widetilde{f}_\infty;\mathcal{D}_n) + C_n, \\ R(\widetilde{f}_2;\mathcal{D}_n) &= R(\widetilde{f}_\infty;\mathcal{D}_n) + \frac{C_n}{2}. \end{aligned}$$

We can thus write $R(\widetilde{f}_\infty;\mathcal{D}_n)$ and C_n in terms of $R(\widetilde{f}^{\mathtt{WR}}_{1,\mathcal{I}_k};\mathcal{D}_n)$ and $R(\widetilde{f}^{\mathtt{WR}}_{2,\mathcal{I}_k};\mathcal{D}_n)$ as

$$\begin{aligned} R(\widetilde{f}_\infty;\mathcal{D}_n) &= 2R(\widetilde{f}_2;\mathcal{D}_n) - R(\widetilde{f}_1;\mathcal{D}_n), \\ C_n &= 2R(\widetilde{f}_1;\mathcal{D}_n) - 2R(\widetilde{f}_2;\mathcal{D}_n). \end{aligned}$$

Substituting in (D.1), we obtain

$$R(\widetilde{f}_M;\mathcal{D}_n) = 2R(\widetilde{f}_2;\mathcal{D}_n) - R(\widetilde{f}_1;\mathcal{D}_n) + \frac{1}{M}\left(2R(\widetilde{f}_1;\mathcal{D}_n) - 2R(\widetilde{f}_2;\mathcal{D}_n)\right)$$

$$= -\left(1 - \frac{2}{M}\right) R(\widetilde{f}_1; \mathcal{D}_n) + \left(2 - \frac{2}{M}\right) R(\widetilde{f}_2; \mathcal{D}_n).$$

Thus, subtracting the desired target in (4.13) for with replacement from both sides, we get

$$R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right] = -\left(1 - \frac{2}{M}\right) \left(R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right) + \left(2 - \frac{2}{M}\right) \left(R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right).$$

Taking absolute values on both sides and using triangle inequality yields

$$\left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \le \left|1 - \frac{2}{M}\right| \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + \left(2 - \frac{2}{M}\right) \left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Taking supremum over M, we have

$$\sup_{M \in \mathbb{N}} \left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \le \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + 2\left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Finally, since we have

$$R(\widetilde{f}_1; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_1, \qquad R(\widetilde{f}_2; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_2,$$

the desired claim in (4.13) for with replacement follows.

SRSWOR. For SRSWOR from $\mathcal{I}_k$, similarly we have

$$\begin{aligned} R(\widetilde{f}_M; \mathcal{D}_n) &= \mathbb{E}_{(\boldsymbol{x},y)}[\mathbb{E}[(\widetilde{f}_M - y)^2 | \mathcal{D}_n, (\boldsymbol{x}, y)]] \\ &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y) \mid \mathcal{D}_n\right] + \frac{|\mathcal{I}_k| - M}{|\mathcal{I}_k| - 1} \frac{1}{M} \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \mid \mathcal{D}_n\right] \\ &= R(\widetilde{f}_\infty; \mathcal{D}_n) + \frac{|\mathcal{I}_k| - M}{|\mathcal{I}_k| - 1} \frac{1}{M} C_n \\ &= R(\widetilde{f}_\infty; \mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k| - 1} + \frac{1}{M} \cdot \frac{|\mathcal{I}_k| C_n}{|\mathcal{I}_k| - 1}, \end{aligned} \tag{D.2}$$

where $C_n = \mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{|\mathcal{I}_k|}\sum_{I \in \mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2 \middle| \mathcal{D}_n\right]$. Since for $M = 1$ and $M = 2$,

$$\begin{aligned} R(\widetilde{f}_1; \mathcal{D}_n) &= R(\widetilde{f}_\infty; \mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k| - 1} + \frac{|\mathcal{I}_k| C_n}{|\mathcal{I}_k| - 1}, \\ R(\widetilde{f}_2; \mathcal{D}_n) &= R(\widetilde{f}_\infty; \mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k| - 1} + \frac{1}{2} \cdot \frac{|\mathcal{I}_k| C_n}{|\mathcal{I}_k| - 1}. \end{aligned}$$

We can thus write $R(\widetilde{f}_\infty; \mathcal{D}_n) - C_n/(|\mathcal{I}_k| - 1)$ and $|\mathcal{I}_k| C_n/(|\mathcal{I}_k| - 1)$ in terms of $R(\widetilde{f}^{\texttt{WR}}_{1,\mathcal{I}_k}; \mathcal{D}_n)$ and $R(\widetilde{f}^{\texttt{WR}}_{2,\mathcal{I}_k}; \mathcal{D}_n)$ as

$$\begin{aligned} R(\widetilde{f}_\infty; \mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k| - 1} &= 2R(\widetilde{f}_2; \mathcal{D}_n) - R(\widetilde{f}_1; \mathcal{D}_n), \\ \frac{|\mathcal{I}_k| C_n}{|\mathcal{I}_k| - 1} &= 2(R(\widetilde{f}_1; \mathcal{D}_n) - R(\widetilde{f}_2; \mathcal{D}_n)). \end{aligned}$$

Substituting in (D.2), we obtain

$$\begin{aligned} R(\widetilde{f}_M; \mathcal{D}_n) &= 2R(\widetilde{f}_2; \mathcal{D}_n) - R(\widetilde{f}_1; \mathcal{D}_n) + \frac{1}{M} \cdot 2(R(\widetilde{f}_1; \mathcal{D}_n) - R(\widetilde{f}_2; \mathcal{D}_n)) \\ &= -\left(1 - \frac{2}{M}\right) R(\widetilde{f}_1; \mathcal{D}_n) + 2\left(1 - \frac{1}{M}\right) R(\widetilde{f}_2; \mathcal{D}_n). \end{aligned}$$

Thus, subtracting the desired target in (4.13) for with replacement from both sides, we get

$$R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right] = -\left(1 - \frac{2}{M}\right)\left(R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right) + \left(2 - \frac{2}{M}\right)\left(R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right).$$

Taking absolute values on both sides and using triangle inequality yields

$$\left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \leq \left|1 - \frac{2}{M}\right| \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + \left(2 - \frac{2}{M}\right)\left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Taking supremum over M, we have

$$\sup_{M \in \mathbb{N}} \left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \leq \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + 2\left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Finally, since we have

$$R(\widetilde{f}_1; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_1, \qquad R(\widetilde{f}_2; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_2,$$

the desired claim in (4.13) for the case of sampling without replacement follows. □

D.2.2 Asymptotic subsample conditional risk, squared loss

Before we present the proof for Proposition 4.3.3, we first show the upper bound of the squared subsample conditional risk for general M.

Lemma D.2.1 (Bounding the squared subsample conditional risk)**.** *The subsample conditional prediction risk defined in* (4.8) *for the bagged predictor* $\widehat{f}_{M,\mathcal{I}_k}$ *can be bounded as:*

$$\begin{aligned}&\left|R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left\{(2b_2 - b_1) + \frac{2(b_1 - b_2)}{M}\right\}\right| \\ &\leq \left|\frac{1}{M}\sum_{\ell=1}^M R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1\right| + 2\left|\frac{1}{M(M-1)}\sum_{i,j \in [M], i \neq j} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2\right|.\end{aligned} \tag{D.3}$$

Proof of Lemma D.2.1. We start by expanding the squared risk as:

$$\begin{aligned}&R(\widetilde{f}_{M,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \\ &= \int \left(y - \frac{1}{M}\sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})\right)^2 \, \mathrm{d}P(\boldsymbol{x}, y) \\ &= \int \left(\frac{1}{M}\sum_{\ell=1}^M \left(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})\right)\right)^2 \, \mathrm{d}P(\boldsymbol{x}, y) \\ &= \frac{1}{M^2}\sum_{\ell=1}^M \int \left(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})\right)^2 \mathrm{d}P(\boldsymbol{x}, y) + \frac{1}{M^2}\sum_{i=1}^M \sum_{\substack{j=1 \\ j \neq i}}^M \int \left(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i})\right)\left(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j})\right) \mathrm{d}P(\boldsymbol{x}, y) \\ &= \frac{1}{M^2}\sum_{\ell=1}^M R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) + \frac{1}{M^2}\sum_{i=1}^M \sum_{\substack{j=1 \\ j \neq i}}^M \int (y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}))(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j})) \, \mathrm{d}P(\boldsymbol{x}, y) \\ &\overset{(i)}{=} \frac{1}{M^2}\sum_{\ell=1}^M R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell)\end{aligned}$$

$$
\begin{aligned}
&\quad + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \int \frac{1}{2} \left\{ 4\Big(y - \frac{1}{2}\big(\widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}) + \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j})\big)\Big)^2 - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i})\big)^2 - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j})\big)^2 \right\} \mathrm{d}P(\boldsymbol{x}, y) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) \\
&\quad + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{1}{2} \left\{ 4R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) - R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_i) - R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_j) \right\} \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) - \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_i) - \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_j) + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} 2R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_\ell) - \frac{1}{2M^2} \cdot 2 \cdot (M-1) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_\ell) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) \\
&= \left(\frac{1}{M^2} - \frac{(M-1)}{M^2} \right) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_\ell) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) \\
&= -\left(\frac{1}{M} - \frac{2}{M^2} \right) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k};\, \mathcal{D}_n, \{I_i, I_j\}).
\end{aligned}
$$

In the expansion above, for equality (i), we used the fact that $ab = \{4(a/2 + b/2)^2 - a^2 - b^2\}/2$.

Now, subtracting the desired limit on both sides yields

$$
\begin{aligned}
&\left| R(\widetilde{f}_{M,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) - \left\{ (2b_2 - b_1) + \frac{2(b_1 - b_2)}{M} \right\} \right| \\
&= \left| -\left(\frac{1}{M} - \frac{2}{M^2} \right) \sum_{\ell=1}^{M} (R(\widetilde{f}_{1,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}) - b_1) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} (R(\widetilde{f}_{2,\mathcal{I}_k};\, \mathcal{D}_n, \{I_i, I_j\}) - b_2) \right| \\
&\leq \left| 1 - \frac{2}{M} \right| \cdot \left| \frac{1}{M} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}) - b_1 \right| + \frac{2(M-1)}{M} \left| \frac{1}{M(M-1)} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k};\, \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right| \\
&\leq \left| \frac{1}{M} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}) - b_1 \right| + 2 \left| \frac{1}{M(M-1)} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k};\, \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right|.
\end{aligned}
$$

This completes the proof of the upper bound.

□

Next, we present the proof of Proposition 4.3.3.

Proof of Proposition 4.3.3. Lemma 4.3.8 implies the asymptotics for the data conditional risk. Now, consider the asymptotics for the subsample conditional risk of the bagged predictors. From (D.3) of

Lemma D.2.1, it holds that

$$\begin{aligned}
&\left|R(\widetilde{f}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \left\{(2b_2-b_1)+\frac{2(b_1-b_2)}{M}\right\}\right| \\
&\le \left|\frac{1}{M}\sum_{\ell=1}^M R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}) - b_1\right| + 2\left|\frac{1}{M(M-1)}\sum_{i,j\in[M],i\neq j} R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_i,I_j\}) - b_2\right|.
\end{aligned} \tag{D.4}$$

This implies that

$$\begin{aligned}
&\sup_{M\in\mathbb{N}}\left|R(\widetilde{f}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \left\{(2b_2-b_1)+\frac{2(b_1-b_2)}{M}\right\}\right| \\
&\le \sup_{I\in\mathcal{I}_k}|R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I\}) - b_1| + 2\sup_{M\ge 2}\left|\frac{1}{M(M-1)}\sum_{i,j\in[M],i\neq j} R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_i,I_j\}) - b_2\right|.
\end{aligned}$$

The first term on the right hand side converges almost surely to zero by Lemma D.9.6 (1). To prove that the second term converges to zero, we start by noting that

$$U_M = \frac{1}{M(M-1)}\sum_{i,j\in[M],i\neq j}\left\{R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_i,I_j\}) - b_2\right\},$$

is a U-statistics based on either an SRSWR or an SRSWOR sample $I_1,\ldots,I_M$ conditional on $\mathcal{D}_n$. Theorem 2 in Section 3.4.2 of Lee (1990) implies that $\{U_M\}_{M\ge 2}$ is a reverse martingale conditional on $\mathcal{D}_n$ with respect to the some filtration, when we have an SRSWR sample (which is same as an i.i.d. sample). Lemma 2.1 of Sen (1970) proves the same result, when we have an SRSWOR sample. This combined with Theorem 3 (maximal inequality for reverse martingales) in Section 3.4.1 of Lee (1990) (for $r = 1$[1]) yields

$$\begin{aligned}
\mathbb{P}\left(\sup_{M\ge 2}|U_M|\ge\delta \;\middle|\; \mathcal{D}_n\right) &\le \frac{1}{\delta}\mathbb{E}\left[|U_2| \;\middle|\; \mathcal{D}_n\right] \\
&= \frac{1}{\delta}\mathbb{E}\left[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \;\middle|\; \mathcal{D}_n\right].
\end{aligned}$$

The right hand side we know converges to zero almost surely. To see this, we first write as before the right hand side as $\mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \mid \mathcal{D}_n = \mathcal{D}_n(\omega)] = \mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n(\omega),\{I_1,I_2\}) - b_2|]$. We know that for all $\omega\in\mathcal{A}$, $R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n(\omega),\{I_1,I_2\}) \xrightarrow{\text{a.s.}} b_2$ as $n\to\infty$ (from the given assumption). Also, we know (D.10) and that the right hand side of (D.10) converges in L_1 to its probability limit. Hence, Vitali's theorem (Bogachev, 2007, Theorem 4.5.4) implies that $\mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \mid \mathcal{D}_n = \mathcal{D}_n(\omega)]$ converges to zero for all $\omega\in\mathcal{A}$ as $n\to\infty$. Therefore, as $n\to\infty$, for all $\omega\in\mathcal{A}$,

$$\mathbb{P}\left(\sup_{M\ge 2}|U_M|\ge\delta \;\middle|\; \mathcal{D}_n = \mathcal{D}_n(\omega)\right)\to 0.$$

Because probabilities are bounded by one, dominated convergence theorem implies that

$$\mathbb{P}\left(\sup_{M\ge 2}|U_M|\ge\delta\right)\to 0, \quad \text{as} \quad n\to\infty.$$

Therefore,

$$\sup_{M\in\mathbb{N}}\left|R(\widetilde{f}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \left\{(2b_2-b_1)+\frac{2(b_1-b_2)}{M}\right\}\right| \xrightarrow{\text{p}} 0.$$

□

[1] Theorem 3 of Section 3.4.1 is only stated with $r > 1$, but from the proof, it is clear that $r = 1$ is a valid choice.

D.2.3 Conditional risk bounds for convex, strongly-convex, and smooth losses

Proof of Proposition 4.3.6. We split the proof in two parts, depending on the assumption imposed on the loss function L.

Part 1. For any loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ convex in the second argument, one can trivially obtain

$$\begin{aligned} R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) &= \mathbb{E}[L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \mid \mathcal{D}_n] \\ &= \mathbb{E}[\mathbb{E}[L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \mid \{I_\ell\}_{\ell=1}^M] \mid \mathcal{D}_n] \\ &\geq \mathbb{E}[L(y, \mathbb{E}[\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) \mid \{I_\ell\}_{\ell=1}^M]) \mid \mathcal{D}_n]. \end{aligned} \tag{D.5}$$

Here the last inequality follows from Jensen's inequality. Because $\mathbb{E}[\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) \mid \{I_\ell\}_{\ell=1}^M] = \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})$, we get for any $M \geq 1$,

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n).$$

On the other hand, we have by Jensen's inequality

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) = \mathbb{E}\left[L\left(y, \frac{1}{M}\sum_{\ell=1}^M \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})\right) \,\Big|\, \mathcal{D}_n\right] \leq \mathbb{E}\left[\frac{1}{M}\sum_{\ell=1}^M L(y, \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})) \,\Big|\, \mathcal{D}_n\right] = R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n).$$

Summarizing, we get that for any $M \geq 1$,

$$R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n).$$

One can further obtain the monotonicity property by noting that for any $M \geq 1$,

$$\widetilde{f}_{M+1,\mathcal{I}_k}(\boldsymbol{x}, \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M+1}) = \frac{1}{M+1}\sum_{\ell=1}^{M+1} \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) = \frac{1}{(M+1)!}\sum_{\pi'}\left(\frac{1}{M}\sum_{\ell=1}^M \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_{\pi'(\ell)}})\right),$$

where π' represents a permutation of $\{1, 2, \ldots, M+1\}$. Therefore, for any loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is convex in the second argument, we get

$$L(y, \widetilde{f}_{M+1,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M+1})) \leq \frac{1}{(M+1)!}\sum_{\pi'} L\left(y, \widetilde{f}(\boldsymbol{x}; \{\mathcal{D}_{I_{\pi'(\ell)}}\}_{\ell=1}^M)\right).$$

Because any (non-random) subset of a simple random sample with/without replacement is itself a simple random sample with/without replacement, taking conditional expectation on both sides conditional on $\mathcal{D}_n$ yields

$$R(\widetilde{f}_{M+1,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n).$$

This, in particular, implies that $R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n)$ for any $M \geq 1$. This finishes the proof of the first part of the statement.

Part 2. If we assume that the loss function is strongly convex and differentiable in the second argument, then we can improve the lower bound of Part 1 in terms of $\widetilde{f}_\infty$. Formally, if $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is $\underline{m}$-strongly convex, i.e., $L(a, b) - \underline{m}/2b^2$ is convex in b (for every a), then

$$L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \geq L(y, \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})) + \frac{\partial L(y, \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))}{\partial b}(\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})) + \frac{\underline{m}}{2}(\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2.$$

Applying Proposition 4.3.1 and taking the expectation $(\boldsymbol{x}, y)$ conditional on $\mathcal{D}_n$, we obtain

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) + \frac{\underline{m}}{2}\frac{1}{M}\int \frac{1}{|\mathcal{I}_k|}\sum_{I \in \mathcal{I}_k} (\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2 \, \mathrm{d}P(\boldsymbol{x}, y). \tag{D.6}$$

On the other hand, if we assume that the loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is $\overline{m}$ smooth in the second argument, then

$$L(a, b) \leq L(a, b') + \frac{\partial L(a, b')}{\partial b}(b - b') + \frac{\overline{m}}{2}(b - b')^2.$$

It follows that

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) + \frac{\overline{m}}{2} \frac{K_{|\mathcal{I}_k|,M}}{M} \int \sum_{I \in \mathcal{I}_k} (\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2 \, \mathrm{d}P(\boldsymbol{x}, y). \tag{D.7}$$

Combining the lower bound from (D.6) and the upper bound from (D.7) finishes the proof of the second part of the statement. □

D.2.4 From subsample conditional to data conditional risk, $M = 1, 2$

Proof of Lemma 4.3.8. Let us first prove the result when sampling with/without replacement from $\mathcal{I}_k$. The proof for $\mathcal{I}_k^\pi$ would be analogous. Note that $R(\widetilde{f}_1; \mathcal{D}_n) = \mathbb{E}[R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_1\}) \mid \mathcal{D}_n]$ where the expectation is taken over a random draw I_1 from $\mathcal{I}_k$. We are given that $R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1 \xrightarrow{\text{a.s.}} 0$ for every $I \in \mathcal{I}_k$. Under this condition, let us note that

$$\begin{aligned}
\left|\mathbb{E}[R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_1\}) \mid \mathcal{D}_n] - b_1\right| &= \left|\frac{1}{|\mathcal{I}_k|} \sum_{I \in \mathcal{I}_k} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1\right| \\
&\leq \frac{1}{|\mathcal{I}_k|} \sum_{I \in \mathcal{I}_k} |R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1| \\
&\leq \max_{I \in \mathcal{I}_k} |R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

by Lemma D.9.6 (1). Hence, we proved that

$$R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) \xrightarrow{\text{a.s.}} b_1, \quad \text{as} \quad n \to \infty. \tag{D.8}$$

Now, observe that

$$R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \leq \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_1\}) + \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_2\}). \tag{D.9}$$

We will now apply Pratt's lemma (see, e.g., Gut, 2005, Theorem 5.5) to prove almost sure convergence of $\mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \mid \mathcal{D}_n]$. Usually Pratt's lemma is applied unconditionally and here we apply it conditional on $\mathcal{D}_n$. For easier understanding of the proof, let us write $\mathcal{D}_n(\omega)$ in place of $\mathcal{D}_n$ in order to make it clear that we are conditioning on $\mathcal{D}_n$. Recall that $\mathcal{D}_n$ is independent of the subsamples $\{I_\ell\}_{\ell=1}^M$ for any $M \geq 1$. In this notation, inequality (D.9) becomes

$$0 \leq R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2) \leq \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_1\}) + \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_2\}). \tag{D.10}$$

Because $R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} b_1$ for every $I \in \mathcal{I}_k$, there exists a set $\mathcal{A} \subseteq \Omega$ such that $\mathbb{P}(\mathcal{A}) = 1$ and for all $\omega \in \mathcal{A}$, $R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I\}) \xrightarrow{\text{a.s.}} b_1$ for every $I \in \mathcal{I}_k$. Applying Pratt's lemma for every $\omega \in \mathcal{A}$, as $n \to \infty$ and using the fact (D.8) as well as the assumption $R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2) \xrightarrow{\text{a.s.}} b_2$, we get that

$$\mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2)] \;\to\; b_2, \quad \text{for all} \quad \omega \in \mathcal{A}.$$

Note that $R(\widetilde{f}_{2,\mathcal{I}_2}; \mathcal{D}_n(\omega)) = \mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \mid \mathcal{D}_n = \mathcal{D}_n(\omega)]$. Therefore, we conclude

$$R(\widetilde{f}_{2,\mathcal{I}_2}; \mathcal{D}_n) \xrightarrow{\text{a.s.}} b_2, \quad \text{as} \quad n \to \infty. \tag{D.11}$$

Therefore, (4.12) applies to yield asymptotics for the data conditional risk uniformly over $M \in \mathbb{N}$. □

D.2.5 From subsample conditional to data conditional risk, general M

Proof of Theorem 4.3.9. The proof follows by combining Propositions 4.3.2 and 4.3.3, and Lemma 4.3.8. □

D.3 Proof of Theorem 4.4.1 (subagging with replacement, ridge predictor)

For $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ defined in Theorem 4.4.1, we present the proof for ridge and ridgeless predictors in Theorems D.3.1 and D.4.1. For $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ defined in Theorem 4.4.1, the conclusion still holds since the limits of the proportions of intersection between two SRSWR and SRSWOR draws from $\mathcal{I}_k$ are the same from Lemma D.9.3. For proving the asymptotic conditional risks, we will treat $\mathcal{I}_k$ as fixed and use $\widetilde{f}^{\texttt{WR}}_{\lambda,M}$ to denote the ingredient predictor associated with regularization parameter λ.

D.3.1 Proof assembly

Before we present the proof, recall the nonnegative constants defined in (4.19) and (4.20): $v(-\lambda;\theta) \geq 0$ is the unique solution to the fixed-point equation

$$v(-\lambda;\theta)^{-1} = \lambda + \theta \int r(1+v(-\lambda;\theta)r)^{-1}\,\mathrm{d}H(r), \tag{D.12}$$

and the nonnegative constants $\widetilde{v}(-\lambda;\vartheta,\theta)$, and $\widetilde{c}(-\lambda;\theta)$ are defined via the following equations

$$\widetilde{v}(-\lambda,\vartheta,\theta) = \frac{\vartheta\int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}{v(-\lambda;\theta)^{-2} - \vartheta\int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)},\ \widetilde{c}(-\lambda;\theta) = \int r(1+v(-\lambda;\theta))r)^{-2}\,\mathrm{d}G(r). \tag{D.13}$$

It helps to first slightly rewrite the statement of Theorem 4.4.1 for $\lambda > 0$ as follows. Though it suffices to analyze the case $M = 2$ according to Theorem 4.3.9, below we will do the risk decomposition for general M.

Theorem D.3.1 (Risk characterization of subagged ridge predictor)**.** *Let $\widetilde{f}^{\texttt{WR}}_{\lambda,M}$ be the ingredient predictor as defined in (4.18) for $\lambda > 0$. Suppose that Assumptions 4.1-4.5 hold, then for $M = \{1,2,3,\ldots\}$, as $k,n,p\to\infty$, $p/n\to\phi\in[0,\infty)$ and $p/k\to\phi_s\in[\phi,\infty]$, there exists a deterministic function $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)$ such that for $I_1,\ldots,I_M \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)| \xrightarrow{\text{p}} 0,$$

and

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)| \xrightarrow{\text{a.s.}} 0.$$

Furthermore, $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)$ decomposes as

$$\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) := \sigma^2 + \mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) + \mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s),$$

where $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi,\phi_s) + (1-M^{-1})B_\lambda(\phi,\phi_s)$, and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s) + (1-M^{-1})V_\lambda(\phi,\phi_s)$ with

$$B_\lambda(\vartheta,\theta) = \rho^2(1+\widetilde{v}(-\lambda;\vartheta,\theta))\widetilde{c}(-\lambda;\theta), \qquad V_\lambda(\vartheta,\theta) = \sigma^2\widetilde{v}(-\lambda;\vartheta,\theta), \qquad \theta\in(0,\infty],\ \vartheta\leq\theta,$$

where $\widetilde{v}(-\lambda;\vartheta,\theta)$ and $\widetilde{c}(-\lambda;\theta)$ are as defined in (D.13).

Proof of Theorem D.3.1. In what follows, we will prove the results for n,k,p being a sequence of integers $\{n_m\}_{m=1}^\infty$, $\{k_m\}_{m=1}^\infty$, $\{p_m\}_{m=1}^\infty$. For simplicity, we drop the subscript when it is clear from the context.

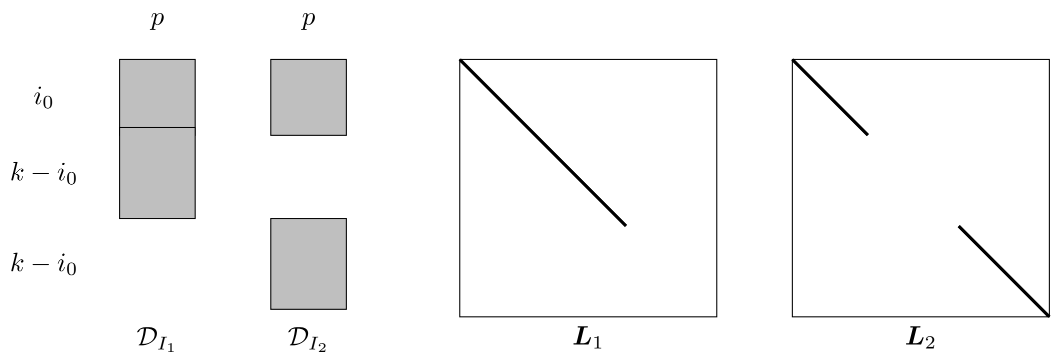


Figure D.1: Illustration of subsampled datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ from $\mathcal{D}_n$. The design matrix of each of them can be represented as $\boldsymbol{L}_j\boldsymbol{X}$ $(j = 1, 2)$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the full design matrix.

For any $m \in [M]$, let I_m be a sample from $\mathcal{I}_k$, and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. An illustration of these notations for $M = 2$ is shown in Figure D.1. The proof will reduce to analyze the individual terms concerning one dataset $\mathcal{D}_{I_m}$, or the cross terms concerning $\mathcal{D}_{I_m}$ and $\mathcal{D}_{I_l}$ for $m \neq l$.

The ingredient estimator takes the form:

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) &= \frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_{I_m}) \\
&= \frac{1}{M}\sum_{m=1}^M (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{y}/k) \\
&= \frac{1}{M}\sum_{m=1}^M \left[\left(\frac{\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}}{k} + \lambda \boldsymbol{I}_p\right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\boldsymbol{\beta}_0 + \left(\frac{\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}}{k} + \lambda \boldsymbol{I}_p\right)^{-1} \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\boldsymbol{\epsilon}\right].
\end{aligned}$$

Denote $\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$ by $\widetilde{\boldsymbol{\beta}}_{\lambda,M}$ for simplicity. Let $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$ for $m \in [M]$, we have

$$\widetilde{\boldsymbol{\beta}}_{\lambda,M} = \frac{1}{M}\sum_{m=1}^M (\boldsymbol{I}_p - \lambda \boldsymbol{M}_m)\boldsymbol{\beta}_0 + \frac{1}{M}\sum_{m=1}^M \boldsymbol{M}_m(\boldsymbol{X}^\top \boldsymbol{L}_m/k)\boldsymbol{\epsilon},$$

which yields

$$\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M} = \frac{1}{M}\sum_{m=1}^M \lambda \boldsymbol{M}_m \boldsymbol{\beta}_0 - \frac{1}{M}\sum_{m=1}^M \boldsymbol{M}_m(\boldsymbol{X}^\top \boldsymbol{L}_m/k)\boldsymbol{\epsilon}.$$

Thus, the conditional risk is given by

$$\begin{aligned}
R(\widetilde{f}_{M,\lambda}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) &= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(y_0 - \boldsymbol{x}_0^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M})^2] \\
&= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M})^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}) \\
&= \sigma^2 + T_C + T_B + T_V,
\end{aligned}$$

where the constant term T_C, bias term T_B, and the variance term T_V are given by

$$T_C = -\frac{2\lambda}{M^2}\cdot \boldsymbol{\epsilon}^\top \left(\sum_{m=1}^M \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{\Sigma}\left(\sum_{m=1}^M \boldsymbol{M}_m\right)\boldsymbol{\beta}_0, \tag{D.14}$$

$$T_B = \frac{\lambda^2}{M^2}\cdot \boldsymbol{\beta}_0^\top \left(\sum_{m=1}^M \boldsymbol{M}_m\right)\boldsymbol{\Sigma}\left(\sum_{m=1}^M \boldsymbol{M}_m\right)\boldsymbol{\beta}_0, \tag{D.15}$$

$$T_V = \frac{1}{M^2} \cdot \boldsymbol{\epsilon}^\top \left(\sum_{m=1}^{M} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right) \boldsymbol{\epsilon}. \tag{D.16}$$

Next we analyze the three terms separately for $M \in \{1, 2\}$. From Lemmas D.3.2 and D.3.3, we have that $T_C \xrightarrow{\text{a.s.}} 0$, and

$$\begin{aligned} T_V &= \frac{1}{M^2} \sum_{m=1}^{M} \boldsymbol{\epsilon}^\top \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\Sigma} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\epsilon} + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \boldsymbol{\epsilon}^\top \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\Sigma} \boldsymbol{M}_l \frac{\boldsymbol{X}^\top \boldsymbol{L}_l}{k} \boldsymbol{\epsilon} \\ &\xrightarrow{\text{a.s.}} \frac{1}{M^2} \sum_{m=1}^{M} \frac{\sigma^2}{k} \operatorname{tr}(\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \boldsymbol{\Sigma}) + \frac{1}{M^2} \sum_{m \neq 1} \frac{\sigma^2}{k^2} \operatorname{tr}(\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma}) := T_V'. \end{aligned}$$

Thus, it remains to obtain the deterministic equivalent for the bias term T_B and the trace term T_V'. From Lemma D.3.4 and Lemma D.3.5, we have that for all $I_1 \in \mathcal{I}_k$ when $M = 1$ and for all $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$ when $M = 2$, it holds that

$$\begin{aligned} T_B &= \frac{\lambda^2}{M^2} \sum_{m=1}^{M} \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{\beta}_0 + \frac{\lambda^2}{M^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_l \boldsymbol{\beta}_0 \\ &\xrightarrow{\text{a.s.}} \frac{\rho^2}{M} (1 + \widetilde{v}(-\lambda; \phi, \phi_s)) \widetilde{c}(-\lambda; \phi_s) + \frac{\rho^2 (M-1)}{M} (1 + \widetilde{v}(-\lambda; \phi, \phi_s)) \widetilde{c}(-\lambda; \phi_s) \\ T_V' &\xrightarrow{\text{a.s.}} \frac{\sigma^2}{M} \widetilde{v}(-\lambda; \phi_s, \phi_s) + \frac{\sigma^2 (M-1)}{M} \widetilde{v}(-\lambda; \phi, \phi_s), \end{aligned}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty)$, where the nonnegative constants $\widetilde{v}(-\lambda; \phi, \phi_s)$ and $\widetilde{c}(-\lambda; \phi_s)$ are as defined in (D.13). Therefore, we have shown that for all $I \in \mathcal{I}_k$,

$$R(\widetilde{f}_{\lambda,1}; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s),$$

and for all $I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,

$$R(\widetilde{f}_{\lambda,2}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{sub}}_{\lambda,2}(\phi, \phi_s),$$

where

$$\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \frac{1}{M} (B_\lambda(\phi_s, \phi_s) + V_\lambda(\phi_s, \phi_s)) + \frac{M-1}{M} (B_\lambda(\phi, \phi_s) + V_\lambda(\phi, \phi_s)),$$

and the components are:

$$B_\lambda(\phi, \phi_s) = \rho^2 (1 + \widetilde{v}(-\lambda; \phi, \phi_s)) \widetilde{c}(-\lambda; \phi_s), \qquad V_\lambda(\phi, \phi_s) = \sigma^2 \widetilde{v}(-\lambda; \phi, \phi_s).$$

The proof for the boundary case when $\phi_s = \infty$ follows from Proposition D.3.6. Then, we have that the function $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ is continuous on $[\phi, \infty]$.

Finally, the risk expression for general M and the uniformity claim over $M \in \mathbb{N}$ follow from Theorem 4.3.9. □

D.3.2 Component concentrations

In this subsection, we will show that the cross-term T_C converges to zero and the variance term T_V converge to its corresponding trace expectation.

D.3.2.1 Convergence of the cross term

Lemma D.3.2 (Convergence of the cross term)**.** *Under Assumptions 4.1-4.5, for T_C as defined in* (D.14)*, we have $T_C \xrightarrow{\text{a.s.}} 0$ as $k, p \to \infty$ and $p/k \to \phi_s$.*

Proof of Lemma D.3.2. Note that

$$T_C = -\frac{2\lambda}{M^2} \cdot \frac{1}{k} \left\langle \left(\sum_{m=1}^{M} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \right) \boldsymbol{\beta}_0, \boldsymbol{\epsilon} \right\rangle .$$

We next bound the squared norm

$$\begin{aligned}
&\frac{1}{k} \left\| \frac{1}{M} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \right) \boldsymbol{\beta}_0 \right\|_2^2 \\
&\leq \sum_{j=1}^{M} \sum_{l=1}^{M} \frac{1}{M^2 k} \left\| (\boldsymbol{M}_j \boldsymbol{X}^\top \boldsymbol{L}_j)^\top \boldsymbol{\Sigma} \boldsymbol{M}_l \boldsymbol{\beta}_0 \right\|_2^2 \\
&\leq \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{j=1}^{M} \sum_{l=1}^{M} \frac{1}{k} \left\| \boldsymbol{M}_l \boldsymbol{\Sigma} \boldsymbol{M}_j \boldsymbol{X}^\top \boldsymbol{L}_j \boldsymbol{X} \boldsymbol{M}_j \boldsymbol{\Sigma} \boldsymbol{M}_l \right\|_{\mathrm{op}} \\
&\leq \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{j=1}^{M} \sum_{l=1}^{M} \|\boldsymbol{M}_l\|_{\mathrm{op}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^2 \left\| \boldsymbol{M}_j (\boldsymbol{X}^\top \boldsymbol{L}_j \boldsymbol{X}/k) \boldsymbol{M}_j \right\|_{\mathrm{op}} \\
&= \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{j=1}^{M} \sum_{l=1}^{M} \|\boldsymbol{M}_l\|_{\mathrm{op}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}}^2 \|\boldsymbol{M}_j\|_{\mathrm{op}} \|\boldsymbol{I}_p - \lambda \boldsymbol{M}_j\|_{\mathrm{op}} \\
&\leq \frac{\|\boldsymbol{\beta}_0\|_2^2 \, r_{\max}^2}{\lambda^3},
\end{aligned}$$

where the last inequality is due to Assumption 4.4 and the fact that $\|\boldsymbol{M}_j\|_{\mathrm{op}} \leq 1/\lambda$. By Assumption 4.3, the above quantity is uniformly bounded in p. Applying Lemma D.9.4, we thus have that $T_C \xrightarrow{\text{a.s.}} 0$. □

D.3.2.2 Convergence of the variance term

Lemma D.3.3 (Convergence of the variance term)**.** *Under Assumptions 4.1-4.5, let $M \in \mathbb{N}$. For all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n \times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. Then, for all $m, l \in [M]$ and $m \neq l$, it holds that*

$$\begin{aligned}
\frac{1}{k^2} \boldsymbol{\epsilon}^\top \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{\epsilon} - \frac{\sigma^2}{k} \operatorname{tr}(\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \boldsymbol{\Sigma}) &\xrightarrow{\text{a.s.}} 0, \\
\frac{1}{k^2} \boldsymbol{\epsilon}^\top \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{\epsilon} - \frac{\sigma^2}{k^2} \operatorname{tr}(\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma}) &\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty)$.

Proof of Lemma D.3.3. Note that the first term is the same as the variance term for ridge predictor trained on k i.i.d. samples $(\boldsymbol{L}_m \boldsymbol{X}, \boldsymbol{L}_m \boldsymbol{y})$. Notice that $\boldsymbol{L}_m \boldsymbol{\epsilon}$ is independent of $\boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m$, and

$$\begin{aligned}
\frac{1}{k} \left\| \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m \right\|_{\mathrm{op}} &\leq \left\| \widehat{\boldsymbol{\Sigma}}_m \right\|_{\mathrm{op}}^{\frac{1}{2}} \|\boldsymbol{M}_m\|_{\mathrm{op}} \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \|\boldsymbol{M}_m\|_{\mathrm{op}} \left\| \widehat{\boldsymbol{\Sigma}}_m \right\|_{\mathrm{op}}^{\frac{1}{2}} \\
&= \left\| \widehat{\boldsymbol{\Sigma}}_m \right\|_{\mathrm{op}} \|\boldsymbol{M}_m\|_{\mathrm{op}}^2 \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \\
&\leq \frac{r_{\max}}{\lambda^2} \left\| \widehat{\boldsymbol{\Sigma}}_m \right\|_{\mathrm{op}} .
\end{aligned}$$

Now, we have $\limsup \left\| \widehat{\boldsymbol{\Sigma}}_m \right\|_{\mathrm{op}} \leq \limsup \max_{1 \leq i \leq p} s_i^2 \leq r_{\max} (1 + \sqrt{\phi_s})^2$ almost surely as $k, p \to \infty$ and $p/k \to \phi_s \in (0, \infty)$ from Bai and Silverstein (2010). From Lemma D.9.5, it follows that

$$\boldsymbol{\epsilon}^\top \frac{\boldsymbol{L}_m^\top \boldsymbol{X}}{k} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\epsilon} - \frac{\sigma^2}{k^2} \operatorname{tr}(\boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_m) \xrightarrow{\text{a.s.}} 0.$$

Since $\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_m\boldsymbol{X}^\top\boldsymbol{L}_m)/k^2 = \operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma})/k = \operatorname{tr}(\boldsymbol{M}_m^2\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{\Sigma})/k$, we further have $\forall\ m \in [M]$,

$$\boldsymbol{\epsilon}^\top \frac{\boldsymbol{L}_m^\top\boldsymbol{X}}{k}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\boldsymbol{\epsilon} - \frac{\sigma^2}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0. \tag{D.17}$$

The second term involves the cross-term $\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l$. Note that

$$\frac{1}{n}\left\|\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l\boldsymbol{X}^\top\boldsymbol{L}_l\right\|_{\text{op}} \leq \frac{k}{n}\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^{\frac{1}{2}}\|\boldsymbol{M}_m\|_{\text{op}}\|\boldsymbol{M}_l\|_{\text{op}}\|\boldsymbol{\Sigma}\|_{\text{op}} \leq \frac{r_{\max}}{\lambda^2}\frac{k}{n}\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^{\frac{1}{2}}.$$

Because $\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}$ for $m \in [M]$ are uniformly bounded almost surely, again by Lemma D.9.5, it follows that

$$\frac{1}{n}\boldsymbol{\epsilon}^\top\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l\boldsymbol{X}^\top\boldsymbol{L}_l\boldsymbol{\epsilon} - \frac{\sigma^2}{n}\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l\boldsymbol{X}^\top\boldsymbol{L}_l) \xrightarrow{\text{a.s.}} 0.$$

Since $k/n \to \phi_s/\phi$, we have

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon} - \frac{\sigma^2}{k^2}\operatorname{tr}(\boldsymbol{M}_l\boldsymbol{X}^\top\boldsymbol{L}_l\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0. \tag{D.18}$$

□

D.3.3 Component deterministic approximations

D.3.3.1 Deterministic approximation of the bias functional

Lemma D.3.4 (Deterministic approximation of the bias functional)**.** *Under Assumptions 4.1-4.5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}$. Then, it holds that*

1. *for all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_m\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s),$$

2. *for all $m, l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s),$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)$, where $\phi_0 = \phi_s^2/\phi$, T_B is as defined in (D.15), and the nonnegative constants $\widetilde{v}(-\lambda;\phi,\phi_s)$ and $\widetilde{c}(-\lambda;\phi_s)$ are as defined in (D.13).

Proof of Lemma D.3.4. From Lemma D.8.9 (1) we have that for $m \in [M]$,

$$\lambda^2\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_m \simeq (\widetilde{v}_b(-\lambda;\phi_s)+1)\cdot(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}. \tag{D.19}$$

By the definition of deterministic equivalent, we have

$$\begin{aligned}
\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_m\boldsymbol{\beta}_0 &\xrightarrow{\text{a.s.}} \lim_{p\to\infty}(1+\widetilde{v}_b(-\lambda;\phi_s))\sum_{i=1}^p\frac{r_i}{(1+r_iv(-\lambda;\phi_s))^2}(\boldsymbol{\beta}_0^\top w_i)^2\\
&= \lim_{p\to\infty}\|\boldsymbol{\beta}_0\|_2^2(1+\widetilde{v}_b(-\lambda;\phi_s))\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G_p(r)\\
&= \rho^2(1+\widetilde{v}_b(-\lambda;\phi_s))\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G(r),
\end{aligned} \tag{D.20}$$

where the last equality holds since G_p and G have compact supports and Assumptions 4.3 and 4.5.

For the cross term, it suffices to derive the deterministic equivalent of $\boldsymbol{\beta}_0^\top \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{\beta}_0/2$. We begin with analyze the deterministic equivalent of $\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2$. Let $i_0 = \operatorname{tr}(\boldsymbol{L}_1\boldsymbol{L}_2)$ be the number of shared samples between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$, we use the decomposition

$$\boldsymbol{M}_j^{-1} = \frac{i_0}{k}(\widehat{\boldsymbol{\Sigma}}_0 + \lambda \boldsymbol{I}_p) + \frac{k-i_0}{k}(\widehat{\boldsymbol{\Sigma}}_j^{\mathrm{ind}} + \lambda \boldsymbol{I}_p), \qquad j=1,2,$$

where $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1\boldsymbol{L}_2\boldsymbol{X}/i_0$ and $\widehat{\boldsymbol{\Sigma}}_j^{\mathrm{ind}} = \boldsymbol{X}^\top(\boldsymbol{L}_j - \boldsymbol{L}_1\boldsymbol{L}_2)\boldsymbol{X}/(k-i_0)$ are the common and individual covariance estimators of the two datasets. Let $\boldsymbol{N}_0 = (\widehat{\boldsymbol{\Sigma}}_0 + \lambda\boldsymbol{I}_p)^{-1}$ and $\boldsymbol{N}_j = (\widehat{\boldsymbol{\Sigma}}_j^{\mathrm{ind}} + \lambda\boldsymbol{I}_p)^{-1}$ for $j=1,2$. Then

$$\boldsymbol{M}_j = \left(\frac{i_0}{k}\boldsymbol{N}_0^{-1} + \frac{k-i_0}{k}\boldsymbol{N}_j^{-1}\right)^{-1}, \qquad j=1,2, \tag{D.21}$$

where the equalities hold because $\boldsymbol{N}_0$ is invertible when $\lambda>0$. Conditioning on i_0, we will show a sequence of deterministic equivalents

$$\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \overset{(a)}{\simeq} \lambda\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}_2 \overset{(b)}{\simeq} \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0} \overset{(c)}{\simeq} (\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2)^{\mathrm{det}}_{i_0},$$

where in each step we consider randomness from $\boldsymbol{N}_1$, $\boldsymbol{N}_2$, $\boldsymbol{N}_0$, respectively, since they are independent to each other conditioning on i_0. The subscript of the deterministic equivalent indicates the dependence on the corresponding random variables, and we will specify each deterministic equivalent in the following proof.

When $i_0 = k$, we have $\boldsymbol{M}_1 = \boldsymbol{M}_2$ and the above asymptotic equation reduces to (D.19). We next prove the case when $i_0 < k$.

Part (a). Since $\boldsymbol{N}_1$ is independent to $\boldsymbol{N}_0$ conditioning on i_0, from Definition D.8.5 and Lemma D.8.10 (1) we have

$$\lambda\boldsymbol{M}_1 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0} := \frac{k}{k-i_0}\left(v_1\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \Big| \, i_0,$$

where $v_1 = v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})$, $\boldsymbol{\Sigma}_{\boldsymbol{C}_1} = (\boldsymbol{I}_p+\boldsymbol{C}_1)^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C}_1)^{-\frac{1}{2}}$, $\boldsymbol{C}_1 = i_0(\lambda(k-i_0))^{-1}\boldsymbol{N}_0^{-1}$, and $\gamma_1 = p/(k-i_0)$. Here the subscripts of v_1 and $\boldsymbol{C}_1$ are related to the aspect ratio γ_1. Because of the sub-multiplicativity of operator norm, we have

$$\|\boldsymbol{\Sigma}\boldsymbol{M}_2\|_{\mathrm{op}} \leq \|\boldsymbol{\Sigma}\|_{\mathrm{op}}\,\|\boldsymbol{M}_2\|_{\mathrm{op}} \leq \frac{r_{\max}}{\lambda}.$$

By Proposition D.8.6 (2), we have $\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq \lambda\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}_2 \mid i_0$.

Part (b). Analoguously, we have

$$\begin{aligned}\lambda\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}_2 &\simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\\ &\simeq \left(\frac{k}{k-i_0}\right)^2 \left(v_1\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}_1\right)^{-1}\boldsymbol{\Sigma}\left(v_1\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}_1\right)^{-1} \mid i_0,\end{aligned}$$

as $\|\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\|_{\mathrm{op}} \leq 1$.

Part (c). As we have symmetrized the expression, we have

$$\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0} = \frac{k^2}{i_0^2}\lambda^2(\boldsymbol{N}_0^{-1}+\lambda\boldsymbol{C}_0)^{-1}\boldsymbol{\Sigma}(\boldsymbol{N}_0^{-1}+\lambda\boldsymbol{C}_0)^{-1} \mid i_0,$$

where $\boldsymbol{C}_0 = (k - i_0)/i_0 \cdot (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)$. Define $\boldsymbol{\Sigma}_{\boldsymbol{C}_0} = (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}}$. Conditioning on i_0, by Lemma D.8.10 (1), we have

$$\begin{aligned}
\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1 \boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}] &= \operatorname{tr}[\boldsymbol{\Sigma}(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1)^{-1}] \\
&= \frac{\lambda(k - i_0)}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{N}_0^{-1} + \frac{\lambda(k-i_0)}{i_0}(v_1\boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&\overset{a.s.}{=} \frac{k - i_0}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(v_0 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \frac{k - i_0}{i_0}(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&= \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k - i_0} v_0 + v_1\right)\boldsymbol{\Sigma} + \frac{k}{k - i_0}\boldsymbol{I}_p\right)^{-1}\right],
\end{aligned}$$

where $v_0 = v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}_0})$and $\gamma_0 = p/i_0$. Note that the fixed-point solution v_0 depends on v_1. The fixed-point equations reduce to

$$\begin{aligned}
\frac{1}{v_0} &= \lambda + \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_0}(v_0 \boldsymbol{\Sigma}_{\boldsymbol{C}_0} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k} v_0 + \frac{k - i_0}{k} v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p \\
\frac{1}{v_1} &= \lambda + \gamma_1 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1 \boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k} v_0 + \frac{k - i_0}{k} v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p
\end{aligned}$$

almost surely. Note that the solution (v_0, v_1) to the above equations is a pair of positive numbers and does not depend on samples. If (v_0, v_1) is a solution to the above system, then (v_1, v_0) is also a solution. Thus, any solution to the above equations must be unique. On the other hand, since $v_0 = v_1 = v(-\lambda; p/k)$ satisfies the above equations, it is the unique solution. By Lemma D.8.16, we can replace $v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$ by the solution $v_0 = v_1 = v(-\lambda; p/k)$ of the above system, which does not depend on samples. Thus,

$$\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq \frac{k^2}{i_0^2} \lambda^2 (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}^*)^{-1} \boldsymbol{\Sigma} (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}^*)^{-1} \mid i_0, \tag{D.22}$$

where $\boldsymbol{C}^* = (k - i_0)/i_0 \cdot (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. By Lemma D.8.10 (2), we have

$$\boldsymbol{M}^{\det}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0, i_0} \simeq (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\det}_{i_0} := \frac{k^2}{i_0^2} (\widetilde{v}_b(-\lambda; \gamma_0, \boldsymbol{C}^*) + 1)(v(-\lambda; \gamma_0, \boldsymbol{C}^*)\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2} \boldsymbol{\Sigma} \mid i_0, \tag{D.23}$$

where $\gamma_0 = p/i_0$, and $v(-\lambda; \gamma_0, \boldsymbol{C}^*)$ and $\widetilde{v}_b(-\lambda; \gamma_0, \boldsymbol{C}^*)$ are defined through the following equations:

$$\begin{aligned}
\frac{1}{v(-\lambda; \gamma_0, \boldsymbol{C}^*)} &= \lambda + \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}(v(-\lambda; \gamma_0, \boldsymbol{C}^*)\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-1}]/p \\
\frac{1}{\widetilde{v}_b(-\lambda; \gamma_0, \boldsymbol{C}^*)} &= \frac{\gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}^2 (v(-\lambda; \gamma_0, \boldsymbol{C}^*)\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2}]/p}{v(-\lambda; \gamma_0, \boldsymbol{C}^*)^{-2} - \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}^2 (v(-\lambda; \gamma_0, \boldsymbol{C}^*)\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2}]/p}.
\end{aligned}$$

From Parts (a) to (c), we have shown that $\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\det}_{i_0} \mid i_0$ for $i_0 < k$. Note that the above equivalence also holds for $i_0 = k$. That is, this holds for all $i_0 \in \{0, 1, \cdots, k\}$. By Proposition D.8.6 (1), we can obtain the unconditioned asymptotic equivalence $\boldsymbol{M}^{\det}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0, i_0} \simeq (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\det}_{i_0}$.

Note that from Lemma D.8.13 $\widetilde{v}_b(-\lambda; \gamma)$ and $v(-\lambda; \gamma)$ are continuous on γ, and from Lemma D.9.3, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$, where $\phi_s \in (0, \infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k, p \to \infty$. We have

$$(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\det}_{i_0} \simeq \frac{\phi_s^2}{\phi^2} (\widetilde{v}_b(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'}) + 1)(v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}')^{-2} \boldsymbol{\Sigma},$$

where $\boldsymbol{C}' = (\phi_s - \phi)/\phi \cdot (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$ and $\phi_0 = \phi_s^2/\phi$. Note that

$$\frac{1}{v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'})} = \lambda + \phi_0 \int \frac{r}{1 + r v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'})} \, \mathrm{d}H(r; \boldsymbol{\Sigma}_{\boldsymbol{C}'})$$

$$= \lambda + \phi_s \lim_{p\to\infty} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\frac{\phi}{\phi_s}(v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma}+\boldsymbol{I}_p)+\left(1-\frac{\phi}{\phi_s}\right)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)\right)^{-1}\right]/p$$

$$\frac{1}{v(-\lambda;\phi_s)} = \lambda + \phi_s \lim_{p\to\infty} \operatorname{tr}\left[\boldsymbol{\Sigma}(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\right]/p.$$

We have

$$v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'}) = v(-\lambda;\phi_s) \tag{D.24}$$

is a solution to the first fixed-point equation. From Lemma D.8.12 (2), this solution is also unique. Then, we have

$$\begin{aligned}1+\widetilde{v}_b(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'}) &= \lim_{p\to\infty}\frac{v(-\lambda;\phi_0,\boldsymbol{C}')^{-2}}{v(-\lambda;\phi_0,\boldsymbol{C}')^{-2}-\phi_0\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_0,\boldsymbol{C}')\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}')^{-2}]/p}\\ &= \lim_{p\to\infty}\frac{v(-\lambda;\phi_s)^{-2}}{v(-\lambda;\phi_s)^{-2}-\phi\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}]/p}\\ &= \frac{v(-\lambda;\phi_s)^{-2}}{v(-\lambda;\phi_s)^{-2}-\phi\displaystyle\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)} := 1+\widetilde{v}(-\lambda;\phi,\phi_s).\end{aligned}$$

From Lemma D.8.12 (4), we have that $1+\widetilde{v}(-\lambda;\phi,\phi_s)>0$. To conclude, we have shown that

$$\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq (1+\widetilde{v}(-\lambda;\phi,\phi_s))\,(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\,\boldsymbol{\Sigma}. \tag{D.25}$$

By the definition of deterministic equivalent, we have

$$\begin{aligned}\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{\beta}_0 &\xrightarrow{\text{a.s.}} \lim_{p\to\infty}\sum_{i=1}^{p}\frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r_i}{(1+v(-\lambda;\phi_s)r_i)^2}(\boldsymbol{\beta}_0^\top w_i)^2\\ &= \lim_{p\to\infty}\|\boldsymbol{\beta}_0\|_2^2\int\frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G_p(r)\\ &= \rho^2\int\frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G(r),\end{aligned} \tag{D.26}$$

where in the last line we used the fact that G_p and G have compact supports and Assumptions 4.3 and 4.5. The conclusion follows by combining (D.20) and (D.26). □

D.3.3.2 Deterministic approximation of the variance functional

Lemma D.3.5 (Deterministic approximation of the variance functional)**.** *Under Assumptions 4.1-4.5, for all $m\in[M]$ and $I_m\in\mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m=\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$, $\boldsymbol{L}_m\in\mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll}=1$ if $l\in I_m$ and 0 otherwise, and $\boldsymbol{M}_m=(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^{-1}$. Then, it holds that*

1. *for all $m\in[M]$ and $I_m\in\mathcal{I}_k$,*

$$\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \widetilde{v}(-\lambda;\phi_s,\phi_s),$$

2. *for all $m,l\in[M]$, $m\neq l$ and $I_m,I_l\overset{\texttt{SRSWR}}{\sim}\mathcal{I}_k$,*

$$\frac{1}{k^2}\operatorname{tr}(\boldsymbol{M}_l\boldsymbol{X}^\top\boldsymbol{L}_l\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \widetilde{v}(-\lambda;\phi,\phi_s),$$

as $n,k,p\to\infty$, $p/n\to\phi\in(0,\infty)$, and $p/k\to\phi_s\in[\phi,\infty)$, where the nonnegative constant $\widetilde{v}(\lambda;\phi,\phi_s)$ is as defined in (D.13).

Proof of Lemma D.3.5. From Lemma D.8.9 (2), we have that for $j \in [M]$,

$$\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{M}_j \boldsymbol{\Sigma} \simeq \widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2. \tag{D.27}$$

By the trace rule Lemma D.8.4 (4) , we have

$$\begin{aligned}\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{M}_j \boldsymbol{\Sigma}) &\xrightarrow{\text{a.s.}} \lim_{p\to\infty} \frac{p}{k}\cdot\frac{1}{p}\operatorname{tr}(\widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2)\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s)\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\frac{r_i^2}{(v(-\lambda;\phi_s)r_i+1)^2}\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s)\lim_{p\to\infty}\int\frac{r^2}{(v(-\lambda;\phi_s)r+1)^2}\,\mathrm{d}H_p(r)\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s)\int\frac{r^2}{(v(-\lambda;\phi_s)r+1)^2}\,\mathrm{d}H(r), \qquad j=1,2,\end{aligned} \tag{D.28}$$

where in the last line we used the fact that H_p and H have compact supports and Assumption 4.5.

For the cross term, it suffices to derive the deterministic equivalent of $\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma}$ where $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1 \boldsymbol{L}_2 \boldsymbol{X}/i_0$ and $i_0 = \operatorname{tr}(\boldsymbol{L}_1 \boldsymbol{L}_2)$. We again show a sequence of deterministic equivalents as in the proof for Lemma D.3.4:

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \overset{(a)}{\simeq} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \overset{(b)}{\simeq} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} \overset{(c)}{\simeq} (\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma})^{\det}_{i_0} \mid i_0.$$

When $i_0 = k$, this reduces to (D.27). We next show the case when $i_0 < k$.

Part (a). We use Lemma D.8.10 (1) to obtain

$$\boldsymbol{M}_1 \simeq \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} := \frac{k}{\lambda(k-i_0)}\left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \mid i_0 \tag{D.29}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{C}_1} = (\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}}$, $\boldsymbol{C}_1 = i_0(\lambda(k-i_0))^{-1}\boldsymbol{N}_0^{-1}$, and $\gamma_1 = p/(k-i_0)$. Let $\gamma_0 = p/i_0$. Note that conditioning on i_0, $\limsup \left\|\widehat{\boldsymbol{\Sigma}}_0\right\|_{\mathrm{op}} \le r_{\max}(1+\sqrt{\phi_0})^2$ almost surely as $i_0, p \to \infty$ and $\gamma_0 \to \phi_0 \in (0,\infty)$ from Bai and Silverstein (2010). Then $\widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma}$ has bounded operator norm and we have $\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \simeq \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \mid i_0$ by Proposition D.8.6 (2).

Part (b). Similarly, we have $\boldsymbol{M}_2 \simeq \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \mid i_0$ and $\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \simeq \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} \mid i_0$.

Part (c). Note that

$$\begin{aligned}\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} &= \frac{k^2}{\lambda^2(k-i_0)^2}\left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1}\boldsymbol{\Sigma}\\
&= \frac{k^2}{i_0^2}\left(\boldsymbol{N}_0^{-1} + \lambda\boldsymbol{C}_0\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(\boldsymbol{N}_0^{-1} + \lambda\boldsymbol{C}_0\right)^{-1}\boldsymbol{\Sigma},\end{aligned}$$

where $\boldsymbol{C}_0 = (k-i_0)/i_0 \cdot (v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. Define $\boldsymbol{\Sigma}_{\boldsymbol{C}_0} = (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}}$. Conditioning on i_0, by Lemma D.8.10 (1) we have

$$\begin{aligned}\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1\boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}] &= \operatorname{tr}[\boldsymbol{\Sigma}(v_1\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1)^{-1}]\\
&= \frac{\lambda(k-i_0)}{i_0}\operatorname{tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{N}_0^{-1} + \frac{\lambda(k-i_0)}{i_0}(v_1\boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right]\\
&\overset{a.s.}{=} \frac{k-i_0}{i_0}\operatorname{tr}\left[\boldsymbol{\Sigma}\left(v_0\boldsymbol{\Sigma} + \boldsymbol{I}_p + \frac{k-i_0}{i_0}(v_1\boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right]\end{aligned}$$

$$= \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k-i_0}v_0 + v_1\right)\boldsymbol{\Sigma} + \frac{k}{k-i_0}\boldsymbol{I}_p\right)^{-1}\right]$$

where $v_0 = v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}_0})$ and $\gamma_0 = p/i_0$. Note that the fixed-point solution v_0 depends on v_1. The fixed-point equations reduce to

$$\frac{1}{v_0} = \lambda + \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_0}(v_0\boldsymbol{\Sigma}_{\boldsymbol{C}_0} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k}\operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k}v_0 + \frac{k-i_0}{k}v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p$$
$$\frac{1}{v_1} = \lambda + \gamma_1 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1\boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k}\operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k}v_0 + \frac{k-i_0}{k}v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p$$

almost surely. By the same argument as in the proof for Lemma D.3.4, we have that the solution $v_0 = v_1 = v(-\lambda; p/k)$ of the above system does not depend on samples and equals to $v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$ or $v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}_0})$ almost surely. Thus, by Lemma D.8.16,

$$\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma} \simeq \frac{k^2}{i_0^2}(\boldsymbol{N}_0^{-1} + \lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_0(\boldsymbol{N}_0^{-1} + \lambda\boldsymbol{C}^*)^{-1} \mid i_0,$$

where $\boldsymbol{C}^* = (k-i_0)/i_0 \cdot (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. From Lemma D.8.10 (3), we have

$$\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\det}_{i_0} := \frac{k^2}{i_0^2}\widetilde{v}_v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}^*})(v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}^*})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2}\boldsymbol{\Sigma}^2 \mid i_0,$$

where $\gamma_0 = p/i_0$.

From Parts (a) to (c), we have shown that $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\det}_{i_0} \mid i_0$ for $i_0 < k$. Note that this also holds for $i_0 = k$. Then by Proposition D.8.6, $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\det}_{i_0}$.

Note that from Lemma D.8.13, $\widetilde{v}_b(-\lambda; \gamma)$ and $v(-\lambda; \gamma)$ are continuous on γ, and from Lemma D.9.3, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$ where $\phi_s \in (0, \infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k, p \to \infty$. We have

$$\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \frac{\phi_s^2}{\phi^2}\widetilde{v}_v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'})(v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}')^{-2}\boldsymbol{\Sigma}^2,$$

where $\phi_0 = \phi_s^2/\phi$, $\boldsymbol{\Sigma}_{\boldsymbol{C}'} = (\boldsymbol{I}_p + \boldsymbol{C}')^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C}')^{-\frac{1}{2}}$, and $\boldsymbol{C}' = (\phi_s - \phi)/\phi \cdot (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. From (D.24), we have that $v(-\lambda; \phi_0; \boldsymbol{\Sigma}_{\boldsymbol{C}'}) = v(-\lambda; \phi_s)$, and

$$\begin{aligned}
\phi\widetilde{v}_v(-\lambda; \phi_0, \boldsymbol{\Sigma}_{\boldsymbol{C}'}) &= \lim_{p\to\infty} \frac{\phi}{v(-\lambda; \phi_0, \boldsymbol{C}')^{-2} - \phi_0 \operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda; \phi_0, \boldsymbol{C}')\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}')^{-2}]/p} \\
&= \lim_{p\to\infty} \frac{\phi}{v(-\lambda; \phi_s)^{-2} - \phi \operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}]/p} \\
&= \frac{\phi}{v(-\lambda; \phi_s)^{-2} - \phi \displaystyle\int \frac{r^2}{(1 + v(-\lambda; \phi_s)r)^2}\,\mathrm{d}H(r)} := v_v(-\lambda; \phi, \phi_s).
\end{aligned}$$

From Lemma D.8.12 (4), we have that $v_v(-\lambda; \phi, \phi_s) > 0$. Then we have

$$\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \phi^{-1}v_v(-\lambda; \phi, \phi_s)(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2, \tag{D.30}$$

and thus, we have

$$\begin{aligned}
\frac{i_0}{k^2}\operatorname{tr}(\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})) &\xrightarrow{\text{a.s.}} \lim_{p\to\infty} \frac{i_0 p}{k^2}\frac{1}{\phi} \cdot \frac{1}{p}\operatorname{tr}(v_v(-\lambda; \phi, \phi_s)(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2) \\
&= \lim_{p\to\infty} \frac{1}{p}\sum_{i=1}^{p} \frac{v_v(-\lambda; \phi, \phi_s)r_i^2}{(1 + v(-\lambda; \phi_s)r_i)^2}
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{p\to\infty} \int \frac{v_v(-\lambda;\phi,\phi_s) r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H_p(r) \\
&= \int \frac{\phi v_v(-\lambda;\phi,\phi_s) r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r) := \widetilde{v}(-\lambda;\phi,\phi_s), \qquad \text{(D.31)}
\end{aligned}
$$

where in the last line we used the fact that H_p and H have compact supports and Assumption 4.5. □

D.3.4 Boundary case: diverging subsample aspect ratio

Proposition D.3.6 (Risk approximation when $\phi_s \to +\infty$)**.** *Under Assumptions 4.1-4.5, it holds for all $M \in \mathbb{N}$*

$$
R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\infty),
$$

as $k,n,p\to\infty$, $p/n\to\phi\in(0,\infty)$ and $p/k\to\infty$, where

$$
\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\infty) := \lim_{\phi_s\to+\infty} \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = \sigma^2 + \rho^2 \int r\,\mathrm{d}G(r) \qquad \text{(D.32)}
$$

and $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)$ is defined in Theorem D.3.1.

Proof of Proposition D.3.6. Note that

$$
\begin{aligned}
R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) &= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(y_0 - \boldsymbol{x}_0^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2] \\
&= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\epsilon}_0 + \boldsymbol{x}_0^\top(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)))^2] \\
&= \sigma^2 + \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{x}_0\boldsymbol{x}_0^\top(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))] \\
&= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)).
\end{aligned}
$$

Then, by the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - (\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0 + \sigma^2) &= \|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 - 2\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0 \\
&\le \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2\|\boldsymbol{\Sigma}\|_2 \\
&\le \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2r_{\max}\rho\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2,
\end{aligned}
$$

almost surely as $k,n,p\to$ and $p/k\to\infty$. Thus, we have the following holds almost surely:

$$
\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2 &\le \frac{1}{M}\sum_{m=1}^M \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{y}/k)\|_2 \\
&\le \frac{1}{M}\sum_{m=1}^M \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\| \cdot \|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2 \\
&\le C\sqrt{\rho^2+\sigma^2}\cdot\frac{1}{M}\sum_{m=1}^M \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|,
\end{aligned}
$$

where the last inequality holds eventually almost surely since Assumptions 4.1-4.3 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]} = C\sqrt{\rho^2+\sigma^2}$ for some constant C. Observe that operator norm of the matrix $(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}\boldsymbol{L}_m/\sqrt{k}$ is upper bounded $\max_i s_i/(s_i^2+\lambda) \le 1/s_{\min}$ where s_i's are the singular values of $\boldsymbol{X}$ and $s_{\min}$ is the smallest nonzero singular value. As $k,p\to\infty$ such that $p/k\to\infty$, $s_{\min}\to\infty$ almost surely (e.g., from results in Bloemendal et al. (2016)) and therefore, $\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2 \to 0$ almost surely. Thus, we have shown that

$$
R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \sigma^2 + \boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0.
$$

or equivalently

$$R(\widetilde{f}^{\texttt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \sigma^2 + \rho^2 \int r \,\mathrm{d}G(r).$$

From Lemma D.8.13, we have

$$\lim_{\phi_s \to +\infty} v(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} \widetilde{v}_b(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} \widetilde{v}_v(-\lambda; \phi_s).$$

Thus,

$$\lim_{\phi_s \to +\infty} V_\lambda(\phi, \phi_s) = \lim_{\phi_s \to +\infty} V_\lambda(\phi, \phi_s) = 0$$

and

$$\lim_{\phi_s \to +\infty} B_\lambda(\phi, \phi_s) = \lim_{\phi_s \to +\infty} B_\lambda(\phi, \phi_s) = \rho^2 \int r \,\mathrm{d}G(r).$$

Therefore, we have $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \infty) := \lim_{\phi_s \to +\infty} \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \rho^2 \int r \,\mathrm{d}G(r)$. Thus, $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \infty)$ is well defined and $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ is right continuous at $\phi_s = +\infty$. □

D.4 Proof of Theorem 4.4.1 (subagging with replacement, ridgeless predictor)

As done in Appendix D.3, for proving the asymptotic conditional risks, we will treat $\mathcal{I}_k$ or $\mathcal{I}_k^\sigma$ as fixed. We will use $\widetilde{f}^{\texttt{WR}}_{0,M}$ to denote the ingredient predictor associated with regularization parameter $\lambda = 0$.

D.4.1 Proof assembly

We first explicitly write out the statement of Theorem 4.4.1 for the ridgeless case of $\lambda = 0$. As in Appendix D.3, we obtain the risk decomposition for general M though it suffices to analyze the case $M = 2$ according to Theorem 4.3.9.

For ridgeless predictors ($\lambda = 0$) and $\theta > 1$, the scalar $v(0; \theta)$ is the unique fixed-point solution to the following equation:

$$v(0;\theta)^{-1} = \theta \int r(1 + v(0;\theta)r)^{-1} \,\mathrm{d}H(r). \tag{D.33}$$

and the nonnegative constants $\widetilde{v}(0; \vartheta, \theta)$ and $\widetilde{c}(0; \theta)$ are defined via the following equations:

$$\widetilde{v}(0;\vartheta,\theta) = \frac{\vartheta \int r^2 (1 + v(0;\theta)r)^{-2} \,\mathrm{d}H(r)}{v(0;\theta)^{-2} - \vartheta \int r^2 (1 + v(0;\theta)r)^{-2} \,\mathrm{d}H(r)}, \qquad \widetilde{c}(0;\theta) = \int r(1 + v(0;\theta)r)^{-2} \,\mathrm{d}G(r). \tag{D.34}$$

When $\theta \leq 1$, the quantities defined in (D.33) and (D.34) are interpreted as $\lim_{\lambda \to 0^+} v(-\lambda; \theta) = \infty$, $\lim_{\lambda \to 0^+} \widetilde{c}(-\lambda; \theta) = 0$ and $\lim_{\lambda \to 0^+} \widetilde{v}(-\lambda; \vartheta, \theta) = \vartheta(1 - \vartheta)^{-1}$.

Theorem D.4.1 (Risk characterization of subagged ridgeless predictor)**.** *Let $\widetilde{f}^{\texttt{WR}}_{0,M}$ be the ingredient predictor as defined in (4.18). Suppose Assumptions 4.1-4.5 hold for the dataset $\mathcal{D}_n$. Then, as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty]$ and $\phi_s \neq 1$, there exists a deterministic function $\mathscr{R}^{\texttt{sub}}_{0,M}(\phi, \phi_s)$, $M \in \mathbb{N}$, such that for $I_1, \ldots, I_M \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\sup_{M \in \mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{0,M}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) - \mathscr{R}^{\texttt{sub}}_{0,M}(\phi, \phi_s)| \xrightarrow{\text{p}} 0,$$

and

$$\sup_{M \in \mathbb{N}} |R(\widetilde{f}_{0,M}; \mathcal{D}_n) - \mathscr{R}^{\texttt{sub}}_{0,M}(\phi, \phi_s)| \xrightarrow{\text{a.s.}} 0.$$

Furthermore, the function $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)$ decomposes as $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) = \sigma^2 + \mathscr{B}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) + \mathscr{V}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)$, where the terms are given by $\mathscr{B}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) = M^{-1}B_0(\phi_s,\phi_s) + (1-M^{-1})B_0(\phi,\phi_s)$, and $\mathscr{V}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) = M^{-1}V_0(\phi_s,\phi_s) + (1-M^{-1})V_0(\phi,\phi_s)$, and the functions $B_0(\cdot,\cdot)$ and $V_0(\cdot,\cdot)$ are defined as

$$B_0(\vartheta,\theta) = \begin{cases} 0 & \theta\in(0,1),\ \vartheta\leq\theta \\ \rho^2(1+\widetilde{v}(0;\vartheta,\theta)\widetilde{c}(0;\theta) & \theta\in(1,\infty],\ \vartheta\leq\theta \end{cases}, \qquad V_0(\vartheta,\theta) = \begin{cases} \sigma^2\dfrac{\vartheta}{1-\vartheta} & \theta\in(0,1),\ \vartheta\leq\theta \\ \sigma^2\widetilde{v}(0;\vartheta,\theta) & \theta\in(1,\infty],\ \vartheta\leq\theta \end{cases},$$

where the nonnegative constants $\widetilde{v}(0;\vartheta,\theta)$ and $\widetilde{c}(0;\theta)$ are as defined in (D.34).

Proof of Theorem D.4.1. We use the same notations as in the proof for Theorem D.3.1 and let $\widehat{\mathbf{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$ for all $m\in[M]$. Note that

$$\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \frac{1}{M}\sum_{m=1}^M(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\boldsymbol{\beta}_0 - \frac{1}{M}\sum_{m=1}^M\widehat{\mathbf{\Sigma}}_m^+\frac{\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{\epsilon}}{k}.$$

We have

$$\begin{aligned} R(\widetilde{f}^{\mathtt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) &= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top\mathbf{\Sigma}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)) \\ &= \sigma^2 + T_B + T_V + T_C, \end{aligned}$$

where

$$T_C = -\frac{2}{M^2}\boldsymbol{\epsilon}^\top\left(\sum_{m=1}^M\widehat{\mathbf{\Sigma}}_m^+\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)^\top\mathbf{\Sigma}\left(\sum_{m=1}^M(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\right)\boldsymbol{\beta}_0, \tag{D.35}$$

$$T_B = \frac{1}{M^2}\boldsymbol{\beta}_0^\top\left(\sum_{m=1}^M(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\right)\mathbf{\Sigma}\left(\sum_{m=1}^M(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\right)\boldsymbol{\beta}_0, \tag{D.36}$$

$$T_V = \frac{1}{M^2}\boldsymbol{\epsilon}^\top\left(\sum_{m=1}^M\widehat{\mathbf{\Sigma}}_m^+\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)^\top\mathbf{\Sigma}\left(\sum_{m=1}^M\widehat{\mathbf{\Sigma}}_m^+\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{\epsilon}. \tag{D.37}$$

Next we analyze the three term separately for $M\in\{1,2\}$. From Lemma D.4.2, we have that $T_C \xrightarrow{\text{a.s.}} 0$. Further, from Lemma D.4.4, Lemma D.4.5, and Lemma D.4.6, for all $I_1\in\mathcal{I}_k$ when $M=1$ and for all $I_m, I_l \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k$ when $M=2$, it holds that

$$R(\widetilde{f}_{M,\lambda};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)$$

as $n,k,p\to\infty$, $p/n\to\phi\in(0,\infty)$, and $p/k\to\phi_s\in[\phi,\infty)\setminus\{1\}$, where

$$\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) = \sigma^2 + \frac{1}{M}(B_0(\phi_s,\phi_s) + V_0(\phi_s,\phi_s)) + \frac{M-1}{M}(B_0(\phi,\phi_s) + V_0(\phi,\phi_s)).$$

Here, the components are:

$$B_0(\phi,\phi_s) = \begin{cases} 0, & \phi_s\in(0,1) \\ \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s), & \phi_s\in(1,\infty) \end{cases}, \qquad V_0(\phi,\phi_s) = \begin{cases} \sigma^2\frac{\phi}{1-\phi}, & \phi_s\in(0,1) \\ \sigma^2\widetilde{v}(0;\phi,\phi_s), & \phi_s\in(1,\infty) \end{cases},$$

and the nonnegative constants $\widetilde{v}(0;\phi,\phi_s)$ and $\widetilde{c}(0;\phi_s)$ are as defined in (D.34). The proof for the boundary case when $\phi_s=\infty$ follows from Proposition D.4.7. Then, we have that the function $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)$ is continuous on $[\phi,\infty]\setminus\{1\}$ and lower-semi continuous on $[\phi,\infty]$.

Finally, the risk expression for general M and the uniformity claim over $M\in\mathbb{N}$ follow from Theorem 4.3.9. □

D.4.2 Component concentrations

In this subsection, we will show that the cross-term C_0 converges to zero and the variance term T_V converge to its corresponding trace expectation.

D.4.2.1 Convergence of the cross term

Lemma D.4.2 (Convergence of the cross term)**.** *Under Assumptions 4.1-4.5, for T_C as defined in (D.35), we have $T_C \xrightarrow{\text{a.s.}} 0$ as $k, p \to \infty$ and $p/k \to \phi_s \in (0,1) \cup (1,\infty)$,*

Proof of Lemma D.4.2. Note that

$$T_C = -\frac{2}{M^2} \cdot \frac{1}{k} \left\langle \left(\sum_{m=1}^{M} \widehat{\boldsymbol{\Sigma}}_m^{+} \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^{+} \widehat{\boldsymbol{\Sigma}}_m) \right) \boldsymbol{\beta}_0, \boldsymbol{\epsilon} \right\rangle .$$

We next bound the norm

$$\begin{aligned}
&\frac{1}{k} \left\| \frac{1}{M} \left(\sum_{m=1}^{M} \widehat{\boldsymbol{\Sigma}}_m^{+} \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^{+} \widehat{\boldsymbol{\Sigma}}_m) \right) \boldsymbol{\beta}_0 \right\|_2^2 \\
&\leq \frac{1}{M^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \frac{1}{k} \left\| (\widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{X}^\top \boldsymbol{L}_m)^\top \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l) \boldsymbol{\beta}_0 \right\|_2^2 \\
&\leq \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{m=1}^{M} \sum_{l=1}^{M} \left\| (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l) \boldsymbol{\Sigma} \boldsymbol{\Sigma}_m^{+} \boldsymbol{\Sigma}_m \widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l) \right\|_{\text{op}} \\
&\leq \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{j=1}^{M} \sum_{l=1}^{M} \left\| \boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l \right\|_{\text{op}}^2 \|\boldsymbol{\Sigma}\|_{\text{op}}^2 \left\| \widehat{\boldsymbol{\Sigma}}_j^{+} \boldsymbol{\Sigma}_j \widehat{\boldsymbol{\Sigma}}_j^{+} \right\|_{\text{op}} \\
&= \frac{\|\boldsymbol{\beta}_0\|_2^2}{M^2} \cdot \sum_{j=1}^{M} \sum_{l=1}^{M} \left\| \boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l \right\|_{\text{op}}^2 \|\boldsymbol{\Sigma}\|_{\text{op}}^2 \left\| \widehat{\boldsymbol{\Sigma}}_j^{+} \right\|_{\text{op}} \\
&\leq \|\boldsymbol{\beta}_0\|_2^2 \, r_{\max}^2 \cdot \frac{1}{M} \sum_{j=1}^{M} \left\| \widehat{\boldsymbol{\Sigma}}_m^{+} \right\|_{\text{op}} ,
\end{aligned}$$

where the last inequality is due to Assumption 4.4 and the fact that $\left\| \boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^{+} \widehat{\boldsymbol{\Sigma}}_l \right\|_{\text{op}} \leq 1$. By Assumption 4.3, $\|\boldsymbol{\beta}_0\|_2^2$ is uniformly bounded in p. From Bai and Silverstein (2010), $\liminf \min_{1 \leq i \leq p} s_i^2 \geq r_{\min}(1 - \sqrt{\phi_s})^2$ almost surely for $\phi_s \in (0,1) \cup (1,\infty)$. Thus, $\limsup \left\| \widehat{\boldsymbol{\Sigma}}_m^{+} \right\|_{\text{op}} \leq C$ for some constant $C < \infty$ almost surely. Applying Lemma D.9.4, we thus have that $T_C \xrightarrow{\text{a.s.}} 0$. □

D.4.2.2 Convergence of the variance term

Lemma D.4.3 (Convergence of the variance term)**.** *Under Assumptions 4.1-4.5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$ and $\boldsymbol{L}_m \in \mathbb{R}^{n \times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

1. *for all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$\frac{1}{k^2} \boldsymbol{\epsilon}^\top \boldsymbol{L}_m \boldsymbol{X} \widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{\epsilon} - \frac{\sigma^2}{k} \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j^{+} \boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0,$$

2. *for all $m, l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\frac{1}{k^2} \boldsymbol{\epsilon}^\top \boldsymbol{L}_m \boldsymbol{X} \widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}_l^{+} \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{\epsilon} - \frac{\sigma^2}{k^2} \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_l^{+} \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \widehat{\boldsymbol{\Sigma}}_m^{+} \boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0,$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, *and* $p/k \to \phi_s \in [\phi, \infty) \setminus \{1\}$.

Proof of Lemma D.3.3. Note that the term is the same as the variance terms for ridge predictor trained on k i.i.d. samples $(\boldsymbol{L}_m\boldsymbol{X}, \boldsymbol{L}_m\boldsymbol{y})$. Notice that $\boldsymbol{L}_m\boldsymbol{\epsilon}$ is independent of $\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m$, and

$$\frac{1}{k}\left\|\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m\right\|_{\mathrm{op}} \le \left\|\widehat{\boldsymbol{\Sigma}}_m^+\right\|_{\mathrm{op}}^2\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}}\|\boldsymbol{\Sigma}\|_{\mathrm{op}} \le r_{\max}\left\|\widehat{\boldsymbol{\Sigma}}_m^+\right\|_{\mathrm{op}}^2\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}}.$$

Observe that $\liminf\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}} \ge \liminf\min_{1\le i\le p} s_i^2 \ge r_{\max}(1-\sqrt{\phi_s})^2$ and $\limsup\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}} \le \limsup\max_{1\le i\le p} s_i^2 \le r_{\max}(1+\sqrt{\phi_s})^2$ almost surely as $k, p \to \infty$ and $p/k \to \phi_s \in (0,1)\cup(1,\infty)$ from Bai and Silverstein (2010). We have $\limsup\left\|\widehat{\boldsymbol{\Sigma}}_m^+\right\|_{\mathrm{op}} \le C$ and $\limsup\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}} \le C$ for some constant $C < \infty$ almost surely. From Lemma D.9.5, it follows that

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{\epsilon} - \frac{\sigma^2}{k^2}\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m) \xrightarrow{\text{a.s.}} 0.$$

Since $\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m)/k^2 = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^+\widehat{\boldsymbol{\Sigma}}_m\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma})/k = \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma})/k$, we further have

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{\epsilon} - \frac{\sigma^2}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0. \tag{D.38}$$

The second term involves the cross term $\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_l$. Note that

$$\frac{1}{n}\left\|\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_l\right\|_{\mathrm{op}} \le \frac{k}{n}r_{\max}\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}}^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}}^{\frac{1}{2}}\left\|\widehat{\boldsymbol{\Sigma}}_l^+\right\|_{\mathrm{op}}\left\|\widehat{\boldsymbol{\Sigma}}_l^+\right\|_{\mathrm{op}},$$

because $\left\|\widehat{\boldsymbol{\Sigma}}_m^+\right\|_{\mathrm{op}}$ and $\left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\mathrm{op}}$ for $m \in [M]$ are uniformly bounded almost surely. By Lemma D.9.5, it follows that

$$\frac{1}{n}\boldsymbol{\epsilon}^\top\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_l\boldsymbol{\epsilon} - \frac{\sigma^2}{n}\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_l) \xrightarrow{\text{a.s.}} 0.$$

Since $k/n \to \phi_s/\phi$, we have

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_l\boldsymbol{\epsilon} - \frac{\sigma^2}{k^2}\operatorname{tr}(\boldsymbol{L}_m\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_l) \xrightarrow{\text{a.s.}} 0.$$

□

D.4.3 Component deterministic approximations

D.4.3.1 Deterministic approximation of the bias functional

Lemma D.4.4 (Deterministic approximation of the bias functional)**.** *Under Assumptions 4.1-4.5, for all* $m \in [M]$ *and* $I_m \in \mathcal{I}_k$, *let* $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$ *and* $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ *be a diagonal matrix with* $(\boldsymbol{L}_m)_{ll} = 1$ *if* $l \in I_m$ *and 0 otherwise. Then, it holds that*

1. *for all* $m \in [M]$ *and* $I_m \in \mathcal{I}_k$,

$$\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+\widehat{\boldsymbol{\Sigma}}_m)\boldsymbol{\Sigma}(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+\widehat{\boldsymbol{\Sigma}}_m)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}(0;\phi_s,\phi_s))\widetilde{c}(0;\phi_s) & \phi_s \in (1,\infty), \end{cases}$$

2. *for all* $m, l \in [M]$, $m \neq l$ *and* $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,

$$\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+\widehat{\boldsymbol{\Sigma}}_m)\boldsymbol{\Sigma}(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_l^+\widehat{\boldsymbol{\Sigma}}_l)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s) & \phi_s \in (1,\infty), \end{cases}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty) \setminus \{1\}$, where the nonnegative constants $\widetilde{v}(0; \phi, \phi_s)$ and $\widetilde{c}(0; \phi_s)$ are as defined in (D.34).

Proof of Lemma D.4.4. For the first term, we have that for $m \in [M]$,

$$\beta_0^\top (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m) \beta_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \text{if } \phi_s \in (0,1) \\ \rho^2 (1 + \widetilde{v}_b(0; \phi_s)) \int \frac{r}{(1+v(0;\phi_s)r)^2} \, \mathrm{d}G(r) & \text{if } \phi_s \in (1, \infty). \end{cases} \tag{D.39}$$

Next we analyze the second term, by considering the following two cases separately for $(m, l) = (1, 2)$.
(1) $\phi_s \in (0, 1)$. Since the singular values of $\widehat{\boldsymbol{\Sigma}}_j$'s are almost surely lower bounded away from 0, we have $\widehat{\boldsymbol{\Sigma}}_j^+ \widehat{\boldsymbol{\Sigma}}_j = \boldsymbol{I}_p$ almost surely. Then $\beta_0^\top (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_1^+ \widehat{\boldsymbol{\Sigma}}_1) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_2^+ \widehat{\boldsymbol{\Sigma}}_2) \beta_0 \xrightarrow{\text{a.s.}} 0$ when $k, p \to \infty$ and $p/k \to \phi_s \in (0, 1)$.
(2) $\phi_s \in (1, \infty)$. We begin with analyzing the deterministic equivalent of $(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_1^+ \widehat{\boldsymbol{\Sigma}}_1) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_2^+ \widehat{\boldsymbol{\Sigma}}_2)$. Recall that i_0 is the number of shared samples between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$, and $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1 \boldsymbol{L}_2 \boldsymbol{X}^\top / i_0$ and $\widehat{\boldsymbol{\Sigma}}_j^{\text{ind}} = \boldsymbol{X}^\top (\boldsymbol{L}_j - \boldsymbol{L}_1 \boldsymbol{L}_2) \boldsymbol{X}^\top / (k - i_0)$ are the common and individual covariance estimators of the two datasets. Also note that from (D.25), we have $\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}}$, where

$$(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} = (1 + \widetilde{v}(-\lambda; \phi_s, \phi)) \, (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2} \, \boldsymbol{\Sigma} > 0, \tag{D.40}$$

and $\widetilde{v}(-\lambda; \phi_s, \phi)$ is as defined in (D.34). Let $\lambda \in \Lambda = [0, \lambda_{\max}]$ where $\lambda_{\max} < \infty$. For any matrix $\boldsymbol{T} \in \mathbb{R}^{p \times p}$ with trace norm uniformly bounded by M,

$$|\operatorname{tr}[\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{T}]| \le \lambda^2 \left\| \boldsymbol{M}_1 \right\|_{\text{op}} \left\| \boldsymbol{M}_2 \right\|_{\text{op}} \left\| \boldsymbol{\Sigma} \right\|_{\text{op}} |\operatorname{tr}[(\boldsymbol{T}^\top \boldsymbol{T})^{\frac{1}{2}}]| \le M r_{\max} \left\| \boldsymbol{\Sigma} \right\|_{\text{op}}$$

where the second inequality holds because $\left\| \boldsymbol{M}_1 \right\|_{\text{op}} \le \lambda^{-1}$ and $\left\| \boldsymbol{\Sigma} \right\|_{\text{op}} \le r_{\max}$. Since $\phi_0 \ge \phi_s > 1$, it follows from Patil et al. (2022a, Lemma S.6.14) that, there exists $M' > 0$ such that the magnitudes of $v(-\lambda; \phi_s)$ and $v_b(\lambda, \phi_s, \phi) - 1$, and their derivatives with respect to λ are continuous and bounded by M' for all $\lambda \in \Lambda$. Thus, we get

$$\begin{aligned} \left| \operatorname{tr} \left[(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} \boldsymbol{T} \right] \right| &\le (1 + M') \left\| v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p \right\|_{\text{op}}^{-2} \left\| \boldsymbol{\Sigma} \right\|_{\text{op}} |\operatorname{tr}[(\boldsymbol{T}^\top \boldsymbol{T})^{\frac{1}{2}}]| \\ &\le (1 + M') M r_{\max}. \end{aligned}$$

Similarly, in the same interval the derivatives of $\operatorname{tr}\left[\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{T}\right]$ and $\operatorname{tr}\left[(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} \boldsymbol{T}\right]$ with respect to λ also have bounded magnitudes for $\lambda \in \Lambda$. Therefore, the family of functions

$$\operatorname{tr}[\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{T}] - \operatorname{tr}\left[(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} \boldsymbol{T}\right]$$

forms an equicontinuous family in λ over $\lambda \in \Lambda$. Thus, the convergence in Part 1 of Lemma D.8.9 is uniform in λ. We can now use the Moore-Osgood theorem and the continuity property from Lemma D.8.15 to interchange the limits to obtain

$$\begin{aligned} &\lim_{p \to \infty} \operatorname{tr}\left[(\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_1^+ \widehat{\boldsymbol{\Sigma}}_1) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_2^+ \widehat{\boldsymbol{\Sigma}}_2) \boldsymbol{T}\right] - \operatorname{tr}\left[((\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_1^+ \widehat{\boldsymbol{\Sigma}}_1) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_2^+ \widehat{\boldsymbol{\Sigma}}_2))^{\text{det}} \boldsymbol{T}\right] \\ &= \lim_{p \to \infty} \lim_{\lambda \to 0^+} \operatorname{tr}\left[\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{T}\right] - \operatorname{tr}\left[(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} \boldsymbol{T}\right] \\ &= \lim_{\lambda \to 0^+} \lim_{p \to \infty} \operatorname{tr}\left[\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \boldsymbol{T}\right] - \operatorname{tr}\left[(\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\text{det}} \boldsymbol{T}\right] \\ &= 0, \end{aligned}$$

where

$$((\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_1^+ \widehat{\boldsymbol{\Sigma}}_1) \boldsymbol{\Sigma} (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_2^+ \widehat{\boldsymbol{\Sigma}}_2))^{\text{det}} = (1 + \widetilde{v}(0; \phi, \phi_s))(v(0; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2} \boldsymbol{\Sigma}.$$

As $p \to \infty$, replacing the empirical distribution $G_p(r)$ by limiting distribution $G(r)$ yields the desired results. □

D.4.3.2 Deterministic approximation of the variance functional

Lemma D.4.5 (Deterministic approximation of the variance functional when $\phi_s < 1$)**.** *Under Assumptions 4.1-4.5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$ and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

1. *for all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^{+}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\phi_s}{1-\phi_s},$$

2. *for all $m, l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_l^{+}\boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \widehat{\boldsymbol{\Sigma}}_m^{+}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\phi}{1-\phi}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty) \cap (0, 1)$.

Proof of Lemma D.4.5. For the first term, from Patil et al. (2022a, Proposition S.3.2) we have that for $m \in [M]$,

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^{+}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \begin{cases} \dfrac{\phi_s}{1-\phi_s} & \text{if } \phi_s \in (0,1) \\ \phi_s v_v(0;\phi,\phi_s)\displaystyle\int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r) & \text{if } \phi_s \in (1,\infty) \end{cases}. \tag{D.41}$$

Next we analyze the second term for $\phi_s \in (0, 1)$. It suffices to analyze the case when $(m, l) = (1, 2)$. From Bai and Silverstein (2010), we have

$$r_{\min}(1-\sqrt{\phi_s})^2 \le \liminf \left\|\widehat{\boldsymbol{\Sigma}}_j\right\|_{\text{op}} \le \limsup \left\|\widehat{\boldsymbol{\Sigma}}_j\right\|_{\text{op}} \le r_{\max}(1+\sqrt{\phi_s})^2, \quad j = 1, 2.$$

Then $\widehat{\boldsymbol{\Sigma}}_j$'s are invertible almost surely. From Lemma D.8.11, we have that for $j = 1, 2$,

$$\widehat{\boldsymbol{\Sigma}}_j^{-1} = \left(\frac{i_0}{k}\widehat{\boldsymbol{\Sigma}}_0 + \frac{k-i_0}{k}\widehat{\boldsymbol{\Sigma}}_1^{\text{ind}}\right)^{-1} \simeq \left(\frac{i_0}{k}\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\boldsymbol{\Sigma}\right)^{-1},$$

where $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1 \boldsymbol{L}_2 \boldsymbol{X}/i_0$ and $\widehat{\boldsymbol{\Sigma}}_j^{\text{ind}} = \boldsymbol{X}^\top \boldsymbol{L}_j \boldsymbol{X}/(k-i_0)$ for $j = 1, 2$, defined analogously as in the proof for Theorem D.3.1. Thus, conditional on $\widehat{\boldsymbol{\Sigma}}_0$ and i_0, we have

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_1^{-1}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma} &\simeq \left(\frac{i_0}{k}\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\boldsymbol{\Sigma}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(\frac{i_0}{k}\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\boldsymbol{\Sigma}\right)^{-1} \\ &= \frac{i_0^2}{k^2}\left(\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{i_0}\boldsymbol{\Sigma}\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{i_0}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Sigma} \end{aligned}$$

by applying the conditional product rule from Proposition D.8.6. When $i_0 < k$, let $\widehat{\boldsymbol{\Sigma}}' = c\boldsymbol{\Sigma}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $c = (1-\phi_s)(k-i_0)/i_0$, we further have

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_1^{-1}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma} &\simeq \frac{i_0^2}{k^2c^2}\boldsymbol{\Sigma}^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}' + \boldsymbol{I}_p)^{-1}\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\Sigma}}' + \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma} \\ &\simeq \frac{i_0^2}{k^2}\widetilde{v}_v(-1;\gamma_0, c^{-1}\boldsymbol{I}_p)(v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)+c)^{-2}\boldsymbol{I}_p, \end{aligned}$$

where $\gamma_0 = p/i_0$, the second equality is from Lemma D.8.9 (2) and the fixed point solutions are defined by

$$\frac{1}{v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)} = 1 + \frac{\gamma_0}{c+v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)}$$

$$\frac{1}{\widetilde{v}_v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)} = \frac{1}{v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)^2} - \frac{\gamma_0}{(c+v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p))^2}.$$

When $i_0 = k$, the above equivalent is also valid, which reduces to the case for $\widehat{\boldsymbol{\Sigma}}_j^+\widehat{\boldsymbol{\Sigma}}_j\widehat{\boldsymbol{\Sigma}}_j^+$ as in (D.41). Note that from Lemma D.8.13, $\widetilde{v}_v(-\lambda;\gamma)$ and $v(-\lambda;\gamma)$ are continuous on γ, and from Lemma D.9.3, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$ where $\phi_s \in (0,\infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k,p\to\infty$. We have

$$\widehat{\boldsymbol{\Sigma}}_1^{-1}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma} \simeq \frac{\phi_s^2}{\phi_0^2}\widetilde{v}_v(-1;\phi_0,c_0^{-1}\boldsymbol{I}_p)(v(-1;\phi_0,c_0^{-1}\boldsymbol{I}_p)+c_0)^{-2}\boldsymbol{I}_p,$$

where $c_0 = \lim_{p\to\infty} c = (1-\phi_s)(\phi_s-\phi)/\phi$ and the fixed solutions reduce to

$$v(-1;\gamma_0,c_0^{-1}\boldsymbol{I}_p) = 1-\phi_s, \qquad \widetilde{v}_v(-1;\gamma_0,c_0^{-1}\boldsymbol{I}_p) = \frac{(1-\phi_s)^2}{1-\phi}.$$

Then, we have

$$\frac{i_0}{k^2}\operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_1^+\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\Sigma}] \xrightarrow{\text{a.s.}} \lim_{p\to\infty}\frac{i_0 p}{k^2}\cdot\frac{1}{p}\operatorname{tr}\left[\frac{\phi_s^2}{\phi^2}\frac{(1-\phi_s)^2}{1-\phi}\left(1-\phi_s+\frac{(1-\phi_s)(\phi_s-\phi)}{\phi}\right)^{-2}\boldsymbol{I}_p\right] = \frac{\phi}{1-\phi}. \tag{D.42}$$

Combining (D.41) and (D.42), the conclusion follows. □

Lemma D.4.6 (Deterministic approximation of the variance functional when $\phi_s > 1$)**.** *Under Assumptions 4.1-4.5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$ and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

1. *for all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j^+\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{1}{2}\widetilde{v}(0;\phi_s,\phi_s),$$

2. *for all $m,l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^+\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon} \xrightarrow{\text{a.s.}} \frac{1}{2}\widetilde{v}(0;\phi,\phi_s),$$

as $n,k,p\to\infty$, $p/n\to\phi\in(0,\infty)$, and $p/k\to\phi_s\in[\phi,\infty)\cap(1,\infty)$, where the nonnegative constants $v(0;\phi_s)$ and $\widetilde{v}(0;\phi,\phi_s)$ are as defined in (D.33) and (D.34).

Proof of Lemma D.4.6. From (D.41) we have

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \widetilde{v}(0;\phi,\phi_s). \tag{D.43}$$

For the second term, it suffices to consider the case when $(m,l) = (1,2)$. Let $P_0 = \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_2^+\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2$ and $P_\lambda = \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\ \boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2$ where $\boldsymbol{M}_j = (\widehat{\boldsymbol{\Sigma}}_j+\lambda\boldsymbol{I}_p)^{-1}$. Note that $\lim_{\lambda\to 0^+} P_\lambda = P_0$. Note that $\lim_{\lambda\to 0^+} P_\lambda = P_0$. From Lemma D.3.3 and Lemma D.3.5, we have that for any fixed $\lambda > 0$,

$$P_\lambda \xrightarrow{\text{a.s.}} Q_\lambda := \widetilde{v}(-\lambda;\phi,\phi_s),$$

as $n,k,p\to\infty$, $p/n\to\phi\in(0,\infty)$, and $p/k\to\phi_s\in[\phi,\infty)\setminus\{1\}$, where $\widetilde{v}(\lambda,\phi_s,\phi)$ is as defined in (D.13). Because of the continuity of $\widetilde{v}_v(-\lambda;\phi)$ and $v(-\lambda;\phi)$ in λ from Lemma D.8.15, we have that

$$\lim_{\lambda\to 0^+} Q_\lambda = Q_0 := \widetilde{v}(0;\phi,\phi_s).$$

As $n, p \to \infty$, we have that almost surely

$$|P_\lambda| = \phi|\operatorname{tr}(\boldsymbol{M}_2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_1\boldsymbol{\Sigma})/p| \leq \phi\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\|_{\mathrm{op}}\|\boldsymbol{\Sigma}\|_{\mathrm{op}} \leq \frac{\phi_s^2 r_{\max}}{\phi},$$

where the last inequality is because $\left\|\widehat{\boldsymbol{\Sigma}}_0\right\|_{\mathrm{op}} \leq r_{\max}$, and

$$\left\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\right\|_{\mathrm{op}} \leq \frac{k^2}{i_0^2} \cdot \max_i \frac{l_i}{\left(l_i + \frac{k-i_0}{i_0}\lambda\right)^2} \leq \frac{k^2}{i_0^2}, \tag{D.44}$$

where l_i's are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_0$. Similarly, we have $|P_0|$ is almost surely bounded. Thus, $|P_\lambda|$ is almost surely bounded over $\lambda \in \Lambda[0, \lambda_{\max}]$ for some constant $\lambda_{\max} > 0$. Next we consider the derivative

$$\begin{aligned}\frac{\partial P_\lambda}{\partial \lambda} &= \boldsymbol{\epsilon}^\top \boldsymbol{L}_1\boldsymbol{X}\frac{\partial \boldsymbol{M}_1}{\partial \lambda}\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2 + \boldsymbol{\epsilon}^\top \boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\frac{\partial \boldsymbol{M}_2}{\partial \lambda}\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2\\ &= -\boldsymbol{\epsilon}^\top \boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1^2\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2 - \boldsymbol{\epsilon}^\top \boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2^2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2\end{aligned}$$

Note that for $\lambda \in \Lambda$, we can bound

$$\left\|\boldsymbol{M}_1^2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\right\|_{\mathrm{op}} \leq \frac{k^2}{i_0^2} \cdot \max_i \frac{l_i}{\left(l_i + \frac{k-i_0}{i_0}\lambda\right)^3} \leq \frac{k^2}{i_0^2},$$

where l_i's are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_0$. Similarly, we have that $\left\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2^2\right\|_{\mathrm{op}}$ is almost surely bounded for $\lambda \in \Lambda$. By similar argument as in Lemma D.4.3, the following holds almost surely as $n, p \to \infty$,

$$\left|\frac{\partial P_\lambda}{\partial \lambda}\right| = \phi|\operatorname{tr}(\boldsymbol{M}_1^2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2^2\boldsymbol{\Sigma})| \leq \frac{\phi_s^2 r_{\max}}{\phi}.$$

That is, $|\partial P_\lambda/\partial\lambda|$ is almost surely bounded over $\lambda \in \Lambda[0, \lambda_{\max}]$.

Since $\phi_0 \geq \phi_s > 1$, it follows from Patil et al. (2022a, Lemma S.6.14) that, there exists $M' > 0$ such that the magnitudes of $v(-\lambda; \phi_s)$ and $v_v(\lambda, \phi_s, \phi)/\phi$, and their derivatives with respect to λ are continuous and bounded by M' for all $\lambda \in \Lambda$. Thus, $|Q_\lambda| \leq \phi_0 M' r_{\max}^2$ over $\lambda \in \Lambda$. Similarly, we have $|\partial Q_\lambda/\partial\lambda|_{\lambda=0^+}|$ are uniformly bounded over $\lambda \in \Lambda$. We can now use the Moore-Osgood theorem and the continuity property from Lemma D.8.15 to interchange the limits to obtain

$$\lim_{p\to\infty} P_0 - Q_0 = \lim_{p\to\infty}\lim_{\lambda\to 0^+} P_\lambda - Q_\lambda = \lim_{\lambda\to 0^+}\lim_{p\to\infty} P_\lambda - Q_\lambda = 0,$$

and the conclusion follows. □

D.4.4 Boundary case: diverging subsample aspect ratio

Proposition D.4.7 (Risk approximation when $\phi_s \to \infty$). *Under Assumptions 4.1-4.5, it holds for all $M \in \mathbb{N}$*

$$R(\widetilde{f}_{0,M}^{\texttt{WR}}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}_{0,M}^{\texttt{sub}}(\phi, \infty),$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to \infty$, where

$$\mathscr{R}_{0,M}^{\texttt{sub}}(\phi, \infty) := \lim_{\phi_s\to\infty} \mathscr{R}_{0,M}^{\texttt{sub}}(\phi, \phi_s) = \sigma^2 + \rho^2 \int r \,\mathrm{d}G(r), \tag{D.45}$$

and $\mathscr{R}_{0,M}^{\texttt{sub}}(\phi, \phi_s)$ is as defined in Theorem D.4.1.

Proof of Proposition D.4.7. Note that

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) &= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(y_0-\boldsymbol{x}_0^\top\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2]\\
&= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\epsilon}_0+\boldsymbol{x}_0^\top(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)))^2]\\
&= \sigma^2+\mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top\boldsymbol{x}_0\boldsymbol{x}_0^\top(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))]\\
&= \sigma^2+(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top\boldsymbol{\Sigma}(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)).
\end{aligned}$$

Then, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)-(\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0+\sigma^2) &= \|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2-2\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0\\
&\le \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2+2\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2\|\boldsymbol{\Sigma}\|_2\\
&\le \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2+2r_{\max}\rho\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2
\end{aligned}$$

almost surely as $k,n,p\to$ and $p/k\to\infty$. Thus, we have the following holds almost surely:

$$\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}^0_M(\mathcal{D}_n)\|_2 &\le \frac{1}{M}\sum_{m=1}^M\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{y}/k)\|_2\\
&\le \frac{1}{M}\sum_{m=1}^M\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|\cdot\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2\\
&\le C\sqrt{\rho^2+\sigma^2}\cdot\frac{1}{M}\sum_{m=1}^M\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|
\end{aligned}$$

where the last inequality holds eventually almost surely since Assumptions 4.1-4.3 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]}=C\sqrt{\rho^2+\sigma^2}$ for some constant C. Observe that operator norm of the matrix $(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}\boldsymbol{L}_m/\sqrt{k}$ is upper bounded $1/s_{\min}$, where $s_{\min}$ is the smallest nonzero singular value of $\boldsymbol{X}$. As $k,p\to\infty$ such that $p/k\to\infty$, $s_{\min}\to\infty$ almost surely (e.g., from results in Bloemendal et al. (2016)), and therefore, $\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2\to 0$ almost surely. Thus, we have shown that

$$R(\widetilde{f}^{\mathtt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)\xrightarrow{\text{a.s.}}\sigma^2+\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0,$$

or equivalently,

$$R(\widetilde{f}^{\mathtt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)\xrightarrow{\text{a.s.}}\sigma^2+\rho^2\int r\,\mathrm{d}G(r).$$

From Lemma D.8.14 we have

$$\lim_{\phi_s\to\infty}v(0;\phi_s)=\lim_{\phi_s\to\infty}\widetilde{v}_b(0;\phi_s)=\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s).$$

Thus,

$$\lim_{\phi_s\to\infty}V_0(\phi_s,\phi_s)=\lim_{\phi_s\to\infty}V_0(\phi,\phi_s)=0,$$

and

$$\lim_{\phi_s\to\infty}B_0(\phi_s,\phi_s)=\lim_{\phi_s\to\infty}B_0(\phi,\phi_s)=\rho^2\int r\,\mathrm{d}G(r).$$

Therefore, we have $\mathscr{R}^{\mathtt{WR}}_{0,M}(\phi,\infty):=\lim_{\phi_s\to\infty}\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)=\sigma^2+\rho^2\int r\,\mathrm{d}G(r)$. Thus, $\mathscr{R}^{\mathtt{WR}}_{0,M}(\phi,\infty)$ is well defined and $\mathscr{R}^{\mathtt{WR}}_{0,M}(\phi,\phi_s)$ is right continuous at $\phi_s=\infty$. □

D.5 Proof of Theorem 4.4.6 (splagging without replacement, ridge and ridgeless predictors)

Proof of Theorem 4.4.6. For $M \in \{1, 2, \dots, \lfloor \liminf n/k \rfloor\}$, following the proof in Theorem D.3.1, the conditional risk is given by

$$R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \sigma^2 + T_C + T_B + T_V,$$

where T_C, T_B, and T_V are defined as

$$T_C = -\frac{\lambda}{M} \cdot \boldsymbol{\epsilon}^\top \left(\sum_{m=1}^M \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^M \boldsymbol{M}_m \right) \boldsymbol{\beta}_0, \tag{D.46}$$

$$\begin{aligned} T_B &= \frac{\lambda^2}{M^2} \cdot \boldsymbol{\beta}_0^\top \left(\sum_{i=1}^M \boldsymbol{M}_{I_i} \right) \boldsymbol{\Sigma} \left(\sum_{i=1}^M \boldsymbol{M}_{I_i} \right) \boldsymbol{\beta}_0 && \text{(D.47)} \\ &= \frac{\lambda^2}{M} \sum_{i=1}^M \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i} \boldsymbol{\Sigma} \boldsymbol{M}_{I_i} \boldsymbol{\beta}_0 + \frac{\lambda^2 (M-1)}{M} \sum_{i,j=1}^M \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i} \boldsymbol{\Sigma} \boldsymbol{M}_{I_j} \boldsymbol{\beta}_0, \\ T_V &= \frac{1}{M^2} \cdot \boldsymbol{\epsilon}^\top \left(\sum_{i=1}^M \boldsymbol{M}_{I_i} \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{i=1}^M \boldsymbol{M}_{I_i} \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k} \right) \boldsymbol{\epsilon} \\ &= \frac{1}{M} \sum_{i=1}^M \left(\boldsymbol{M}_{I_i} \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k} \right)^\top \boldsymbol{\Sigma} \left(\boldsymbol{M}_{I_i} \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k} \right) + \frac{M-1}{M} \sum_{i,j=1}^M \left(\boldsymbol{M}_{I_i} \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k} \right)^\top \boldsymbol{\Sigma} \left(\boldsymbol{M}_{I_j} \frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k} \right), \end{aligned}$$

$$\text{(D.48)}$$

where $\boldsymbol{M}_{I_\ell} = (\boldsymbol{X}^\top \boldsymbol{L}_\ell \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$ and $\boldsymbol{L}_\ell$ is a diagonal matrix with diagonal entry being 1 if the ℓth sample X_ℓ is in the sub-sampled dataset $\mathcal{D}_{I_\ell}$ and 0 otherwise. Note that for splagging, $I_i \cap I_j = \varnothing$ for all $i \neq j$.

We analyze each term separately for $M \in \{1, 2\}$. From Lemma D.3.2, we have that $T_C \xrightarrow{\text{a.s.}} 0$. From Lemma D.3.3, we have that

$$T_V - \frac{1}{M} \sum_{j=1}^M \frac{\sigma^2}{k} \operatorname{tr}(\boldsymbol{M}_{I_j} \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{M}_{I_j} \boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0, \tag{D.49}$$

since the datasets have no overlaps and the cross term vanishes because $\boldsymbol{L}_l \boldsymbol{L}_m = \boldsymbol{0}_{n \times n}$ for $l \neq m$. Then, from (D.20) and (D.28), we have that for $\ell \in [M]$,

$$\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_\ell} \boldsymbol{\Sigma} \boldsymbol{M}_{I_\ell} \boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2 \widetilde{v}(-\lambda; \phi_s, \phi_s) \widetilde{c}(-\lambda; \phi_s), \tag{D.50}$$

$$\frac{\sigma^2}{k} \operatorname{tr}(\boldsymbol{M}_{I_\ell} \widehat{\boldsymbol{\Sigma}}_\ell \boldsymbol{M}_{I_\ell} \boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\sigma^2}{2} \widetilde{v}(-\lambda; \phi_s, \phi_s), \tag{D.51}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s = 2\phi$, where the positive constants $\widetilde{v}(\lambda; \phi_s, \phi)$, and $\widetilde{c}(-\lambda; \phi_s)$ are as defined in (D.13). For the cross term ($i \neq j$), setting $i_0 = 0$ in (D.22) yields that

$$\boldsymbol{M}_{I_i} \boldsymbol{\Sigma} \boldsymbol{M}_{I_j} \simeq (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} \boldsymbol{\Sigma} (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}.$$

Thus,

$$\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i} \boldsymbol{\Sigma} \boldsymbol{M}_{I_j} \boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2 \int \frac{r}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}G(r) = \rho^2 \widetilde{c}(0; \phi_s). \tag{D.52}$$

Combining (D.46)-(D.52), we have shown that $R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$, where

$$\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) + \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s),$$

and the components are:

$$\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = \frac{1}{M}B_\lambda(\phi_s,\phi_s) + \left(1-\frac{1}{M}\right)C_\lambda(\phi_s), \qquad \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)\frac{1}{M}V_\lambda(\phi_s,\phi_s),$$

with $B_\lambda(\phi,\phi_s) = \rho^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s)$, $C_\lambda(\phi_s) = \rho^2\widetilde{c}(-\lambda;\phi_s)$, and $V_\lambda(\phi,\phi_s) = \sigma^2\widetilde{v}(-\lambda;\phi,\phi_s)$.

From Proposition D.3.6 and Proposition D.4.7, we have that for all $\lambda \in [0,\infty)$ and $M \in \{1,2\}$,

$$\lim_{\phi_s\to+\infty} R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M};\{\mathcal{D}_k^{(m)}\}_{m=1}^M) = \sigma^2 + \rho^2\int r\,\mathrm{d}G(r),$$

$\lim_{\phi_s\to+\infty} B_\lambda(\phi_s,\phi_s) = \rho^2\int r\,\mathrm{d}G(r)$ and $\lim_{\phi_s\to+\infty} v(-\lambda;\phi_s) = \lim_{\phi_s\to+\infty} V_\lambda(\phi_s,\phi_s) = 0$. Then

$$\lim_{\phi_s\to+\infty}\widetilde{c}(-\lambda;\phi_s) = \lim_{\phi_s\to+\infty}\int r(1+v(-\lambda;\phi_s)r)^{-2}\,\mathrm{d}G(r) = \int r\,\mathrm{d}G(r).$$

Thus, the approximation holds when $\phi_s = \infty$: $\lim_{\phi_s\to+\infty} R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \lim_{\phi_s\to+\infty}\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)$.

Finally, the risk expression for general M and the uniform statement for all $M \leq \lfloor n/k \rfloor$ follow from Theorem 4.3.9. □

D.6 Proofs related to bagged risk properties

D.6.1 Bias-variance monotonicities in the number of bags, subagging with replacement

Proof of Proposition 4.4.5. Recall that from the proof for Theorem D.3.1, we have

$$\begin{aligned}
\mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s) &= \rho^2(1+\widetilde{v}(-\lambda,\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s) \qquad & \mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s) &= \sigma^2\widetilde{v}(-\lambda;\phi_s,\phi_s)\\
\mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s) &= \rho^2(1+\widetilde{v}(-\lambda,\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s) \qquad & \mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s) &= \sigma^2\widetilde{v}(-\lambda;\phi,\phi_s)
\end{aligned}$$

where the nonnegative constants $\widetilde{v}(-\lambda,\phi,\phi_s)$ and $\widetilde{c}(-\lambda;\phi_s)$ are defined in (D.12). Since H has positive support, $\widetilde{v}(-\lambda;\phi,\phi_s)$ is strictly increasing in ϕ, and thus, $\mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s) = \mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s)$ when $\phi_s = \phi$, and $\mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s) > \mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s)$ when $\phi_s > \phi$. Similarly, $\mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s) = \mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s)$ when $\phi_s = \phi$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi,\phi_s) < \mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s)$ when $\phi_s > \phi$. Recall that the definitions of $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = 1/M\cdot B_\lambda(\phi_s,\phi_s) + (1-1/M)B_\lambda(\phi,\phi_s)$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = 1/M\cdot V_\lambda(\phi_s,\phi_s) + (1-1/M)V_\lambda(\phi,\phi_s)$ are a convex combination of $B_\lambda(\phi,\phi_s)$ and $B_\lambda(\phi_s,\phi_s)$, and $V_\lambda(\phi,\phi_s)$ and $V_\lambda(\phi_s,\phi_s)$, respectively. The proof for ridgeless predictor follows by setting $\lambda = 0$ except $B_0(\phi,\phi_s) = B_0(\phi,\phi_s) = 0$ for $\phi_s < 1$. □

D.6.2 Bias-variance monotonicities in the number of bags, splagging without replacement

Proof of Proposition 4.4.10. For the variance term, $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s)$ as a linear function of M^{-1} is strictly decreasing in M if $\phi_s < \infty$ and is zero if $\phi_s = \infty$ or $\sigma^2 = 0$.

For the bias term, when $\phi_s > 1$, since $\widetilde{c}(-\lambda;\phi_s) > 0$, we have that $B_\lambda(\phi_s,\phi_s) \geq C_\lambda(\phi_s)$ with equality holds if and only if $\widetilde{v}(-\lambda;\phi,\phi_s) = 0$ or $\widetilde{c}(-\lambda;\phi_s) = 0$, if and only if $\phi_s = \infty$. Then we have

$$\begin{aligned}
\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) &= \frac{1}{M}B_\lambda(\phi_s,\phi_s) + \left(1-\frac{1}{M}\right)C_\lambda(\phi_s)\\
&= \frac{1}{M}(B_\lambda(\phi_s,\phi_s) - C_\lambda(\phi_s)) + C_\lambda(\phi_s)\\
&\geq \frac{1}{M+1}(B_\lambda(\phi_s,\phi_s) - C_\lambda(\phi_s)) + C_\lambda(\phi_s)\\
&= \frac{1}{M+1}B_\lambda(\phi_s,\phi_s) + \left(1-\frac{1}{M+1}\right)C_\lambda(\phi_s)
\end{aligned}$$

$$= \mathscr{B}^{\texttt{spl}}_{\lambda,M+1}(\phi,\phi_s).$$

with equality holds if $\phi_s = \infty$ or $\rho^2 = 0$. When $\phi_s < 1$, $B_\lambda(\phi_s,\phi_s) \geq C_\lambda(\phi_s)$ with equality holds if and only if $\widetilde{c}(-\lambda;\phi_s) = 0$, if and only if $\lambda = 0$. The monotonicity of $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)$ in M follows analogously.

As $M \leq \phi_s/\phi$, we further have $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) \geq \mathscr{V}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)$ and $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) \geq \mathscr{V}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)$ for all $M = 1,\ldots,\lfloor \liminf n/k \rfloor$. □

D.6.3 Risk monotonization of general bagged predictors by cross-validation

Proof of Theorem 4.5.1. We present the proof for bagging with replacement, and the proof for bagging without replacement follows by restricting the support of $\phi_s \mapsto \mathscr{R}_M(\phi,\phi_s)$ to $[M\phi,\infty]$. From Theorem 4.3.9, we have that for any $M \in \mathbb{N}$ and $\{I_\ell\}_{\ell=1}^M$ simple random samples from $\mathcal{I}_k$ or $\mathcal{I}_k^\pi$, it holds that

$$R(\widetilde{f}_M;\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{p}} \mathscr{R}_M(\phi,\phi_s)$$

as $k,n,p \to$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)$, where

$$\mathscr{R}_M(\phi,\phi_s) := (2\mathscr{R}(\phi,\phi_s) - \mathscr{R}(\phi_s,\phi_s)) + \frac{2}{M}(\mathscr{R}(\phi_s,\phi_s) - \mathscr{R}(\phi,\phi_s)).$$

From Patil et al. (2022a, Lemma 3.8 and Theorem 3.4), we have that

$$\left(R(\widehat{f}^{\text{cv}}_{M,\mathcal{I}_{\widehat{k}}};\mathcal{D}_n) - \mathscr{R}_M(\phi,\phi_s)\right)_+ \xrightarrow{\text{p}} 0.$$

In Patil et al. (2022a), we have assumed that the risk is bounded away from 0 in order to conclude that the relative error converges to 0. But in Theorem 4.5.1, we conclude only the positive part of the absolute error converges to 0, for which we do not require the risk to be bounded away from 0.

Since $\mathscr{R}_M(\phi,\phi_s)$ is increasing in ϕ for any fixed ϕ_s. For $0 < \phi_1 \leq \phi_2 < \infty$,

$$\min_{\phi_s \geq \phi_1} \mathscr{R}_M(\phi_1,\phi_s) \leq \min_{\phi_s \geq \phi_2} \mathscr{R}_M(\phi_1,\phi_s) \leq \min_{\phi_s \geq \phi_2} \mathscr{R}_M(\phi_2,\phi_s)$$

where the first inequality follows because $\{\phi_s : \phi_s \geq \phi_1\} \supseteq \{\phi_s : \phi_s \geq \phi_2\}$, and the second inequality follows because $\mathscr{R}_M(\phi,\phi_s)$ is increasing in ϕ for a fixed ϕ_s. Thus, $\min_{\phi_s \geq \phi} \mathscr{R}_M(\phi,\phi_s)$ is a monotonically increasing function in ϕ. □

D.6.4 Risk monotonization of ridge bagged predictors by cross-validation

Proof of Theorem 4.5.5. It suffices to verify the two conditions (i) and (ii) in Theorem 4.5.1. From Theorem 4.4.1 and Theorem 4.4.6, condition (i) holds naturally with $\mathscr{R}_M(\phi,\phi_s)$ being the limiting risk $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)$ (or $\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)$) for fixed $\lambda \geq 0$. For condition (ii), note that when $\lambda > 0$, $\mathscr{R}_M(\phi,\phi_s)$ is continuous over $[\phi,\infty]$. When $\lambda = 0$, $\mathscr{R}_M(\phi,\phi_s)$ is continuous over $[\phi,\infty] \setminus \{1\}$ and can takes value infinity when ϕ_s tends to 1 from both sides. Thus, $\mathscr{R}_M(\phi,\phi_s)$ is lower semi-continuous over $[\phi,\infty]$ and continuous on the set $\arg\min_{\psi:\psi\geq\phi} \mathscr{R}_M(\phi,\psi) \subseteq [\phi,\infty] \setminus \{1\}$.

Following Remark 4.5.4, the uniform risk closeness condition for $k \in \mathcal{K}_n$ holds. Then by Theorem 4.5.1, we have that

$$\left(R(\widehat{f}^{\text{cv}}_M;\mathcal{D}_n,\{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s \geq \phi} \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)\right)_+ \xrightarrow{\text{p}} 0.$$

Recall that for any fixed θ, the function

$$\widetilde{v}(-\lambda;\vartheta,\theta) = \frac{\vartheta \int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}{v(-\lambda;\phi_s)^{-2} - \vartheta \int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)} \geq 0$$

is increasing in ϑ. Then $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ as a function of $\widetilde{v}(-\lambda;\vartheta,\theta)$ through (4.23) and (4.26) is also increasing in ϕ for any fixed ϕ_s. For $0<\phi_1\leq\phi_2<\infty$,

$$\min_{\phi_s\geq\phi_1}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_1,\phi_s)\leq\min_{\phi_s\geq\phi_2}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_1,\phi_s)\leq\min_{\phi_s\geq\phi_2}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_2,\phi_s)$$

where the first inequality follows because $\{\phi_s:\phi_s\geq\phi_1\}\supseteq\{\phi_s:\phi_s\geq\phi_2\}$, and the second inequality follows because $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ is increasing in ϕ for a fixed ϕ_s. Thus, $\min_{\phi_s\geq\phi}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ is a monotonically increasing function in ϕ. □

D.6.5 Optimal subagging versus optimal splagging

Proof of Proposition 4.5.6. Recall that

$$\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)=\frac{1}{M}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s))+\left(1-\frac{1}{M}\right)(B_\lambda(\phi,\phi_s)+V_\lambda(\phi,\phi_s)),\qquad M\in\mathbb{N}$$

$$\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s)=\frac{1}{M}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s))+\left(1-\frac{1}{M}\right)C_\lambda(\phi_s),\qquad M=1,\dots,\lfloor\frac{n}{k}\rfloor.$$

From Proposition 4.4.5, we have that

$$\begin{aligned}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)&\geq\mathscr{R}^{\mathtt{sub}}_{\lambda,\infty}(\phi,\phi_s)\\&=B_\lambda(\phi,\phi_s)+V_\lambda(\phi,\phi_s)\\&=\rho^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s)+\sigma^2\widetilde{v}(-\lambda;\phi,\phi_s)\\&=\rho^2\widetilde{c}(-\lambda;\phi_s)+\widetilde{v}(-\lambda;\phi,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2).\end{aligned}\tag{D.53}$$

where $\widetilde{c}(-\lambda;\phi_s)=\int r/(1+v(-\lambda;\phi_s)r)^2\,\mathrm{d}G(r)$. From Proposition 4.4.10, we have that for $M\in\mathbb{N}$,

$$\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s)\geq\mathscr{R}^{\mathtt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)=\frac{\phi}{\phi_s}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s))+\left(1-\frac{\phi}{\phi_s}\right)C_\lambda(\phi_s).\tag{D.54}$$

On the other hand,

$$\begin{aligned}\mathscr{R}^{\mathtt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)&=\frac{\phi}{\phi_s}\rho^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s)+\frac{\phi}{\phi_s}\sigma^2\widetilde{v}(-\lambda;\phi_s,\phi_s))+\left(1-\frac{\phi}{\phi_s}\right)\rho^2\widetilde{c}(-\lambda;\phi_s)\\&=\rho^2\widetilde{c}(-\lambda;\phi_s)+\frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2).\end{aligned}\tag{D.55}$$

Since $v(-\lambda;\phi_s)$ is strictly decreasing in ϕ_s from Lemma D.8.13 and G has nonnegative support from Assumption 4.5, we have that $\widetilde{c}(-\lambda;\phi_s)$ is nonnegative and increasing in ϕ_s. Also note that

$$\begin{aligned}\frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s))&=\frac{\phi\displaystyle\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}{v(-\lambda;\phi_s)^{-2}-\phi_s\displaystyle\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}\\&\geq\frac{\phi\displaystyle\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}{v(-\lambda;\phi_s)^{-2}-\phi\displaystyle\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}\\&=\widetilde{v}(-\lambda;\phi,\phi_s).\end{aligned}\tag{D.56}$$

Suppose that $\phi^*\in\arg\min_{\inf_{\phi_s\in[\phi,\infty]}}\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s)$, we have

$$\inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)=\inf_{\phi_s\in[\phi,\infty]}\mathscr{R}^{\mathtt{sub}}_{\lambda,\infty}(\phi,\phi_s)$$

$$
\begin{aligned}
&\leq \mathscr{R}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi^*)\\
&= \rho^2\widetilde{c}(-\lambda;\phi_s^*) + \widetilde{v}(-\lambda;\phi_s^*,\phi))(\rho^2\widetilde{c}(-\lambda;\phi_s^*)+\sigma^2)\\
&\leq \rho^2\widetilde{c}(-\lambda;\phi_s) + \frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2)\\
&= \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi_s,\phi_s^*)\\
&= \inf_{\phi_s\in[\phi,\infty]} \mathscr{R}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)\\
&\leq \inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]} \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)
\end{aligned}
$$

where in the second inequality we use (D.56) and the last inequality is from (D.54). □

D.6.6 Optimal bag size for ridgeless predictors

Proof of Proposition 4.5.7. The proof of Proposition 4.5.7 follows by combining results from Lemma D.6.1 and Lemma D.6.2 for subagged and splagged ridgeless predictors, respectively. □

Lemma D.6.1 (Optimal risk for subagged ridgeless predictor)**.** *Suppose the conditions in Theorem 4.4.1 hold, and $\sigma^2,\rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 4.2 and 4.3. Let $\texttt{SNR} = \rho^2/\sigma^2$. For any $\phi\in(0,\infty)$, the properties of the optimal asymptotic risk $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s^{\texttt{sub}}(\phi))$ in terms of $\texttt{SNR}$ and ϕ are characterized as follows:*

(1) $\texttt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \neq 0)$*: For all $\phi \geq 0$, the global minimum $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s^{\texttt{sub}}(\phi)) = \sigma^2$ is obtained with $\phi_s^{\texttt{sub}}(\phi) = \infty$.*

(2) $\texttt{SNR} > 0$*: For all $\phi \geq 0$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)$ is obtained at $\phi_s^{\texttt{sub}}(\phi)\in(1,\infty)$.*

(3) $\texttt{SNR} = \infty$ $(\rho^2 \neq 0, \sigma^2 = 0)$*: If $\phi\in(0,1]$, the global minimum $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s^{\texttt{sub}}(\phi)) = 0$ is obtained with any $\phi_s^{\texttt{sub}}(\phi)\in[\phi,1]$. If $\phi\in(1,\infty)$, then the global minimum $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s^{\texttt{sub}}(\phi))$ is obtained at $\phi_s^{\texttt{sub}}(\phi)\in[\phi,\infty)$.*

Proof of Lemma D.6.1. From Theorem 4.4.1, the limiting risk for bagged ridgeless with $M=\infty$ is given by

$$\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s) + \sigma^2(1+\widetilde{v}(0;\phi,\phi_s)).$$

Defined in (D.33)-(D.34), $\widetilde{v}(0;\phi,\phi_s)\geq 0$ and $\widetilde{c}(0;\phi_s)\geq 0$ are continuous functions of $v(0;\phi_s)$, which is strictly decreasing over $\phi_s\in(1,\infty)$ and satisfies $\lim_{\phi_s\to\infty} v(0;\phi_s) = 0$ from Lemma D.8.14. Then we have $\widetilde{v}(0;\phi,\phi_s)$ is decreasing in ϕ_s over $(1,\infty)$, $\widetilde{c}(0;\phi_s)$ is increasing in ϕ_s over $(1,\infty)$, and

$$\lim_{\phi_s\to\infty}\widetilde{v}(0;\phi,\phi_s) = 0, \qquad \lim_{\phi_s\to\infty}\widetilde{c}(0;\phi_s) = \int r\,\mathrm{d}G(r).$$

Also, $\widetilde{v}(0;\phi,\phi_s) = \phi/(1-\phi)$ and $\widetilde{c}(0;\phi_s) = 0$ remain constant for $\phi_s\in(0,1]$ from (4.28). Then to determine the global minimum, it suffices to consider the case when $\phi_s\in[1,\infty)$. Next, we consider various cases depending on the value of $\texttt{SNR}$.

- We first consider the case $\texttt{SNR} > 0$. We consider further sub-cases depending the value of the pair (ϕ,ϕ_s).

  1. When $\phi\in(0,1)$ and $\phi_s\in(1,\infty]$,

$$
\begin{aligned}
\frac{\partial\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)}{\partial\phi_s} &= \frac{\partial\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)}{\partial v(0;\phi_s)}\frac{\partial v(0;\phi_s)}{\partial\phi_s}\\
&= \rho^2\frac{\phi\displaystyle\int\frac{v(0;\phi_s)r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)}{\left(1-\phi\displaystyle\int\left(\frac{v(0;\phi_s)r}{(1+v(0;\phi_s)r)}\right)^2\mathrm{d}H(r)\right)^2}\int\frac{r}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}G(r)\cdot\frac{\partial v(0;\phi_s)}{\partial\phi_s}
\end{aligned}
$$

$$
- 2\rho^2 \frac{\dfrac{1}{v(0;\phi_s)^2}}{\dfrac{1}{v(0;\phi_s)^2} - \phi \displaystyle\int \frac{r^2}{(1+v(0;\phi_s)r)^2} \, \mathrm{d}H(r)} \int \frac{r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}G(r) \cdot \frac{\partial v(0;\phi_s)}{\partial \phi_s}
$$

$$
+ \sigma^2 \frac{\phi \displaystyle\int \frac{v(0;\phi_s) r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}H(r)}{\left(1 - \phi \displaystyle\int \left(\frac{v(0;\phi_s) r}{(1+v(0;\phi_s)r)}\right)^2 \mathrm{d}H(r)\right)^2} \cdot \frac{\partial v(0;\phi_s)}{\partial \phi_s}.
$$

Note that from Lemma D.8.13 $v(0;\phi_s)$ is differentiable in $\phi_s \in (0,\infty]$ with

$$
\frac{\partial v(0;\phi_s)}{\partial \phi_s} = - \frac{\displaystyle\int \frac{r}{1+v(0;\phi_s)r} \, \mathrm{d}H(r)}{\dfrac{1}{v(0;\phi_s)^2} - \phi_s \displaystyle\int \frac{r^2}{(1+v(0;\phi_s)r)^2} \, \mathrm{d}H(r)}
$$

being negative over $\phi_s \in (1,\infty)$ and continuous in $\phi_s \in (1,\infty]$, and

$$
\lim_{\phi_s \to 1^+} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = -\infty, \qquad \lim_{\phi_s \to \infty} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = - \lim_{\phi_s \to \infty} \widetilde{v}_v(0;\phi_s) \int \frac{r}{1+v(0;\phi_s)r} \, \mathrm{d}H(r) = 0
$$

by Lemma D.8.14 with $\widetilde{v}_v$ defined therein. We have that $\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ is continuous over $\phi_s \in (1,\infty]$. Since $\lim_{\phi_s\to\infty} v(0;\phi_s) = 0$ from Lemma D.8.14, we have that

$$
\lim_{\phi_s \to \infty} \frac{\phi \displaystyle\int \frac{v(0;\phi_s) r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}H(r)}{\left(1 - \phi \displaystyle\int \left(\frac{v(0;\phi_s) r}{(1+v(0;\phi_s)r)}\right)^2 \mathrm{d}H(r)\right)^2} = 0 \tag{D.57}
$$

$$
\lim_{\phi_s \to \infty} \frac{\dfrac{1}{v(0;\phi_s)^2}}{\dfrac{1}{v(0;\phi_s)^2} - \phi \displaystyle\int \frac{r^2}{(1+v(0;\phi_s)r)^2} \, \mathrm{d}H(r)} \int \frac{r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}G(r) = \frac{1}{1-\phi} \int r^2 \, \mathrm{d}G(r) > 0. \tag{D.58}
$$

Since $\partial v(0;\phi_s)/\partial \phi_s$ is negative over $(1,\infty)$ and $\lim_{\phi_s\to\infty 0} \partial v(0;\phi_s)/\partial\phi_s = 0$, we have

$$
\frac{\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)}{\partial \phi_s} \Big|_{\phi_s=\infty} = -\rho^2 \int r^2 \, \mathrm{d}G(r) \cdot \lim_{\phi_s\to\infty} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = 0. \tag{D.59}
$$

Combining (D.57)-(D.59), we have that when ϕ_s is large, $\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ approaching zero from above as ϕ_s tends to ∞. On the other hand, since for $k=1,2$,

$$
\begin{aligned}
&\lim_{\phi_s\to 1^+} \int \frac{r^k}{(1+v(0;\phi_s)r)^{k+1}} \, \mathrm{d}G(r) \cdot \frac{\partial v(0;\phi_s)}{\partial \phi_s} \\
&= \lim_{\phi_s\to 1^+} \int \frac{v(0;\phi_s) r^k}{(1+v(0;\phi_s)r)^{k+1}} \, \mathrm{d}G(r) \cdot \lim_{\phi_s\to 1^+} \frac{1}{v(0;\phi_s)} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = 0,
\end{aligned}
$$

we have

$$
\frac{\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)}{\partial \phi_s} \Big|_{\phi_s=1^+} = \sigma^2 \frac{\phi}{1-\phi} \cdot \lim_{\phi_s\to 1^+} \frac{\partial v(0;\phi_s)}{\partial \phi_s} < 0.
$$

Thus, there exists $\phi^* \in (1,\infty)$ such that

$$
\mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi^*) < \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,1) = \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi).
$$

2. When $\phi = 1$, $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(1,1) = \infty$ while $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s) < \infty$ for all $\phi_s \in (1,\infty]$. Since $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s)$ is continuous and finite in $(1,\infty]$, by continuity and (D.57)-(D.59) we have $\phi^* \in (1,\infty)$.
3. When $\phi \in (1,\infty)$, the optimal $\phi^* \geq \phi > 1$ must be obtained in $[\phi,\infty)$ because of (D.57)-(D.59).

- Next consider the case when $\mathtt{SNR} = 0$, i.e., $\rho^2 = 0$ and $\sigma^2 \neq 0$, since $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s) = \sigma^2 + \sigma^2\widetilde{v}(0;\phi,\phi_s) > 0$ is increasing in $v(0;\phi_s)$ and $v(0;\phi_s) \geq 0$ is decreasing in ϕ_s, we have that $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s)$ is decreasing in ϕ_s. Thus, the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\infty,\phi_s) = \sigma^2$ is obtained at $\phi_s^* = \infty$.
- Finally, consider the case when $\mathtt{SNR} = \infty$, i.e. $\rho^2 \neq 0$ and $\sigma^2 = 0$, $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s) = \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s)$. As the bias term is zero when $\phi_s \in (0,1]$ and positive when $\phi_s \in (1,\infty]$, we have that $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s) \geq \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi^*) = 0$ for all $\phi_s^* \in [\phi,1]$ when $\phi \in (0,1]$. If $\phi \in (1,\infty)$, since the risk is continuous over $[\phi,\infty]$, the global minimum exists. Since the derivative $\partial\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ is continuous over $\phi_s \in (1,\infty]$ and (D.57)-(D.59), the minimizer satisfies $\phi^* \in [\phi,\infty)$.

□

Lemma D.6.2 (Optimal splagged ridgeless)**.** *Suppose the conditions in Theorem 4.4.6 hold, and $\sigma^2,\rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 4.2 and 4.3. Let $\mathtt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0,\infty)$, the properties of the optimal asymptotic risk $\mathscr{R}^{\mathtt{spl}}_{0,\infty}(\phi,\phi_s^{\mathtt{spl}}(\phi))$ in terms of $\mathtt{SNR}$ and ϕ are characterized as follows:*

(1) $\mathtt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \neq 0)$*: For all $\phi \geq 0$, the global minimum $\mathscr{R}^{\mathtt{spl}}_{0,\infty}(\phi,\phi_s^{\mathtt{spl}}(\phi)) = \sigma^2$ is obtained with $\phi_s^{\mathtt{spl}}(\phi) = \infty$.*

(2) $\mathtt{SNR} > 0$*: For $\phi \geq 1$, there exists global minimum of $\phi_s \mapsto \mathscr{R}^{\mathtt{spl}}_{0,\infty}(\phi,\phi_s)$ in $(1,\infty)$. For $\phi \in (0,1)$, the global minimum is in $\{\phi\} \cup (1,\infty)$.*

(3) $\mathtt{SNR} = \infty$ $(\rho^2 \neq 0, \sigma^2 = 0)$*: If $\phi \in (0,1]$, the global minimum $\mathscr{R}^{\mathtt{spl}}_{0,\infty}(\phi,\phi_s^{\mathtt{spl}}(\phi)) = 0$ is obtained with any $\phi_s^{\mathtt{spl}}(\phi) \in [\phi,1]$. If $\phi \in (1,\infty)$, then the global minimum $\mathscr{R}^{\mathtt{spl}}_{0,\infty}(\phi,\phi_s^{\mathtt{spl}}(\phi))$ is obtained at $\phi_s^{\mathtt{spl}}(\phi) \in [\phi,\infty)$.*

Proof of Lemma D.6.2. From Theorem 4.4.6, the limiting risk for bagged ridgeless with $M = \phi_s/\phi$ is given by

$$\begin{aligned}\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s) &= \sigma^2 + \frac{\phi}{\phi_s}\left[\rho^2(1+\widetilde{v}(0;\phi_s,\phi_s))\widetilde{c}(0;\phi_s) + \sigma^2\widetilde{v}(0;\phi_s,\phi_s)\right] + \left(1-\frac{\phi}{\phi_s}\right)\rho^2\widetilde{c}(0;\phi_s)\\ &= \sigma^2 + \rho^2\widetilde{c}(0;\phi_s) + \phi\frac{\widetilde{v}(0;\phi_s,\phi_s)}{\phi_s}(\rho^2\widetilde{c}(0;\phi_s)+\sigma^2).\end{aligned}$$

Defined in (D.33)-(D.34), $\widetilde{v}(0;\phi_s,\phi_s) \geq 0$ and $\widetilde{c}(0;\phi_s) \geq 0$ are continuous functions of $v(0;\phi_s)$, which is strictly decreasing over $\phi_s \in (1,\infty)$ and satisfies $\lim_{\phi_s\to\infty} v(0;\phi_s) = 0$ from Lemma D.8.14. Then $\widetilde{c}(0;\phi_s)$ is increasing in ϕ_s over $(1,\infty)$ and $\lim_{\phi_s\to\infty}\widetilde{c}(0;\phi_s) = \int r\,\mathrm{d}G(r)$.

- We first consider the case $\mathtt{SNR} > 0$. We consider further sub-cases depending the value of the pair (ϕ,ϕ_s).
  1. When $\phi \in (0,1)$ and $\phi_s \in (1,\infty]$,

     Define functions h_1 and h_2 as follows:

     $$h_1(\phi_s) = \mathtt{SNR}\cdot\widetilde{c}(0;\phi_s), \qquad h_2(\phi_s) = \frac{\widetilde{v}(0;\phi_s,\phi_s)}{\phi_s} = \widetilde{v}_v(0;\phi_s)\int\left(\frac{r}{1+v(0;\phi_s)r}\right)^2\mathrm{d}H(r), \quad \text{(D.60)}$$

     where $\widetilde{v}_v$ is defined in Lemma D.8.14. Then $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s) = \sigma^2 + \sigma^2(h_1(\phi_s) + \phi h_2(\phi_s)(1+h_1(\phi_s)))$, with h_1 increasing in ϕ_s and

     $$\lim_{\phi_s\to1^+} h_1(\phi_s) = 0, \qquad \lim_{\phi_s\to\infty} h_1(\phi_s) = \mathtt{SNR}\int r\,\mathrm{d}G(r), \qquad \lim_{\phi_s\to1^+} h_2(\phi_s) = +\infty, \qquad \lim_{\phi_s\to\infty} h_2(\phi_s) = 0.$$

Next we study the property of h_2. Simple calculation yields that

$$
\begin{aligned}
&\frac{\partial h_2(\phi_s)}{\partial \phi_s}\\
=&\widetilde{v}_v(0;\phi_s)^2\left[\frac{2}{v(0;\phi_s)^3}\int \frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)\cdot\frac{\partial v(0;\phi_s)}{\partial\phi_s}+\left(\int\frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2\right]\\
=&\widetilde{v}_v(0;\phi_s)^2\left[\left(\int\frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2-\frac{2\widetilde{v}_v(0;\phi_s)}{v(0;\phi_s)^3}\int\frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\right].
\end{aligned}
$$

From Lemma D.8.14 (4), we have that $\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s)/v(0;\phi_s)^2\lim_{\phi_s\to\infty}[1+\widetilde{v}_b(0;\phi_s)]=1$ where $\widetilde{v}_b(0;\phi_s)$ is defined in Lemma D.8.14. Analogously, $\lim_{\phi_s\to1^+}\widetilde{v}_v(0;\phi_s)/v(0;\phi_s)^2=+\infty$. Then as in the proof of Proposition 4.5.7, one can verify that

$$
\begin{aligned}
\frac{\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)}{\partial\phi_s}=&-\sigma^2\widetilde{v}_v(0;\phi_s)\Bigg[\texttt{SNR}(1+\phi h_2(\phi_s))\int\frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}G(r)\cdot\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\\
&+\phi(1+h_1(\phi_s))\frac{2\widetilde{v}_v(0;\phi_s)^2}{v(0;\phi_s)^3}\int\frac{r^2}{(1+v(0;\phi_s))^3}\,\mathrm{d}H(r)\cdot\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\\
&-\widetilde{v}_v(0;\phi_s)\phi(1+h_1(\phi_s))\left(\int\frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2\Bigg]
\end{aligned}
$$

satisfies $\lim_{\phi_s\to1^+}\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s=-\infty$ and $\lim_{\phi_s\to\infty}\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s=0$ by utilizing properties in Lemma D.8.14. Furthermore, as

$$
\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s)^{-1}\frac{\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)}{\partial\phi_s}=-\rho^2\int r^2\,\mathrm{d}G(r)\cdot\int r\,\mathrm{d}H(r)<0, \tag{D.61}
$$

we have that when ϕ_s is large, $\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s$ approaching zero from above as ϕ_s tends to ∞. Thus, the minimum of $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ over $[1,\infty]$ is obtained in the open interval $(1,\infty)$.

2. When $\phi<1$ and $\phi_s\in[\phi,1)$, since the term $\widetilde{c}(0;\phi_s)$ is zero, $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)=\sigma^2+\sigma^2\phi(1-\phi_s)^{-1}$ is increasing in ϕ_s. So the minimum over $[\phi,1]$ is obtained at $\phi_s=\phi$.
3. When $\phi=1$, $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(1,1)=\infty$ while $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)<\infty$ for all $\phi_s\in(1,\infty]$. Since $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ is continuous and finite in $(1,\infty]$, by continuity and (D.61) we have $\phi_s^*\in(1,\infty)$.
4. When $\phi\in(1,\infty)$, the optimal $\phi_s^*\geq\phi>1$ must be obtained in $[\phi,\infty)$ because of (D.61).

- Next consider the case when $\texttt{SNR}=0$, i.e., $\rho^2=0$ and $\sigma^2\neq0$. Then $h_1\equiv0$ and $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi_s/\phi,\phi_s)=\sigma^2+\sigma^2\phi\widetilde{v}(0;\phi_s,\phi_s)/\phi_s$. When $\phi_s\in(0,1)$, $\widetilde{v}(0;\phi_s,\phi_s)/\phi_s=(1-\phi_s)^{-1}$ is increasing in ϕ_s; when $\phi_s>1$, $\widetilde{v}(0;\phi_s,\phi_s)/\phi_s\geq0=\lim_{\phi_s\to\infty}\widetilde{v}(0;\phi_s,\phi_s)/\phi_s=0$. Therefore, the global minimum $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\infty,\phi_s)=\sigma^2$ is obtained at $\phi_s^*=\infty$.

- Finally, consider the case when $\texttt{SNR}=\infty$, i.e. $\rho^2\neq0$ and $\sigma^2=0$, $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi_s/\phi,\phi_s)=\rho^2\widetilde{c}(0;\phi_s)+\rho^2\phi\phi_s^{-1}\widetilde{v}(0;\phi,\phi_s)\widetilde{c}(0;\phi_s)$. As the term $\widetilde{c}(0;\phi_s)$ is zero when $\phi_s\in(0,1]$ and positive when $\phi_s\in(1,\infty]$, we have that $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)\geq\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s^*)=0$ for all $\phi_s^*\in[\phi,1]$ when $\phi\in(0,1]$. If $\phi\in(1,\infty)$, since the risk is continuous over $[\phi,\infty]$, the global minimum exists. Since the derivative $\partial\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ is continuous over $\phi_s\in(1,\infty]$ and (D.61), the minimizer satisfies $\phi_s^*\in[\phi,\infty)$.

□

D.7 Proofs in Section 4.6

D.7.1 Bagged risk for ridgeless regression

Proof of Corollary 4.6.1. Since $\mathbf{\Sigma} = \boldsymbol{I}_p$, we have that $\mathrm{d}G = \mathrm{d}H = \delta_1$. Then, $v(0;\phi_s)$, $\widetilde{v}(0;\phi,\phi_s)$ and $\widetilde{c}(0;\phi_s)$ defined in (D.33) and (D.34) for $\phi_s > 1$ reduce to

$$v(0;\phi_s) = \frac{1}{\phi_s - 1}, \qquad \widetilde{v}(0;\phi,\phi_s) = \frac{\phi}{\phi_s^2 - \phi}, \qquad \widetilde{c}(0;\phi_s) = \frac{(\phi_s - 1)^2}{\phi_s^2}.$$

Thus, we have

$$B_0(\phi,\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2 \dfrac{\phi_s - 1}{\phi_s}, & \phi_s \in (1,\infty) \end{cases}, \qquad V_0(\phi,\phi_s) = \begin{cases} \sigma^2 \dfrac{\phi_s}{1-\phi_s}, & \phi_s \in (0,1) \\ \sigma^2 \dfrac{1}{\phi_s - 1}, & \phi_s \in (1,\infty) \end{cases},$$

and

$$C_0(\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2 \dfrac{(\phi_s - 1)^2}{\phi_s^2}, & \phi_s \in (1,\infty) \end{cases}.$$

□

From Corollary 4.6.1, we are able to derive the asymptotic bias and variance for $M = 1$ and $M = \infty$ for ridgeless regression with replacement:

$$\mathscr{B}^{\texttt{sub}}_{0,1}(\phi,\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2 \dfrac{\phi_s - 1}{\phi_s}, & \phi_s \in (1,\infty) \end{cases} \qquad \mathscr{V}^{\texttt{sub}}_{0,1}(\phi,\phi_s) = \begin{cases} \sigma^2 \dfrac{\phi_s}{1-\phi_s}, & \phi_s \in (0,1) \\ \sigma^2 \dfrac{1}{\phi_s - 1}, & \phi_s \in (1,\infty) \end{cases}$$

$$\mathscr{B}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2 \dfrac{(\phi_s - 1)^2}{\phi_s^2 - \phi}, & \phi_s \in (1,\infty) \end{cases} \qquad \mathscr{V}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \begin{cases} \sigma^2 \dfrac{\phi}{1-\phi}, & \phi_s \in (0,1) \\ \sigma^2 \dfrac{\phi}{\phi_s^2 - \phi}, & \phi_s \in (1,\infty) \end{cases}.$$

Then the asymptotic bias and variance for general M would be convex combinations of the above quantities.

On the other hand, the asymptotic bias and variance for splagging without replacement are given by

$$\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi_s,\phi_s) + (1-M^{-1})C_\lambda(\phi_s), \qquad \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s).$$

D.7.2 Optimal subagged ridgeless regression with replacement

Proof of Proposition 4.6.2. For $\phi \in (0,1)$ and $\phi_s \in (1,\infty]$, we have that

$$\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \sigma^2 + \rho^2 \frac{(\phi_s - 1)^2}{\phi_s^2 - \phi} + \sigma^2 \frac{\phi}{\phi_s^2 - \phi}.$$

Taking the derivative of the right hand side with respect to ϕ_s

$$\frac{\partial \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)}{\partial \phi_s} = 2\sigma^2 \frac{\texttt{SNR}(\phi_s - 1)(\phi_s - \phi) - \phi\phi_s}{(\phi_s^2 - \phi)^2}$$

and setting it to zero yields that

$$\phi_s = A \pm \sqrt{A^2 - \phi}. \tag{D.62}$$

where $A = (\phi + 1 + \phi/\texttt{SNR})/2$. Since $A - \sqrt{A^2 - \phi} < \sqrt{\phi} \leq 1$, we have $\phi_s^* = A + \sqrt{A^2 - \phi}$ is a minimizer and

$$
\begin{aligned}
\mathscr{R}_{0,\infty}^{\texttt{sub}}(\phi, \phi^*) &= \sigma^2 + \sigma^2 \frac{\phi + A - \sqrt{A^2 - \phi} + \texttt{SNR}(1-\phi)/\phi(A - \phi - \sqrt{A^2 - \phi})}{2\sqrt{A^2 - \phi}} \\
&= \frac{\sigma^2}{2}\left[1 + \frac{\phi - 1}{\phi}\texttt{SNR} + \frac{2\texttt{SNR}}{\phi}\sqrt{A^2 - \phi}\right] \\
&= \frac{\sigma^2}{2}\left[1 + \frac{\phi - 1}{\phi}\texttt{SNR} + \sqrt{\left(1 - \frac{\phi - 1}{\phi}\texttt{SNR}\right)^2 + 4\texttt{SNR}}\right], \qquad \text{(D.63)}
\end{aligned}
$$

which gives the simplified formula. Note that

$$
\begin{aligned}
\mathscr{R}_{0,\infty}^{\texttt{sub}}(\phi, \phi^*) &= \sigma^2 + \sigma^2\left(\frac{\phi}{2\sqrt{A^2 - \phi}} + \frac{A - \sqrt{A^2 - \phi}}{2\sqrt{A^2 - \phi}} + \frac{1 - \phi}{\phi}\texttt{SNR}\frac{A - \phi - \sqrt{A^2 - \phi}}{2\sqrt{A^2 - \phi}}\right) \\
&= \sigma^2 + \sigma^2 h(\texttt{SNR}) - \sigma^2\delta(\texttt{SNR}) \qquad \text{(D.64)}
\end{aligned}
$$

where for all $r \geq 0$, the functions h and δ are defined as $h(r) = h_1(r) + h_2(r) + h_3(r)$ and $\delta(r) = (1 - \phi)rh_1(r)/\phi$, with $A(r) = (\phi + 1 + \phi/r)/2$ and

$$
\begin{aligned}
h_1(r) &= \frac{\phi}{2\sqrt{A(r)^2 - \phi}} \\
h_2(r) &= \frac{A(r) - \sqrt{A(r)^2 - \phi}}{2\sqrt{A(r)^2 - \phi}} = \frac{1}{2\sqrt{1 - \phi/A(r)^2}} - \frac{1}{2} \\
h_3(r) &= \frac{1 - \phi}{\phi}rh_2(r).
\end{aligned}
$$

Since h_1, h_2, and h_3 are nonngative over $(0, \infty)$, h and δ are also nonnegative. Also noted that

$$
\delta(0) = \frac{1 - \phi}{\phi}\lim_{r \to 0^+} rh_1(r) = 0, \qquad \delta(\infty) = \frac{1 - \phi}{\phi}\lim_{r \to +\infty} rh_1(r) = +\infty,
$$

we obtain the upper bound for $\mathscr{R}_{0,\infty}^{\texttt{sub}}(\phi, \phi^*)$ as follows:

$$
\mathscr{R}_{0,\infty}^{\texttt{sub}}(\phi, \phi^*) \leq \sigma^2 + \sigma^2 h(\texttt{SNR}), \qquad \text{(D.65)}
$$

with equality obtained if and only if $\texttt{SNR} = 0$.

Next we analyze the function $h(r)$. Note that $A(r) > 0$ is decreasing in r, we have that the functions h_1 and h_2 are nonnegative and monotone increasing in $\texttt{SNR}$. Hence h_3 as the product of nonnegative and monotone increasing functions, is also nonnegative and monotone increasing in $\texttt{SNR}$. Thus, h is monotone increasing in $\texttt{SNR}$ and

$$
\begin{aligned}
h(\texttt{SNR}) &\leq \lim_{r \to \infty} h(r) \\
&= \lim_{r \to \infty} h_1(r) + \lim_{r \to \infty} h_2(r) + \lim_{r \to \infty} rh_2(r) \\
&= \frac{\phi}{1 - \phi} + \frac{\phi}{1 - \phi} + \frac{1}{\phi}\lim_{r \to \infty}\frac{A(r) - \sqrt{A(r)^2 - \phi}}{\frac{1}{r}} \\
&= \frac{\phi}{1 - \phi} + \frac{\phi}{1 - \phi} + \frac{1}{\phi}\lim_{r \to \infty}\frac{-\dfrac{\phi}{2r^2} + \dfrac{\dfrac{A(r)\phi}{r^2}}{2\sqrt{A(r)^2 - \phi}}}{-\dfrac{1}{r^2}}
\end{aligned}
$$

$$= \frac{\phi}{1-\phi}$$

where the third equality is due to the L'Hospital's rule. Note that the risk for $\phi_s \in [\phi, 1)$ is given by $\sigma^2 + \sigma^2\phi/(1-\phi)$, we have that ϕ_s^* obtained the global minimum of $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)$ over $\phi_s \in (\phi, \infty]$.

For $\phi \in [1, \infty)$ and $\phi_s \in [\phi, \infty)$, from (D.62) and $A - \sqrt{A^2 - \phi} \leq \sqrt{\phi} \leq \phi$, we have again $\phi_s^* = A + \sqrt{A^2 - \phi}$ is a minimizer.

When $\mathtt{SNR} = 0$, since the bias term is zero and variance term is increasing over $\phi_s < 1$ and increasing over $\phi_s > 1$, we have that when $\phi_s > 1$ (whenever $\phi \leq \phi_s$),

$$\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) = \sigma^2 + V_0(\phi, \phi_s) \geq \sigma^2 + V_0(\phi, \infty) = \sigma^2.$$

When $\phi < 1$, we have $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) \geq \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi) = \sigma^2/(1-\phi) > \sigma^2$. Therefore, $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) \geq \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \infty) = \sigma^2$ for all $\phi \in (1, \infty]$.

When $\mathtt{SNR} = \infty$, the variance term $V_0(\phi, \phi_s) = 0$ for all $\phi_s \in [\phi, \infty]$. If $\phi \in (0, 1]$, then $B_0(\phi, \phi_s) = 0$ for all $\phi_s \in [\phi, 1]$. If $\phi \in (1, \infty)$, then $B_0(\phi, \phi_s)$ is increasing over $\phi_s \in [\phi, \infty]$. Hence, the conclusions follow. □

D.7.3 Comparison between subagged and optimal ridge regression

Proof of Theorem 4.6.3. As $n, p \to \infty$ and $p/n \to \phi$, the optimal regularization parameter is given by $\lambda^* = \phi\sigma^2/\rho^2$ under the isotopic model (Dobriban and Wager, 2018). The limiting risk of the optimal ridge regression is given by

$$\mathscr{R}^{\mathtt{WR}}_{\lambda^*,1}(\phi, \phi) = \frac{\sigma^2}{2}\left[1 + \frac{\phi - 1}{\phi}\mathtt{SNR} + \sqrt{\left(1 - \frac{\phi-1}{\phi}\mathtt{SNR}\right)^2 + 4\mathtt{SNR}}\right]$$

which is the same the formula given in Proposition 4.6.2. Thus, the conclusion follows. □

Fixed-point equation details for ridge regression

For isotopic features $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\mathrm{d}G = \mathrm{d}H = \delta_1$. When $n, p \to$ and $p/n \to \phi \in (0, \infty)$, (D.73)-(D.75) reduce to

$$v(-\lambda; \phi)^{-1} = \lambda + \phi(1 + v(-\lambda; \phi))^{-1}$$
$$\widetilde{v}_b(-\lambda; \phi) = \frac{\phi(1 + v(-\lambda; \phi))^{-2}}{v(-\lambda; \phi)^{-2} - \phi(1 + v(-\lambda; \phi))^{-2}}$$
$$\widetilde{v}_v(-\lambda; \phi)^{-1} = v(-\lambda; \phi)^{-2} - \phi(1 + v(-\lambda; \phi))^{-2}.$$

Solving the first equation for $v(-\lambda; \phi) \geq 0$ gives

$$v(-\lambda; \phi) = \frac{1}{2\lambda}(-(\phi + \lambda - 1) + \sqrt{(\phi + \lambda - 1)^2 + 4\lambda}). \tag{D.66}$$

Then the asymptotic bias and variance defined in Theorem D.3.1 can be evaluated accordingly.

D.8 Auxiliary asymptotic equivalency results

D.8.1 Preliminaries

We use the notion of asymptotic equivalence of sequences of random matrices in various proofs. In this section, we provide a basic review of the related definitions and corresponding calculus rules.

Definition D.8.1 (Asymptotic equivalence: deterministic version)**.** Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$ and $\{\boldsymbol{B}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are equivalent and write $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$ if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely for any sequence of matrices $\boldsymbol{C}_p$ with bounded trace norm such that $\limsup_{p\to\infty} \|\boldsymbol{C}_p\|_{\mathrm{tr}} < \infty$.

We emphasize that recent work (Dobriban and Sheng, 2021; Patil et al., 2022a) used the deterministic version of the asymptotic equivalence, implicitly assuming that $\boldsymbol{C}_p$ in the definition is deterministic. However, in this work we need to investigate the asymptotic equivalence relationship conditional on some other sequences. In that direction, we first extend Definition D.8.1 to allow for random $\boldsymbol{C}_p$, as in Definition D.8.2.

Definition D.8.2 (Asymptotic equivalence: random version)**.** Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$ and $\{\boldsymbol{B}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are equivalent and write $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely for any sequence of random matrices $\boldsymbol{C}_p$ independent to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$, with bounded trace norm such that $\limsup_{p\to\infty} \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ almost surely.

Even though Definition D.8.1 seems to be more restrictive than Definition D.8.2, they are indeed equivalent as shown in Proposition D.8.3. The latter definition allows for more general definition for "conditional" asymptotic equivalents.

Proposition D.8.3 (Equivalence of $\simeq_D$ and $\simeq_R$)**.** *The asymptotic equivalent relations $\simeq_D$ in Definition D.8.1 and $\simeq_R$ in Definition D.8.2 are equivalent.*

Proof of Proposition D.8.3. Let $\{\boldsymbol{A}_p\}$ and $\{\boldsymbol{B}\}_p$ be two sequences of random matrices. Suppose that $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$. We next show that $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ holds. For any sequence of random matrices $\boldsymbol{C}_p$ that is independent to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ for all $p \in \mathbb{N}$, and has bounded trace norm such that $\limsup \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ as $p \to \infty$ almost surely. Let A denote the event that $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$. Then

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A] \overset{(a)}{=} \mathbb{E}[\mathbb{E}[\mathbb{1}_A \mid \{\boldsymbol{C}_p\}_{p\geq 1}]] \overset{(b)}{=} \mathbb{E}[1] = 1.$$

Above, equality (a) follows from the law of total expectation. Inequality (b) holds almost surely because $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$ and $\boldsymbol{C}_p$ is independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$. This can be seen as follows. Note that $\mathbb{1}_A(\{\boldsymbol{C}_p\}, (\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\}))$ is a function of random variables $\{\boldsymbol{C}_p\}$ and $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$. Let $\mathbb{E}[\mathbb{1}_A(\{\boldsymbol{c}_p\}, (\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\}))] = h(\{\boldsymbol{c}_p\})$ where the expectation is taken over the randomness in $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$. Since $\{\boldsymbol{C}_p\}$ and $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$ are independent and $\mathbb{E}[|\, \mathbb{1}_A |] \leq 1 < \infty$, we have that (see, e.g., Shiryaev (2016, Chapter 2, Section 7, Equation (16)), or Durrett (2019, Example 5.1.5))

$$\mathbb{E}[\mathbb{1}_A \mid \{\boldsymbol{C}_p\}] = h(\{\boldsymbol{C}_p\}),$$

and from Definition D.8.1, we have $h(\{\boldsymbol{C}_p\}) = 1$ almost surely. Thus, we can conclude that $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$.

On the other hand, by definition, $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ directly implies $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$, which completes the proof. □

The properties for the two types of deterministic equivalents are summarized in Lemma D.8.4. Though most of the calculus rules are the direct consequences from Dobriban and Wager (2018); Dobriban and Sheng (2021), the product rule involving random matrices $\boldsymbol{C}_p$ does not immediately follow from previous work.

Lemma D.8.4 (Calculus of deterministic equivalents)**.** *Let $\boldsymbol{A}_p$, $\boldsymbol{B}_p$, $\boldsymbol{C}_p$ and $\boldsymbol{D}_p$ be sequences of random matrices. The calculus of deterministic equivalents ($\simeq_D$ and $\simeq_R$) satisfies the following properties:*

(1) Equivalence: The relation $\simeq$ is an equivalence relation.

(2) Sum: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$ and $\boldsymbol{C}_p \simeq \boldsymbol{D}_p$, then $\boldsymbol{A}_p + \boldsymbol{C}_p \simeq \boldsymbol{B}_p + \boldsymbol{D}_p$.

(3) Product: If $\boldsymbol{A}_p$ has uniformly bounded operator norms such that $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\operatorname{op}} < \infty$, $\boldsymbol{A}_p$ is independent to $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$ for $p \geq 1$, and $\boldsymbol{B}_p \simeq \boldsymbol{C}_p$, then $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p$.

(4) Trace: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$, then $\operatorname{tr}[\boldsymbol{A}_p]/p - \operatorname{tr}[\boldsymbol{B}_p]/p \to 0$ almost surely.

(5) Differentiation: Suppose $f(z, \boldsymbol{A}_p) \simeq g(z, \boldsymbol{B}_p)$ where the entries of f and g are analytic functions in $z \in S$ and S is an open connected subset of $\mathbb{C}$. Suppose for any sequence $\boldsymbol{C}_p$ of deterministic matrices with bounded trace norm we have $|\operatorname{tr}[\boldsymbol{C}_p(f(z, \boldsymbol{A}_p) - g(z, \boldsymbol{B}_p))]| \leq M$ for every p and $z \in S$. Then we have $f'(z, \boldsymbol{A}_p) \simeq g'(z, \boldsymbol{B}_p)$ for every $z \in S$, where the derivatives are taken entrywise with respect to z.

Proof. The conclusions for $\simeq_D$ directly follow from Dobriban and Wager (2018); Dobriban and Sheng (2021). Then, the proof of property (1), (2), (4), and (5) for $\simeq_R$ follows from Proposition D.8.3. It remains to show that the product rule holds for $\simeq_R$. Since $\boldsymbol{B}_p \simeq_R \boldsymbol{C}_p$, we have $\boldsymbol{B}_p \simeq_D \boldsymbol{C}_p$. Then for any sequence of random matrices $\{\boldsymbol{D}_p\}_{p\geq 1}$ that have bounded trace norm and are independent to $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$, we have

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0\right) = 1.$$

Because $|\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| \leq \|\boldsymbol{A}_p\|_{\mathrm{op}} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]|$ and $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\mathrm{op}} < \infty$, we have that $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0$ implies $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| = 0$ conditioning on $\{\boldsymbol{A}_p\}_{p\geq 1}$. Thus,

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| = 0 \,\middle|\, \{\boldsymbol{A}_p\}_{p\geq 1}\right) = 1.$$

and by law of total expectation

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p\boldsymbol{A}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0\right) = 1,$$

which holds for any sequence of random matrices $\{\boldsymbol{D}_p\boldsymbol{A}_p\}_{p\geq 1}$ that have bounded trace norm and are independent to $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$. By definition, we have $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p$. □

Since the asymptotic equivalent relation $\simeq_D$ is equivalent to $\simeq_R$, we will just ignore the subscript and use the notation "$\simeq$" for simplicity. The subscript will be specified when needed.

D.8.2 Conditioning and calculus

In this section, we extend the notion of asymptotic equivalence of two sequences of random matrices from Definitions D.8.1 and D.8.2 to incorporate conditioning on another sequence of random matrices.

Definition D.8.5 (Conditional asymptotic equivalence)**.** Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$, $\{\boldsymbol{B}_p\}_{p\geq 1}$ and $\{\boldsymbol{D}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are equivalent given $\boldsymbol{D}_p$ and write $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{D}_p$ if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely conditional on $\{\boldsymbol{D}_p\}_{p\geq 1}$, i.e.,

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0 \,\middle|\, \{\boldsymbol{D}_p\}_{p\geq 1}\right) = 1,$$

for any sequence of random matrices $\boldsymbol{C}_p$ independent to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ conditional on $\boldsymbol{D}_p$, with bounded trace norm such that $\limsup \|\boldsymbol{C}_p\|_{\mathrm{tr}} < \infty$ as $p \to \infty$.

Below we formalize additional calculus rules that hold for conditional asymptotic equivalence Definition D.8.5.

Proposition D.8.6 (Calculus of conditional asymptotic equivalents)**.** *Let $\boldsymbol{A}_p$, $\boldsymbol{B}_p$, $\boldsymbol{C}_p$, and $\boldsymbol{E}_p$ be sequences of random matrices.*

(1) Unconditioning: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$, then $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$.

(2) Product: If $\boldsymbol{A}_p$ has bounded operator norms such that $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\mathrm{op}} < \infty$, $\boldsymbol{A}_p$ is conditional independent to $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$ given $\boldsymbol{E}_p$ for $p \geq 1$, and $\boldsymbol{B}_p \simeq \boldsymbol{C}_p \mid \boldsymbol{E}_p$, then $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p \mid \boldsymbol{E}_p$.

Proof of Proposition D.8.6. Proofs for the two parts appear below.

Part (1) For any sequence of deterministic matrices $\boldsymbol{C}_p$ with bounded trace norm, we have

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0 \,\middle|\, \{\boldsymbol{D}_p\}_{p\geq 1}\right) = 1$$

because $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$. By the law of total expectation, we have

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0\right) = 1.$$

Thus, $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$. By Proposition D.8.3, we further have $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$.

Part (2) For any sequence of random matrices $\boldsymbol{D}_p$, let E_1 and E_2 denote the event $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0$ and $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| = 0$, respectively. Because $\boldsymbol{B}_p \simeq \boldsymbol{C}_p \mid \boldsymbol{E}_p$, by definition we have

$$\mathbb{P}\left(E_1 \,|\, \{\boldsymbol{E}_p\}_{p\geq 1}\right) = 1$$

Because $|\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| \leq \|\boldsymbol{A}_p\|_{\mathrm{op}} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]|$ and $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\mathrm{op}} < \infty$, we have E_1 implies E_2 conditioning on $\{\boldsymbol{E}_p\}_{p\geq 1}$. Thus we have

$$\mathbb{P}\left(E_2 \,|\, \{\boldsymbol{E}_p\}_{p\geq 1}\right) = 1$$

holds for any $\{\boldsymbol{D}_p\}_{p\geq 1}$. This implies that $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p \mid \boldsymbol{E}_p$. □

Other rules in Lemma D.8.4 also hold for conditional asymptotic equivalents. A direct implication of this is that the deterministic equivalents for resolvents we will derive in Appendix D.8.3 based on these rules can be naturally generalized to allow for conditional asymptotic equivalents given a common sequence of random matrices that are independent to the source sequence.

D.8.3 Standard ridge resolvents and extensions

In this section, we collect various asymptotic equivalents that are used in the proofs of Lemmas D.3.4 and D.3.5, and Lemmas D.4.4 to D.4.6, which serve to prove Theorem 4.4.1. These equivalents are also subsequently used in the proof of Theorem 4.4.6.

D.8.3.1 Standard ridge resolvents

The following lemma provides a deterministic equivalent for the standard ridge resolvent and implies Corollary D.8.8. It is adapted from Theorem 1 of Rubio and Mestre (2011). See also Theorem 3 of Dobriban and Sheng (2021).

Lemma D.8.7 (Deterministic equivalent for standard ridge resolvent)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are i.i.d. random vectors such that each $\boldsymbol{x}_i = \boldsymbol{z}_i\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{z}_i$ is a random vector consisting of i.i.d. entries z_{ij}, $1 \leq j \leq p$, satisfying $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leq M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is a positive semidefinite matrix satisfying $0 \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$ for some constant $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ the concatenated matrix with $\boldsymbol{x}_i^\top$, $1 \leq i \leq n$, as rows, and let $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$ denote the random matrix $\boldsymbol{X}^\top\boldsymbol{X}/n$. Let $\gamma := p/n$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \gamma \leq \limsup \gamma < \infty$, we have*

$$(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-1} \simeq (c(e(z;\gamma))\boldsymbol{\Sigma} - z\boldsymbol{I}_p)^{-1}, \tag{D.67}$$

where the scalar $c(e(z;\gamma))$ is defined in terms of another scalar $e(z;\gamma)$ by the equation

$$c(e(z;\gamma)) = \frac{1}{1 + \gamma e(z;\gamma)}, \tag{D.68}$$

and $e(z;\gamma)$ is the unique solution in $\mathbb{C}^+$ to the fixed-point equation

$$e(z;\gamma) = \operatorname{tr}[\boldsymbol{\Sigma}(c(e(z;\gamma))\boldsymbol{\Sigma} - zI_p)^{-1}]/p. \tag{D.69}$$

Note that both the scalars $c(e(z;\gamma))$ and $e(z;\gamma)$ also implicitly depend on $\mathbf{\Sigma}$. For notation brevity, we do not always explicitly indicate this dependence. However, we will be explicit in such dependence for certain extensions to follow. See the remark after Lemma D.8.9 for more details. Additionally, observe that one can eliminate $e(z;\gamma)$ in the statement of Lemma D.8.7 by combining (D.68) and (D.69) so that for $z \in \mathbb{C}^+$, one has

$$(\widehat{\mathbf{\Sigma}} - z\boldsymbol{I}_p)^{-1} \simeq (c(z;\gamma)\mathbf{\Sigma} - z\boldsymbol{I}_p)^{-1},$$

where $c(z)$ is the unique solution in $\mathbb{C}^-$ to the fixed-point equation

$$\frac{1}{c(z;\gamma)} = 1 + \gamma \operatorname{tr}[\mathbf{\Sigma}(c(z;\gamma)\mathbf{\Sigma} - z\boldsymbol{I}_p)^{-1}]/p.$$

The following corollary is a simple consequence of Lemma D.8.7, which supplies a deterministic equivalent for the (regularization) scaled ridge resolvent. We will work with a real regularization parameter λ from here on.

Corollary D.8.8 (Deterministic equivalent for scaled ridge resolvent)**.** *Assume the setting of Lemma D.8.7. For $\lambda > 0$, we have*

$$\lambda(\widehat{\mathbf{\Sigma}} + \lambda\boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-1},$$

where $v(-\lambda;\gamma) > 0$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma)} = \lambda + \gamma \int \frac{r}{1 + v(-\lambda;\gamma)r} \, \mathrm{d}H_n(r). \tag{D.70}$$

Here H_n is the empirical distribution (supported on $\mathbb{R}_{\geq 0}$) of the eigenvalues of $\mathbf{\Sigma}$.

As a side note, the parameter $v(-\lambda;\gamma)$ in Corollary D.8.8 is also the companion Stieltjes transform of the spectral distribution of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$, which is also the Stieltjes transform of the spectral distribution of the gram matrix $\boldsymbol{X}\boldsymbol{X}^\top/n$.

The following lemma uses Corollary D.8.8 along with calculus of deterministic equivalents (from Lemma D.8.4), and provides deterministic equivalents for resolvents needed to obtain limiting bias and variance of standard ridge regression. It is adapted from Lemma S.6.10 of Patil et al. (2022a).

Lemma D.8.9 (Deterministic equivalents for ridge resolvents associated with generalized bias and variance)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are i.i.d. random vectors with each $\boldsymbol{x}_i = \boldsymbol{z}_i\mathbf{\Sigma}^{1/2}$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector that contains i.i.d. random variables z_{ij}, $1 \leq j \leq p$, each with $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leq M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ is a positive semidefinite matrix with $r_{\min}\boldsymbol{I}_p \preceq \mathbf{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$ for some constants $r_{\min} > 0$ and $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the concatenated random matrix with $\boldsymbol{x}_i$, $1 \leq i \leq n$, as its rows, and define $\widehat{\mathbf{\Sigma}} := \boldsymbol{X}^\top\boldsymbol{X}/n \in \mathbb{R}^{p\times p}$. Let $\gamma := p/n$. Then, for $\lambda > 0$, as $n, p \to \infty$ with $0 < \liminf \gamma \leq \limsup \gamma < \infty$, the following statements hold:*

(1) Bias of ridge regression:

$$\lambda^2(\widehat{\mathbf{\Sigma}} + \lambda\boldsymbol{I}_p)^{-1}\mathbf{\Sigma}(\widehat{\mathbf{\Sigma}} + \lambda\boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-1}(\widetilde{v}_b(-\lambda;\gamma)\mathbf{\Sigma} + \mathbf{\Sigma})(v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-1}. \tag{D.71}$$

(2) Variance of ridge regression:

$$(\widehat{\mathbf{\Sigma}} + \lambda\boldsymbol{I}_p)^{-2}\widehat{\mathbf{\Sigma}}\mathbf{\Sigma} \simeq \widetilde{v}_v(-\lambda;\gamma)(v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-2}\mathbf{\Sigma}\mathbf{\Sigma}. \tag{D.72}$$

Here $v(-\lambda;\gamma,\mathbf{\Sigma}) > 0$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma,\mathbf{\Sigma})} = \lambda + \int \frac{\gamma r}{1 + v(-\lambda;\gamma,\mathbf{\Sigma})r} \, \mathrm{d}H_n(r;\mathbf{\Sigma}), \tag{D.73}$$

and $\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma})$ and $\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma})$ are defined through $v(-\lambda;\gamma,\mathbf{\Sigma})$ by the following equations:

$$\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma}) = \frac{\displaystyle\int \gamma r^2 (1+v(-\lambda;\gamma,\mathbf{\Sigma})r)^{-2}\,\mathrm{d}H_n(r;\mathbf{\Sigma})}{v(-\lambda;\gamma,\mathbf{\Sigma})^{-2} - \displaystyle\int \gamma r^2 (1+v(-\lambda;\gamma,\mathbf{\Sigma})r)^{-2}\,\mathrm{d}H_n(r;\mathbf{\Sigma})}, \tag{D.74}$$

$$\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma})^{-1} = v(-\lambda;\gamma,\mathbf{\Sigma})^{-2} - \int \gamma r^2 (1+v(-\lambda;\gamma,\mathbf{\Sigma})r)^{-2}\,\mathrm{d}H_n(r;\mathbf{\Sigma}), \tag{D.75}$$

where $H_n(\cdot;\mathbf{\Sigma})$ is the empirical distribution (supported on $[r_{\min}, r_{\max}]$) of the eigenvalues of $\mathbf{\Sigma}$.

A few remarks on Lemma D.8.9 follow.

- The dependency of various scalar parameters appearing in Lemma D.8.9 on the matrix $\mathbf{\Sigma}$ is explicitly highlighted the statement. This is because when we extend the current results later in Lemma D.8.10, these parameters depend on the distributions of eigenvalues of matrices other than $\mathbf{\Sigma}$. In places where it is clear from context, we will write $H_n(r)$, $v(-\lambda;\gamma)$, $\widetilde{v}_b(-\lambda;\gamma)$, and $\widetilde{v}_v(-\lambda;\gamma)$ to denote $H_n(r;\mathbf{\Sigma})$, $v(-\lambda;\gamma,\mathbf{\Sigma})$, $\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma})$, and $\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma})$, respectively, for notational simplicity.
- Lemmas D.8.7 and D.8.9 assume existence of moments of order $8+\alpha$ for some $\alpha>0$ on the entries of $\boldsymbol{z}_i$, $1\le i\le k_m$, mentioned in assumption 4.2. As done in the proof of Theorem 6 of Hastie et al. (2022) (in Appendix A.4 therein), this can be relaxed to only requiring existence of moments of order $4+\alpha$ by a truncation argument. We omit the details and refer the readers to Hastie et al. (2022).

D.8.3.2 Extended ridge resolvents

The lemma below extends the deterministic equivalents of the ridge resolvents in Lemma D.8.9 to provide deterministic equivalents for Tikhonov resolvents, where the regularization matrix $\lambda\boldsymbol{I}_p$ is replaced with $\lambda(\boldsymbol{I}_p+\boldsymbol{C})$ and $\boldsymbol{C}\in\mathbb{R}^{p\times p}$ is an arbitrary positive semidefinite random matrix.

Lemma D.8.10 (Tikhonov resolvents)**.** *Suppose the conditions in Lemma D.8.9 holds. Let $\boldsymbol{C}\in\mathbb{R}^{p\times p}$ be any symmetric and positive semidefinite random matrix with uniformly bounded operator norm in p that is independent to $\boldsymbol{X}$ for all $n,p\in\mathbb{N}$, and let $\boldsymbol{N}=(\widehat{\mathbf{\Sigma}}+\lambda\boldsymbol{I}_p)^{-1}$. Then the following statements hold:*

(1) Tikhonov resolvent:

$$\lambda(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1} \simeq \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}. \tag{D.76}$$

(2) Bias of Tikhonov regression:

$$\lambda^2(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\mathbf{\Sigma}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1} \simeq \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}(\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\mathbf{\Sigma}+\mathbf{\Sigma})\widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}. \tag{D.77}$$

(3) Variance of Tikhonov regression:

$$(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\widehat{\mathbf{\Sigma}}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\mathbf{\Sigma} \simeq \widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}\mathbf{\Sigma}\widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}\mathbf{\Sigma}, \tag{D.78}$$

where $\mathbf{\Sigma}_{\boldsymbol{C}}=(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\mathbf{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}$, $\widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}=v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\mathbf{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}$. Here, $v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})$, $\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})$, and $\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})$ defined by (D.73)-(D.75) simplify to

$$\frac{1}{v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})} = \lambda+\gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\mathbf{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}\mathbf{\Sigma}]/p, \tag{D.79}$$

$$\frac{1}{\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})} = \frac{1}{v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})^2}-\gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\mathbf{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-2}\mathbf{\Sigma}^2]/p, \tag{D.80}$$

$$\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}}) = \gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}})\mathbf{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-2}\mathbf{\Sigma}^2]/p\cdot\widetilde{v}_v(-\lambda;\gamma,\mathbf{\Sigma}_{\boldsymbol{C}}). \tag{D.81}$$

If $\gamma\to\phi\in(0,\infty)$, then γ in (1)-(3) can be replaced by ϕ, with the empirical distribution H_n of eigenvalues replaced by the limiting distribution H.

Proof of Lemma D.8.10. Proofs for the different parts are separated below.

Part (1) Note that

$$\lambda(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}=\lambda(\widehat{\boldsymbol{\Sigma}}+\lambda(\boldsymbol{I}_p+\boldsymbol{C}))^{-1}=(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\lambda(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}, \tag{D.82}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}=\boldsymbol{\Sigma}_{\boldsymbol{C}}^{\frac{1}{2}}(\boldsymbol{Z}^\top\boldsymbol{Z}/n)\boldsymbol{\Sigma}_{\boldsymbol{C}}^{\frac{1}{2}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{C}}=(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}$. Using Lemma D.8.7, we have

$$\lambda(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}\simeq(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}, \tag{D.83}$$

where $v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$ is the unique solution to the fixed point equation (D.70) such that

$$\frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})}=\lambda+\gamma\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}]/p=\lambda+\gamma\operatorname{tr}[\boldsymbol{\Sigma}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}]/p.$$

Note that $\left\|(\boldsymbol{I}_p+\boldsymbol{C})^{-1}\right\|_{\rm op}\leq 1$. We can apply the product rule from Lemma D.8.4 (3) and get

$$\lambda(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\simeq(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+I_p)^{-1}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}=(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1},$$

by combining (D.82)-(D.83).

Part (2) From Lemma D.8.9 (1), we have

$$\begin{aligned}
&\lambda^2(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\\
&=\lambda^2(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\cdot[(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\cdot\boldsymbol{\Sigma}\cdot(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}]\cdot(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\\
&\simeq(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\cdot[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}\\
&\qquad\cdot(\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}})\cdot(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}]\cdot(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\\
&=(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}(\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma})(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}.
\end{aligned}$$

Part (3) Similar to Part (2), from Lemma D.8.9 (2), we have

$$\begin{aligned}
&(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma}\\
&=(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\cdot(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}+\lambda\boldsymbol{I}_p)^{-1}\cdot(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}\\
&\simeq(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\cdot\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}_{\boldsymbol{C}}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}}+\boldsymbol{I}_p)^{-1}\cdot(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}\\
&=\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}\boldsymbol{\Sigma}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}\boldsymbol{\Sigma}.
\end{aligned}$$

Note that the distribution of $\boldsymbol{\Sigma}_{\boldsymbol{C}}$'s eigenvalue has positive support. By the continuity of $v(-\lambda;\cdot,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, $\widetilde{v}_b(-\lambda;\cdot,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, and $\widetilde{v}_v(-\lambda;\cdot,\boldsymbol{\Sigma}_{\boldsymbol{C}})$ from Lemma D.8.13 (2), (4) and (3), γ can by replaced by its limit ϕ as $n,p\to\infty$. □

The following lemma concerns the deterministic equivalents of the precision matrix as the weighted average of two sample covariance matrices of subsamples, when the full sample covariance matrix is invertible almost surely. It is useful when we aim to condition on one of the subsampled covariance matrix, which is used in the proof of Lemma D.4.5.

Lemma D.8.11 (Deterministic equivalent of subsamples in the underparameterized regime)**.** *Suppose the conditions in Lemma D.8.9 holds. Let $\widehat{\boldsymbol{\Sigma}}_0$ be the sample covariance matrix computed using i observations of $\boldsymbol{X}$, and $\widehat{\boldsymbol{\Sigma}}_1$ be the sample covariance matrix computed using the remaining $n-i$ samples. Let $\pi_0=i/n$ and $\pi_1=(n-i)/n$. Suppose that $p/n\to\phi\in(0,1)$ as $n,p\to\infty$. Then, we have*

$$(\pi_0\widehat{\boldsymbol{\Sigma}}_0+\pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1}\simeq(\pi_0\widehat{\boldsymbol{\Sigma}}_0+(1-\phi)\pi_1\boldsymbol{\Sigma})^{-1}.$$

Proof. We first note that when $\phi \in (0,1)$, the eigenvalues of $\widehat{\boldsymbol{\Sigma}} = \pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \pi_1 \widehat{\boldsymbol{\Sigma}}_1$ are bounded away from zero almost surely (Bai and Silverstein, 2010) and hence the inverse is well defined almost surely as $n, p \to \infty$.

The idea for the proof is to consider the perturbed resolvent $(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1}$ for $\mu > 0$. Note that since the matrix $(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)$ is almost surely invertible. Then,

$$\lim_{\mu \to 0^+} (\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1} = (\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1}.$$

Conditioned on $(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p)$, we have

$$\begin{aligned}
(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1} &= a(\boldsymbol{A} + \widehat{\boldsymbol{\Sigma}}_1)^{-1} \\
&= a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + \boldsymbol{A}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{A}^{-\frac{1}{2}})\boldsymbol{A}^{-\frac{1}{2}} \\
&= a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + \widehat{\boldsymbol{\Sigma}}_{1,\boldsymbol{A}})^{-1}\boldsymbol{A}^{-\frac{1}{2}} \\
&\simeq a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + c\boldsymbol{\Sigma}_{\boldsymbol{A}})\boldsymbol{A}^{-\frac{1}{2}} \\
&= a(\boldsymbol{A} + c\boldsymbol{\Sigma})^{-1} \\
&= (\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + c\pi_1 \boldsymbol{\Sigma})^{-1},
\end{aligned}$$

where the intermediate constants are $a = \pi_1^{-1}$, $\boldsymbol{A} = a\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + a\mu \boldsymbol{I}_p$, $\widehat{\boldsymbol{\Sigma}}_{1,\boldsymbol{A}} = \boldsymbol{A}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}} \boldsymbol{A}^{-\frac{1}{2}}$, $\boldsymbol{\Sigma}_{\boldsymbol{A}} = \boldsymbol{A}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{A}^{-\frac{1}{2}}$, and c satisfy the fixed-point equation

$$\begin{aligned}
\frac{1}{c} &= 1 + \frac{p}{n-i} \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{A}}(c\boldsymbol{\Sigma}_{\boldsymbol{A}} + \boldsymbol{I}_p)^{-1}]/p \\
&= 1 + \frac{p}{k}\frac{k}{n-i} \operatorname{tr}[\boldsymbol{A}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{A}^{-1/2}(c\boldsymbol{A}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{A}^{-1/2} + \boldsymbol{I}_p)^{-1}]/p \\
&= 1 + \phi a \operatorname{tr}[\boldsymbol{\Sigma}(c\boldsymbol{\Sigma} + \boldsymbol{A})^{-1}]/p \\
&= 1 + \phi \operatorname{tr}[\boldsymbol{\Sigma}(c\pi_1\boldsymbol{\Sigma} + \pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p)^{-1}]/p \\
&= 1 + \phi \operatorname{tr}[\boldsymbol{\Sigma}(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1}]/p,
\end{aligned}$$

where in final equality, we used the trace property of the asymptotic equivalence

$$(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + c\pi_1 \boldsymbol{\Sigma})^{-1} \simeq (\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1}.$$

Now note that

$$(\pi_0 \widehat{\boldsymbol{\Sigma}}_0 + \mu \boldsymbol{I}_p + \pi_1 \widehat{\boldsymbol{\Sigma}}_1)^{-1} = (\widehat{\boldsymbol{\Sigma}} + \mu \boldsymbol{I}_p)^{-1} \simeq (c'\boldsymbol{\Sigma} + \mu \boldsymbol{I}_p)^{-1}$$

where c' solves the fixed-point equation

$$\frac{1}{c'} = 1 + \phi \operatorname{tr}[\boldsymbol{\Sigma}(c'\boldsymbol{\Sigma} + \mu \boldsymbol{I}_p)^{-1}]/p.$$

Thus, the fixed-point in c can be written as

$$\frac{1}{c} = 1 + \phi \operatorname{tr}[\boldsymbol{\Sigma}(c'\boldsymbol{\Sigma} + \mu \boldsymbol{I}_p)^{-1}]/p.$$

We note that $c = c'$ satisfy the fixed-point equation for c (from the fixed-point equation for c'). Since c is a unique solution, this must be the solution. Letting $\mu \to 0^+$, we observe that $c' = 1 - \phi$ is the solution for the fixed-point equation in c'. Thus, we also have $c = 1 - \phi$. □

D.8.4 Analytic properties of associated fixed-point equations

In this section, we gather results on the properties of the fixed-point solution $v(-\lambda; \phi)$ defined in (D.70).

The following lemma provides the existence and uniqueness of the solution $v(-\lambda; \phi)$. The properties of the derivatives in Lemma D.8.12 are related to the properties of $\widetilde{v}_v(-\lambda; \phi)$ defined in (D.75), which equals $-f'(x)$, where the function f is defined in (D.84).

Lemma D.8.12 (Properties of the solution to the fixed-point equation)**.** *Let $\lambda, \phi, a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Define the function*

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{1 + rx} \mathrm{d}P(r) - \lambda. \tag{D.84}$$

Then the following properties hold:

(1) For $\lambda = 0$ and $\phi \in (1, \infty)$, there is a unique $x_0 \in (0, \infty)$ such that $f(x_0) = 0$. The function f is positive and strictly decreasing over $(0, x_0)$ and negative over (x_0, ∞), with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = 0$.

(2) For $\lambda > 0$ and $\phi \in (0, \infty)$, there is a unique $x_0^\lambda \in (0, \infty)$ such that $f(x_0^\lambda) = 0$. The function f is positive and strictly decreasing over $(0, x_0^\lambda)$ and negative over (x_0^λ, ∞), with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = -\lambda$.

(3) For $\lambda = 0$ and $\phi \in (1, \infty)$, f is differentiable on $(0, \infty)$ and its derivative f' is strictly increasing over $(0, x_0)$, with $\lim_{x \to 0^+} f'(x) = -\infty$ and $f'(x_0) < 0$.

(4) For $\lambda > 0$ and $\phi \in (0, \infty)$, f is differentiable on $(0, \infty)$ and its derivative f' is strictly increasing over $(0, \infty)$, with $\lim_{x \to 0^+} f'(x) = -\infty$ and $f'(x_0^\lambda) < 0$.

Proof of Lemma D.8.12. We consider different parts separately below.

Part (1) Observe that

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{xr + 1} \, \mathrm{d}P(r) = g_1(x) h_1(x),$$

where

$$g_1(x) = \frac{1}{x}, \qquad h_1(x) = 1 - \phi \int \frac{xr}{xr + 1} \, \mathrm{d}P(r).$$

Note that g_1 is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g_1(x) = \infty$ and $\lim_{x \to \infty} g_1(x) = 0$, while h_1 is strictly decreasing over $(0, \infty)$ with $h_1(0) = 1$ and $\lim_{x \to \infty} h_1(x) = 1 - \phi < 0$. Thus, there is a unique $0 < x_0 < \infty$ such that $h_1(x_0) = 0$, and consequently $f(x_0) = 0$. Because h_1 is positive over $(0, x_0)$, and negative over (x_0, ∞), f is positive strictly decreasing over $(0, x_0)$ and negative over (x_0, ∞), with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = 0$.

Part (2) Note that $f(x) = g_1(x) h_1(x) - \lambda$. Since from (1) $\lim_{x \to 0} g_1(x) h_1(x) = \infty$ and $\lim_{x \to 0} g_1(x) h_1(x) = 0$, we have that $\lim_{x \to 0^+} f(x) = +\infty$ and $\lim_{x \to \infty} f(x) = -\lambda < 0$.

For $\phi > 1$, since $g_1(x) h_1(x)$ is positive and strictly decreasing over $(0, x_0)$ and negative over (x_0, ∞), and $\lim_{x \to 0^+} g_1(x) h_1(x) = \infty$, we have that there exists $x_0^\lambda \in (0, x_0)$ such that $f(x_0^\lambda) = 0$. The properties of f over $(0, x_0^\lambda)$ and (x_0^λ, ∞) follow analogously as in (1).

For $\phi \in (0, 1]$, since $g_1 h_1$ is continuous, positive and strictly decreasing over $(0, \infty)$, by intermediate value theorem, there exists $x_0^\lambda \in (0, \infty)$ such that $f(x_0^\lambda) = 0$, f is positive and strictly decreasing for $x < x_0^\lambda$ and negative for $x > x_0^\lambda$, with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = -\lambda$.

Part (3) Since f is monotone and continuous, it is differentiable. The derivative f' at x is given by

$$f'(x) = -\frac{1}{x^2} + \phi \int \frac{r^2}{(xr + 1)^2} \, \mathrm{d}P(r) = -g_2(x) h_2(x),$$

where

$$g_2(x) = \frac{1}{x^2}, \qquad h_2(x) = \left(1 - \phi \int \left(\frac{xr}{xr + 1} \right)^2 \mathrm{d}P(r). \right)$$

Note that the function g_2 is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g_2(x) = \infty$ and $\lim_{x \to \infty} g_2(x) = 0$. On the other hand, the function h_2 is strictly decreasing over $(0, \infty)$ with $h_2(0) = 1$ and $h_2(x_0) > 0$. This follows because for $x \in (0, x_0]$,

$$\phi \int \left(\frac{xr}{xr+1} \right)^2 \mathrm{d}P(r) \leq \frac{x_0 b}{x_0 b + 1} \phi \int \left(\frac{xr}{xr+1} \right) \mathrm{d}P(r) < \int \frac{\phi xr}{xr+1} \mathrm{d}P(r) \leq \int \frac{\phi x_0 r}{x_0 r + 1} \mathrm{d}P(r) = 1,$$

where the first inequality in the chain above follows as the support of P is $[a, b]$, and the last inequality follows since $f(x_0) = 0$ and $x_0 > 0$, which implies that

$$\frac{1}{x_0} = \phi \int \frac{r}{x_0 r + 1} \mathrm{d}P(r), \quad \text{or equivalently that} \quad 1 = \phi \int \frac{x_0 r}{x_0 r + 1} \mathrm{d}P(r).$$

Thus, $-f'$, a product of two positive strictly decreasing functions, is strictly decreasing, and in turn, f' is strictly increasing. Moreover, $\lim_{x \to 0^+} f'(x) = -\infty$ and $f'(x_0) < 0$.

Part (4) The conclusion follows by noting that $h_2(x_0^\lambda) > h_2(x_0) > 0$ from (3). □

The continuity and limiting behavior of the function $\phi \mapsto v(-\lambda; \phi)$ is given for ridge regression ($\lambda > 0$) in Lemma D.8.13 and for ridgeless regression ($\lambda = 0$) in Lemma D.8.14.

Lemma D.8.13 (Continuity in the aspect ratio for ridge regression)**.** *Let $\lambda, a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(-\lambda; \cdot) : \phi \mapsto v(-\lambda; \phi)$, over $(0, \infty)$, where $v(-\lambda; \phi) > 0$ is the unique solution to the fixed-point equation*

$$\frac{1}{v(-\lambda; \phi)} = \lambda + \phi \int \frac{r}{1 + rv(-\lambda; \phi)} \mathrm{d}P(r) \tag{D.85}$$

Then the following properties hold:

(1) The range of the function $v(-\lambda; \cdot)$ is a subset of $(0, \lambda^{-1})$.

(2) The function $v(-\lambda; \cdot)$ is continuous and strictly decreasing over $(0, \infty)$. Furthermore, $\lim_{\phi \to 0^+} v(-\lambda; \phi) = \lambda^{-1}$, and $\lim_{\phi \to \infty} v(-\lambda; \phi) = 0$.

(3) The function $\widetilde{v}_v(-\lambda; \cdot) : \phi \mapsto \widetilde{v}_v(-\lambda; \phi)$, where

$$\widetilde{v}_v(-\lambda; \phi) = \left(v(-\lambda; \phi)^{-2} - \int \phi r^2 (1 + rv(-\lambda; \phi))^{-2} \, \mathrm{d}P(r) \right)^{-1},$$

is positive and continuous over $(0, \infty)$. Furthermore, $\lim_{\phi \to 0^+} \widetilde{v}_v(-\lambda; \phi) = \lambda^{-2}$, and $\lim_{\phi \to \infty} \widetilde{v}_v(-\lambda; \phi) = 0$.

(4) The function $\widetilde{v}_b(-\lambda; \cdot) : \phi \mapsto \widetilde{v}_b(-\lambda; \phi)$, where

$$\widetilde{v}_b(-\lambda; \phi) = \widetilde{v}_v(-\lambda; \phi) \int \phi r^2 (1 + v(-\lambda; \phi) r)^{-2} \, \mathrm{d}P(r),$$

is positive and continuous over $(0, \infty)$. Furthermore, $\lim_{\phi \to 0^+} \widetilde{v}_b(-\lambda; \phi) = \lim_{\phi \to \infty} \widetilde{v}_b(-\lambda; \phi) = 0$.

Proof of Lemma D.8.13. Proofs for the different parts appear below.

Part (1) Since P has positive support, we have

$$\frac{1}{v(-\lambda; \phi)} = \lambda + \phi \int \frac{r}{1 + rv(-\lambda; \phi)} \mathrm{d}P(r) > \lambda,$$
$$\frac{1}{v(-\lambda; \phi)} = \lambda + \phi \int \frac{r}{1 + rv(-\lambda; \phi)} \mathrm{d}P(r) < \lambda + \phi b$$

which implies that $0 < (\lambda + \phi b)^{-1} < v(-\lambda; \phi) < \lambda^{-1}$.

Part (2) Rearranging (D.85) yields

$$\frac{1}{\phi} = \frac{1}{1 - \lambda v(-\lambda;\phi)} \left(1 - \int \frac{1}{1 + rv(-\lambda;\phi)} \mathrm{d}P(r)\right).$$

From Patil et al. (2022a, Lemma S.6.13), the function

$$h_1 : t \mapsto 1 - \int \frac{1}{1 + rt} \mathrm{d}P(r)$$

is strictly increasing and continuous over $(0, \infty)$, $\lim_{t\to 0} h_1(t) = 0$, and $\lim_{t\to\infty} h_1(t) = 1$. It is also positive from (1). Since $h_2 : t \mapsto 1/(1 - \lambda t)$ is positive, strictly increasing and continuous over $t \in (0, \lambda^{-1})$, we have that the function

$$T : t \mapsto \frac{1}{1 - \lambda t} \left(1 - \int \frac{1}{1 + rt} \mathrm{d}P(r)\right)$$

is strictly increasing and continuous over $(0, \lambda^{-1})$. By the continuous inverse theorem, we have T^{-1} is strictly increasing and continuous. Note that $v(-\lambda;\phi) = T^{-1}(\phi^{-1})$. Since $\phi \mapsto \phi^{-1}$ is continuous and strictly decreasing in ϕ, we have $\phi \mapsto v(-\lambda;\phi)$ is continuous and strictly decreasing over $\phi \in (0, \infty)$. Moreover, $\lim_{\phi\to 0^+} T^{-1}(\phi^{-1}) = \lambda^{-1}$ and $\lim_{\phi\to\infty} T^{-1}(\phi^{-1}) = 0$.

Part (3) From (2), $\phi \mapsto v(-\lambda;\phi)^{-2}$ is continuous in ϕ and

$$T_2 : \phi \mapsto \phi \int \frac{r^2}{(1 + rv(-\lambda;\phi))^2} \,\mathrm{d}P(r)$$

is also continuous in ϕ. Thus, the function $\widetilde{v}_v(-\lambda;\cdot)^{-1}$ is continuous. Note that

$$\frac{v(-\lambda;\phi)^2}{\widetilde{v}_v(-\lambda;\phi)} = 1 - \phi \int \frac{r^2 v(-\lambda;\phi)^2}{(1 + rv(-\lambda;\phi))^2} \mathrm{d}P(r) > 1 - \phi \int \frac{rv(-\lambda;\phi)}{1 + rv(-\lambda;\phi)} \mathrm{d}P(r) = 0,$$

where the inequality holds because $rv(-\lambda;\phi)/(1 + rv(-\lambda;\phi))$ is strictly positive and $P(r)$ has positive support. Then we have that $\phi \mapsto \widetilde{v}_v(-\lambda;\phi)^{-1} > 0$ and $\widetilde{v}_v(-\lambda;\cdot)$ is continuous over $(0, \infty)$. Since $\lim_{\phi\to 0^+} v(-\lambda;\phi) = \lambda^{-1}$, it follows that $\lim_{\phi\to 0^+} \widetilde{v}_v(-\lambda;\phi) = \lambda^{-2}$. Similarly, from $\lim_{\phi\to\infty} v(-\lambda;\phi) = 0$, $\lim_{\phi\to\infty} \phi v(-\lambda;\phi) = 1$ and the fact that

$$\lim_{\phi\to\infty} \int \frac{r^2}{(1 + rv(-\lambda;\phi))^2} \,\mathrm{d}P(r) \in [a^2, b^2],$$

it follows that

$$\lim_{\phi\to\infty} \widetilde{v}_v(-\lambda;\phi) = \lim_{\phi\to\infty} v(-\lambda;\phi)^2 \cdot \left(1 - v(-\lambda;\phi) \cdot \phi v(-\lambda;\phi) \cdot \int r^2 (1 + rv(-\lambda;\phi))^{-2} \,\mathrm{d}P(r)\right)^{-1} = 0.$$

Part (4) The continuity of $\widetilde{v}_b(-\lambda;\cdot)$ follows from the continuity of $v(-\lambda;\cdot)$ and $\widetilde{v}_v(-\lambda;\cdot)$. Note that

$$\frac{1}{1 + \widetilde{v}_b(-\lambda;\phi)} = 1 - v(-\lambda;\phi) \cdot \phi v(-\lambda;\phi) \cdot \int \frac{r^2}{(1 + rv(-\lambda;\phi))^2} \,\mathrm{d}P(r).$$

From the proof in (3), we have

$$\begin{aligned} \lim_{\phi\to 0^+} \frac{1}{1 + \widetilde{v}_b(-\lambda;\phi)} &= 1 - \lim_{\phi\to 0^+} v(-\lambda;\phi) \cdot \phi v(-\lambda;\phi) \cdot \int \frac{r^2}{(1 + rv(-\lambda;\phi))^2} \,\mathrm{d}P(r) = 1 \\ \lim_{\phi\to\infty} \frac{1}{1 + \widetilde{v}_b(-\lambda;\phi)} &= 1 - \lim_{\phi\to\infty} v(-\lambda;\phi) \cdot \phi v(-\lambda;\phi) \cdot \int \frac{r^2}{(1 + rv(-\lambda;\phi))^2} \,\mathrm{d}P(r) = 1 \end{aligned}$$

and thus, $\lim_{\phi\to 0^+} \widetilde{v}_b(-\lambda;\phi) = \lim_{\phi\to\infty} \widetilde{v}_b(-\lambda;\phi) = 0$. □

Lemma D.8.14 (Continuity in the aspect ratio for ridgeless regression, adapted from Patil et al. (2022a))**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(0; \cdot) : \phi \mapsto v(0; \phi)$, over $(1, \infty)$, where $v(0; \phi) > 0$ is the unique solution to the fixed-point equation*

$$\frac{1}{\phi} = \int \frac{v(0;\phi) r}{1 + v(0;\phi) r} \, \mathrm{d}P(r). \tag{D.86}$$

Then the following properties hold:

(1) The function $v(0; \cdot)$ is continuous and strictly decreasing over $(1, \infty)$. Furthermore, $\lim_{\phi \to 1^+} v(0; \phi) = \infty$, and $\lim_{\phi \to \infty} v(0; \phi) = 0$.

(2) The function $\phi \mapsto (\phi v(0; \phi))^{-1}$ is strictly increasing over $(1, \infty)$. Furthermore, $\lim_{\phi \to 1^+} (\phi v(0; \phi))^{-1} = 0$ and $\lim_{\phi \to \infty} (\phi v(0; \phi))^{-1} = 1$.

(3) The function $\widetilde{v}_v(0; \cdot) : \phi \mapsto \widetilde{v}_v(0; \phi)$, where

$$\widetilde{v}_v(0;\phi) = \left(v(0;\phi)^{-2} - \phi \int r^2 (1 + r v(0;\phi))^{-2} \, \mathrm{d}P(r) \right)^{-1},$$

is positive and continuous over $(1, \infty)$. Furthermore, $\lim_{\phi \to 1^+} \widetilde{v}_v(0; \phi) = \infty$, and $\lim_{\phi \to \infty} \widetilde{v}_v(0; \phi) = 0$.

(4) The function $\widetilde{v}_b(0; \cdot) : \phi \mapsto \widetilde{v}_b(0; \phi)$, where

$$\widetilde{v}_b(0;\phi) = \widetilde{v}_v(0;\phi) \int r^2 (1 + v(0;\phi) r)^{-2} \, \mathrm{d}P(r),$$

is positive and continuous over $(1, \infty)$. Furthermore, $\lim_{\phi \to 1^+} \widetilde{v}_b(0; \phi) = \infty$, and $\lim_{\phi \to \infty} \widetilde{v}_b(0; \phi) = 0$.

The continuity and differentiabilty of the function $\lambda \mapsto v(-\lambda; \phi)$ on a closed interval $[0, \lambda_{\max}]$ for some constant $\lambda_{\max}$ is given for $\phi \in (1, \infty)$ in Lemma D.8.15 adapted from Patil et al. (2022a). This ensures that $v(0; \phi) = \lim_{\lambda \to 0^+} v(-\lambda; \phi)$ is well-defined for $\phi > 1$ and also implies that related functions are bounded.

Lemma D.8.15 (Differentiability in the regularization parameter for $\phi \in (1, \infty)$, adapted from Patil et al. (2022a))**.** *Let $0 < a \leq b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Let $\phi \in (1, \infty)$ be a real number. Let $\Lambda = [0, \lambda_{\max}]$ for some constant $\lambda_{\max} \in (0, \infty)$. For $\lambda \in \Lambda$, let $v(-\lambda; \phi) > 0$ denote the solution to the fixed-point equation*

$$\frac{1}{v(-\lambda;\phi)} = \lambda + \phi \int \frac{r}{v(-\lambda;\phi) r + 1} \, \mathrm{d}P(r).$$

Then, the function $\lambda \mapsto v(-\lambda; \phi)$ is twice differentiable over Λ. Furthermore, over Λ, $v(-\lambda; \phi)$, $\partial/\partial\lambda[v(-\lambda; \phi)]$, and $\partial^2/\partial\lambda^2[v(-\lambda; \phi)]$ are bounded.

Lemma D.8.16 (Substitutability of the fixed-point solution)**.** *Let $v : \mathbb{R}^{p \times p} \to \mathbb{R}$ and $f(v(\boldsymbol{C}), \boldsymbol{C}) : \mathbb{R}^{p \times p} \to \mathbb{R}^{p \times p}$ be a matrix function for matrix $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ and $p \in \mathbb{N}$, that is continuous in the first augment with respect to operator norm. If $v(\boldsymbol{C}) \overset{a.s.}{=} v(\boldsymbol{D})$ such that $\boldsymbol{C}$ is independent to $\boldsymbol{D}$, then $f(v(\boldsymbol{C}), \boldsymbol{C}) \simeq f(v(\boldsymbol{D}), \boldsymbol{C}) \mid \boldsymbol{C}$.*

Proof. For any matrix $\boldsymbol{T}$ whose trace norm is bounded by M, conditioning on $\{\boldsymbol{C}\}_{p \geq 1}$, we have

$$\begin{aligned} |\operatorname{tr}[(f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C}))\boldsymbol{T}]| &\leq \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}} \operatorname{tr}(\boldsymbol{T}) \\ &\leq M \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}}. \end{aligned}$$

Since $v(\boldsymbol{C}) \xrightarrow{\text{a.s.}} v(\boldsymbol{D})$ and f is continuous in the first argument with respect to operator norm, we have $\lim_{p \to \infty} \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}} = 0$. Thus,

$$\lim_{p \to \infty} |\operatorname{tr}[(f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C}))\boldsymbol{T}]| = 0,$$

conditioning on $\{\boldsymbol{C}\}_{p \geq 1}$. □

The lemma below specializes the solution to the fixed-point equations under the isotopic model.

Lemma D.8.17 (Properties of the fixed-point solution with isotopic features)**.** *Let P be a probability measure supported on $\{a\}$ for $a > 0$. For $\lambda > 0$ and $\phi > 0$, the fixed-point equation*

$$\frac{1}{v(-\lambda;\phi)} = \lambda + \phi \int \frac{r}{v(-\lambda;\phi)r + 1} \, \mathrm{d}P(r) = \lambda + \frac{\phi a}{1 + v(-\lambda;\phi)a}$$

has a closed-form solution

$$v(-\lambda;\phi) = \frac{-(\lambda/a + \phi - 1) + \sqrt{(\lambda/a + \phi - 1)^2 + 4\lambda/a}}{2\lambda}.$$

Define $\widetilde{v}_b(-\lambda;\phi)$ and $\widetilde{v}_v(-\lambda;\phi)$ via the follow equations:

$$\widetilde{v}_b(-\lambda;\phi) = \frac{\int \phi r^2 (1 + v(-\lambda;\phi)r)^{-2} \mathrm{d}P(r)}{v(-\lambda;\phi)^{-2} - \int \phi r^2 (1 + v(-\lambda;\phi)r)^{-2} \mathrm{d}P(r)},$$

$$\widetilde{v}_v(-\lambda;\phi)^{-1} = v(-\lambda;\phi)^{-2} - \int \phi r^2 (1 + v(-\lambda;\phi)r)^{-2} \mathrm{d}P(r).$$

As $\lambda \to 0^+$, we have

$$\begin{array}{lllll}
(1) & \phi \in (0,1): & v(0;\phi) = \infty, & \widetilde{v}_b(0;\phi) = \dfrac{\phi}{1-\phi}, & \widetilde{v}_v(0;\phi) = \infty, \\
(2) & \phi = 1: & v(0;\phi) = \infty, & \widetilde{v}_b(0;\phi) = \infty, & \widetilde{v}_v(0;\phi) = \infty, \\
(3) & \phi \in (1,\infty): & v(0;\phi) = \dfrac{1}{a(\phi - 1)}, & \widetilde{v}_b(0;\phi) = \dfrac{1}{\phi - 1}, & \widetilde{v}_v(0;\phi) = \dfrac{\phi}{a^2(\phi-1)^3}, \\
(4) & \phi = \infty: & v(0;\phi) = 0, & \widetilde{v}_b(0;\phi) = 0, & \widetilde{v}_v(0;\phi) = 0,
\end{array}$$

Proof of Lemma D.8.17. For $\phi \in (0,1)$, we have $v(0;\phi) = \lim_{\lambda\to 0^+} v(-\lambda;\phi) = \infty$. For $\phi > 1$,

$$v(0;\phi) = \lim_{\lambda\to 0^+} v(-\lambda;\phi) = \frac{1}{2a} \lim_{\lambda\to 0^+} \left(-1 + \frac{\lambda/a + \phi + 1}{\sqrt{(\lambda/a + \phi - 1)^2 + 4\lambda/a}} \right) = \frac{1}{a(\phi - 1)},$$

by applying the L'Hospital's rule for indeterminate forms. When $\phi = 1$, we have

$$v(0;1) = \lim_{\lambda\to 0^+} v(-\lambda;1) = \lim_{\lambda\to 0^+} \frac{1}{2a} \left(-1 + \sqrt{1 + \frac{a}{\lambda}} \right) = \infty.$$

Since $\widetilde{v}_b(0;\phi)$ and $\widetilde{v}_v(0;\phi)$ are continuous functions of $v(0;\phi)$, we have

$$\widetilde{v}_v(0;\phi) = \begin{cases} \infty, & \phi \in (0,1] \\ \dfrac{\phi}{a^2(\phi-1)^3}, & \phi \in (1,\infty) \end{cases}$$

and $\widetilde{v}_b(0;\phi) = 1/(\phi - 1)$ for $\phi \in (1,\infty)$. For $\phi \in (0,1]$, we apply the L'Hospital' rule to obtain $\widetilde{v}_b(0;\phi) = \phi/(1-\phi)$.

□

D.9 Helper concentration results

D.9.1 Size of intersection of randomly sampled datasets

In this section, we collect various helper results concerned with concentrations and convergences that are used in the proofs of Lemma 4.3.8, Lemmas D.3.4, D.3.5 and D.4.5.

Below we recall the definition of a hypergeometric random variable, along with its mean and variance. See, e.g., Greene and Wellner (2017) for more related details.

Definition D.9.1 (Hypergeometric random variable)**.** A random variable X follows the hypergeometric distribution $X \sim \mathrm{Hypergeometric}(n, K, N)$ if its probability mass function is given by

$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad \max\{0, n + K - N\} \leq k \leq \min\{n, K\}.$$

The expectation and variance of X are given by

$$\mathbb{E}[X] = \frac{nK}{N}, \qquad \mathrm{Var}(X) = \frac{nK(N-K)(N-n)}{N^2(N-1)}.$$

The following lemma provides tail bounds for the number of shared observations in two simple random samples, which is adapted from (Hoeffding, 1963; Serfling, 1974). See also Greene and Wellner (2017).

Lemma D.9.2 (Concentration bounds for the number of shared observations)**.** *For $n \in \mathbb{N}$, define $\mathcal{I}_k := \{\{i_1, i_2, \ldots, i_k\} : 1 \leq i_1 < i_2 < \ldots < i_k \leq n\}$. Let $I_1, I_2 \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k$, define the random variable $i_0^{\mathtt{SRSWR}} := |I_1 \cap I_2|$ to be the number of shared samples, and define $i_0^{\mathtt{SRSWOR}}$ accordingly. Then we have*

(1) $i_0^{\mathtt{SRSWR}}$ follows a binomial distribution, $i_0^{\mathtt{SRSWR}} \sim \mathrm{Binomial}(k, k/n)$ with mean $\mathbb{E}[i_0^{\mathtt{SRSWR}}] = k^2/n$. It holds that for all $t > 0$,

$$\mathbb{P}\left(i_0^{\mathtt{SRSWR}} - \mathbb{E}[i_0^{\mathtt{SRSWR}}] \geq kt\right) \leq \exp\left(-2kt^2\right).$$

(2) $i_0^{\mathtt{SRSWOR}}$ follows a hypergeometric distribution, $i_0^{\mathtt{SRSWOR}} \sim \mathrm{Hypergeometric}(k, k, n)$ with mean $\mathbb{E}[i_0^{\mathtt{SRSWOR}}] = k^2/n$. It holds that for all $t > 0$,

$$\mathbb{P}\left(i_0^{\mathtt{SRSWOR}} - \mathbb{E}[i_0^{\mathtt{SRSWOR}}] \geq kt\right) \leq \exp\left(-\frac{2nkt^2}{n-k+1}\right). \tag{D.87}$$

The following lemma characterize the limiting proportions of shared observations in two simple random samples under proportional asymptotics, when both the subsample size and the full data size tend to infinity.

Lemma D.9.3 (Asymptotic proportions of shared observations)**.** *Consider the setting in Lemma D.9.2. Let $\{k_m\}_{m=1}^{\infty}$ and $\{n_m\}_{m=1}^{\infty}$ be two sequences of positive integers such that n_m is strictly increasing in m, $n_m^{\nu} \leq k_m \leq n_m$ for some constant $\nu \in (0,1)$ and $k_m/n_m \to \omega_s \in [0,1]$. Then, $i_0^{\mathtt{SRSWR}} \xrightarrow{\text{a.s.}} \omega_s$, and $i_0^{\mathtt{SRSWOR}} \xrightarrow{\text{a.s.}} \omega_s$.*

Proof. The two parts are split below.

Part 1. For all $\delta > 0$,

$$\sum_{m=1}^{\infty} \mathbb{P}\left(\frac{1}{k_m}|i_0^{\mathtt{SRSWR}} - \mathbb{E}[i_0^{\mathtt{SRSWR}}]| > \delta\right) \leq 2\sum_{m=1}^{\infty} \exp\left(-2k_m\delta^2\right).$$

Because $k_m, n_m \to \infty$ and $k_m = \Omega(n_m^{\nu})$, there exists $m_0 \in \mathbb{N}$, such that for all $m > m_0$, $\exp(-2k_m\delta^2) \leq n_m^{-(1+\nu)}$. Thus,

$$\sum_{m=1}^{\infty} \mathbb{P}\left(\frac{1}{k_m}|i_0^{\mathtt{SRSWR}} - \mathbb{E}[i_0^{\mathtt{SRSWR}}]| > \delta\right) \leq 2\sum_{m=1}^{m_0} \exp\left(-2k_m\delta^2\right) + 2\sum_{m=m_0}^{\infty} \frac{1}{n_m^{1+\nu}} < \infty.$$

By the Borel–Cantelli lemma, we have

$$\frac{i_0^{\mathtt{SRSWR}}}{k_m} - \frac{\mathbb{E}[i_0^{\mathtt{SRSWR}}]}{k_m} \xrightarrow{\text{a.s.}} 0.$$

As $\lim_{m\to\infty} \mathbb{E}[i_0^{\mathtt{SRSWR}}]/k_m = \lim_{m\to\infty} k_m/n_m = \omega_s$, we further have $i_0^{\mathtt{SRSWR}}/k_m \xrightarrow{\text{a.s.}} \omega_s$.

Part 2. Note that

$$\mathbb{P}\left(i_0^{\texttt{SRSWOR}} - \mathbb{E}[i_0^{\texttt{SRSWOR}}] \geq kt\right) \leq \exp\left(-\frac{2nkt^2}{n-k+1}\right) \leq \exp\left(-2kt^2\right).$$

The conclusion then follows analogously as in Part 1. □

D.9.2 Convergence of random linear and quadratic forms

In this section, we collect helper lemmas on concentration of linear and quadratic forms of random vectors that are used in the proofs of Lemmas D.3.2, D.3.3, D.4.2 and D.4.3.

The following lemma provides concentration of a linear form of a random vector with independent components. It follows from a moment bound from Lemma 7.8 of Erdos and Yau (2017), along with the Borel-Cantelli lemma, and is adapted from Lemma S.8.5 of Patil et al. (2022a).

Lemma D.9.4 (Concentration of linear form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \dots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \leq M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $\boldsymbol{a}_p \in \mathbb{R}^p$ be a sequence of random vectors independent of $\boldsymbol{z}_p$ such that $\limsup_p \|\boldsymbol{a}_p\|^2/p \leq M_0$ almost surely for a constant $M_0 < \infty$. Then, $\boldsymbol{a}_p^\top \boldsymbol{z}_p/p \to 0$ almost surely as $p \to \infty$.*

The following lemma provides concentration of a quadratic form of a random vector with independent components. It follows from a moment bound from Lemma B.26 of Bai and Silverstein (2010), along with the Borel-Cantelli lemma, and is adapted from Lemma S.8.6 of Patil et al. (2022a).

Lemma D.9.5 (Concentration of quadratic form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \dots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \leq M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $\boldsymbol{D}_p \in \mathbb{R}^{p\times p}$ be a sequence of random matrix such that $\limsup \|\boldsymbol{D}_p\|_{\mathrm{op}} \leq M_0$ almost surely as $p \to \infty$ for some constant $M_0 < \infty$. Then, $\boldsymbol{z}_p^\top \boldsymbol{D}_p \boldsymbol{z}_p/p - \mathrm{tr}[\boldsymbol{D}_p]/p \to 0$ almost surely as $p \to \infty$.*

D.9.3 Convergence of Ceŝaro-type mean and max for triangular array

In this section, we collect a helper lemma on deducing almost sure convergence of a Ceŝaro-type mean from almost sure convergence of the original sequence. It is used in the proof of Proposition 4.3.3 and Lemma 4.3.8.

Lemma D.9.6 (Convergence of conditional expectation)**.** *For $n \in \mathbb{N}$, suppose $\{R_{n,\ell}\}_{\ell=1}^{N_n}$ is a set of N_n random variables defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $1 < N_n < \infty$ almost surely. If there exists a constant c such that $R_{n,p_n} \xrightarrow{\text{a.s.}} c$ for all deterministic sequences $\{p_n \in [N_n]\}_{n=1}^\infty$, then the following statements hold:*

(1) $\max_{\ell\in[N_n]} |R_{n,\ell}(\omega) - c| \xrightarrow{\text{a.s.}} 0$,

(2) $N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} \xrightarrow{\text{a.s.}} c$.

Proof of Lemma D.9.6. Proofs for the two parts are split below.

Part (1) We concatenate the sets $\{R_{n,\ell}\}_{\ell=1}^{N_n}$ for all $n \in \mathbb{N}$ to form a new sequence

$$W = (W_1, W_2, \cdots) = (R_{1,1}, \cdots, R_{1,N_1}, R_{2,1}, \cdots, R_{2,N_2}, \cdots).$$

That is, $W_t = R_{n,\ell}$ for $t = \sum_{j=1}^n N_j + \ell$. See Figure D.2 for an illustration. Because $N_n \to \infty$ if and only if $n \to \infty$ if and only if $t \to \infty$, it holds that $W_t \xrightarrow{\text{a.s.}} c$ as $t \to \infty$ Then, by Shiryaev (2016, Chapter 2, Section 10, Theorem 1), we have that for all $\epsilon > 0$,

$$\lim_{s\to\infty} \mathbb{P}\left(\bigcup_{t=s}^{\infty} \{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\}\right) = 0.$$

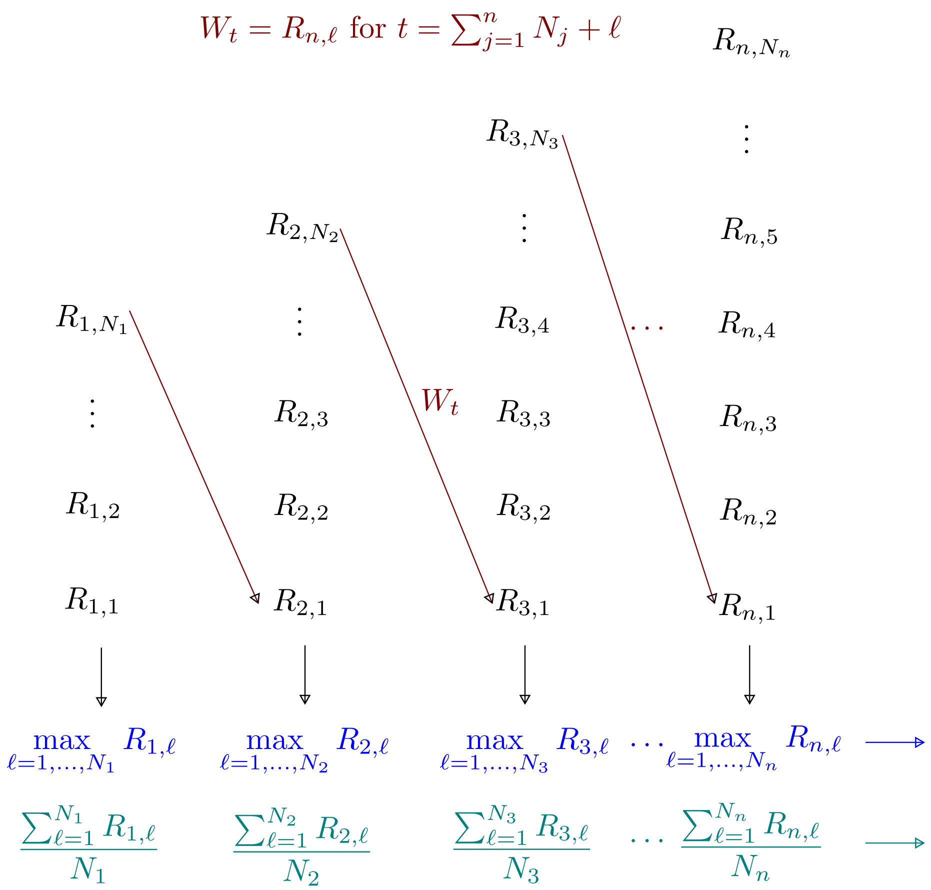

$W_t = R_{n,\ell}$ for $t = \sum_{j=1}^{n} N_j + \ell$
R_{1,N_1} $\vdots$ $R_{1,2}$ $R_{1,1}$
R_{2,N_2} $\vdots$ $R_{2,3}$ $R_{2,2}$ $R_{2,1}$
W_t
R_{3,N_3} $\vdots$ $R_{3,4}$ $R_{3,3}$ $R_{3,2}$ $R_{3,1}$
$\dots$
R_{n,N_n} $\vdots$ $R_{n,5}$ $R_{n,4}$ $R_{n,3}$ $R_{n,2}$ $R_{n,1}$
$\max_{\ell=1,\dots,N_1} R_{1,\ell}$ $\max_{\ell=1,\dots,N_2} R_{2,\ell}$ $\max_{\ell=1,\dots,N_3} R_{3,\ell}$ $\dots$ $\max_{\ell=1,\dots,N_n} R_{n,\ell}$
$\frac{\sum_{\ell=1}^{N_1} R_{1,\ell}}{N_1}$ $\frac{\sum_{\ell=1}^{N_2} R_{2,\ell}}{N_2}$ $\frac{\sum_{\ell=1}^{N_3} R_{3,\ell}}{N_3}$ $\dots$ $\frac{\sum_{\ell=1}^{N_n} R_{n,\ell}}{N_n}$

Figure D.2: Illustration of the concatenated sequence $\{W_t\}$ (in maroon) constructed from the triangle array $\{R_{n,\ell}\}_{\ell=1}^{N_n}, n \in \mathbb{N}$ (in black), used in the proof of Lemma D.9.6, along with the max sequence (in blue) and the average sequence (in teal).

Now, for $s \in \mathbb{N}$, let m be the smallest natural number such that $\sum_{j=1}^{m} N_j \geq s$. Since

$$\begin{aligned}\bigcup_{t=s}^{\infty}\{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\} &\supseteq \bigcup_{n=m}^{\infty} \bigcup_{\ell=1}^{N_n} \{\omega \in \Omega : |R_{n,\ell}(\omega) - c| > \epsilon\} \\ &= \bigcup_{n=m}^{\infty} \left\{\omega \in \Omega : \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| > \epsilon\right\}.\end{aligned}$$

We further have

$$0 \leq \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} \left\{\omega \in \Omega : \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| > \epsilon\right\}\right) \leq \lim_{s\to\infty} \mathbb{P}\left(\bigcup_{t=s}^{\infty}\{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\}\right) = 0,$$

or in other words,

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} \left\{\omega \in \Omega : \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| > \epsilon\right\}\right) = 0.$$

Thus, we have that $\max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| \xrightarrow{\text{a.s.}} 0$ by Shiryaev (2016, Chapter 2, Section 10, Theorem 1).

Part (2) We will use the first part. Note that by triangle inequality,

$$\left| N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} - c \right| \leq N_n^{-1} \sum_{\ell=1}^{N_n} |R_{n,\ell} - c| \leq \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| \, .$$

Invoking the first part, we have that $N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} \xrightarrow{\text{a.s.}} c$.

□

D.10 Additional numerical illustrations

D.10.1 Additional illustrations for Theorem 4.4.1

D.10.1.1 Prediction risk curves for subagged ridgeless and ridge predictors with varying M

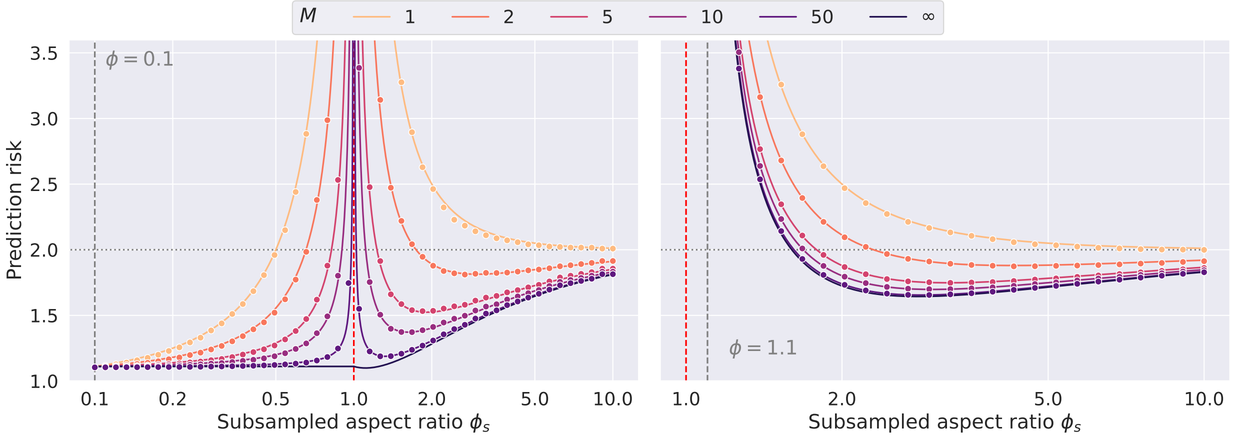


Figure D.3: Asymptotic prediction risk curves in (4.23) for ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

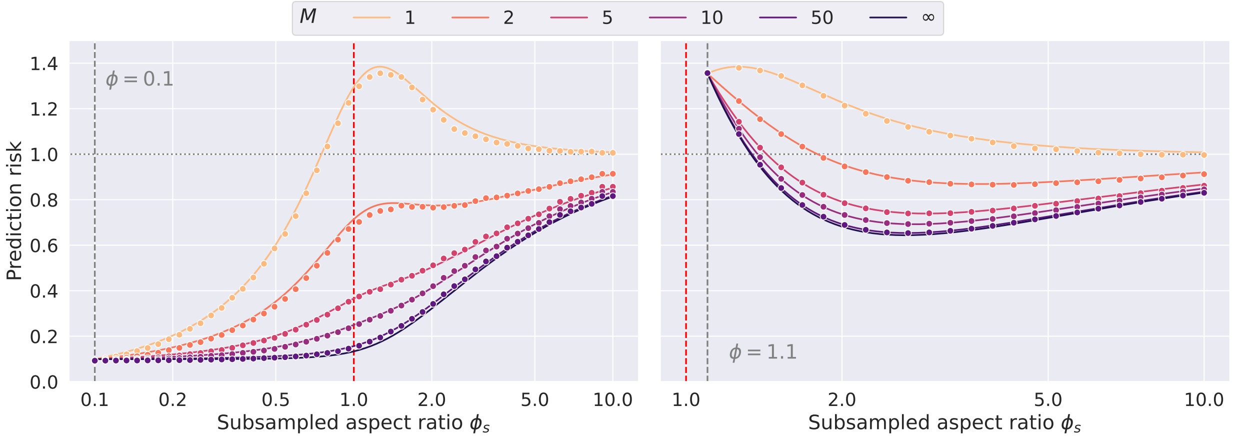


Figure D.4: Asymptotic prediction risk curves in (4.23) for subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

D.10.1.2 Bias-variance curves for subagged ridgeless and ridge predictors with varying M

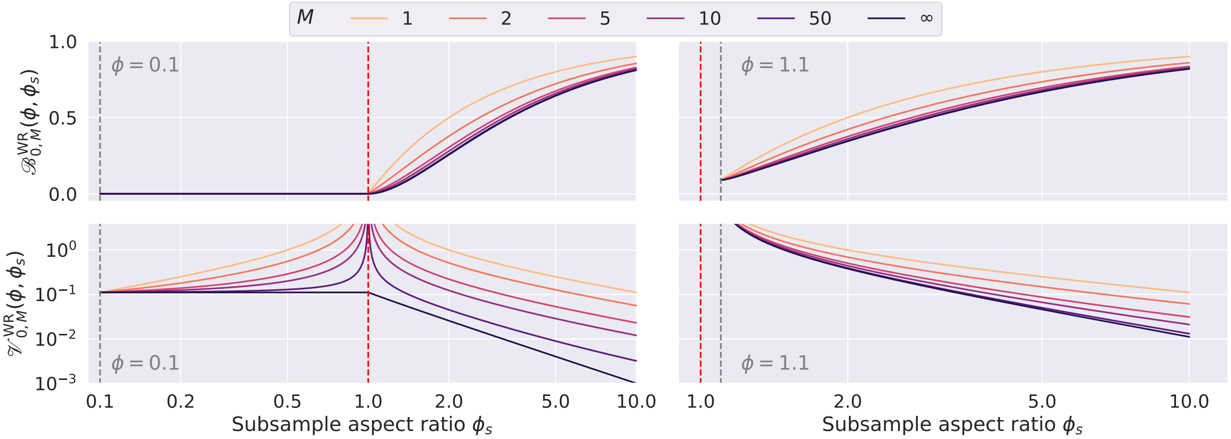


Figure D.5: Asymptotic bias and variance curves in (4.26) for subagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathsf{sub}}_{0,M}(\phi, \phi_s)$ are shown on a log-10 scale.

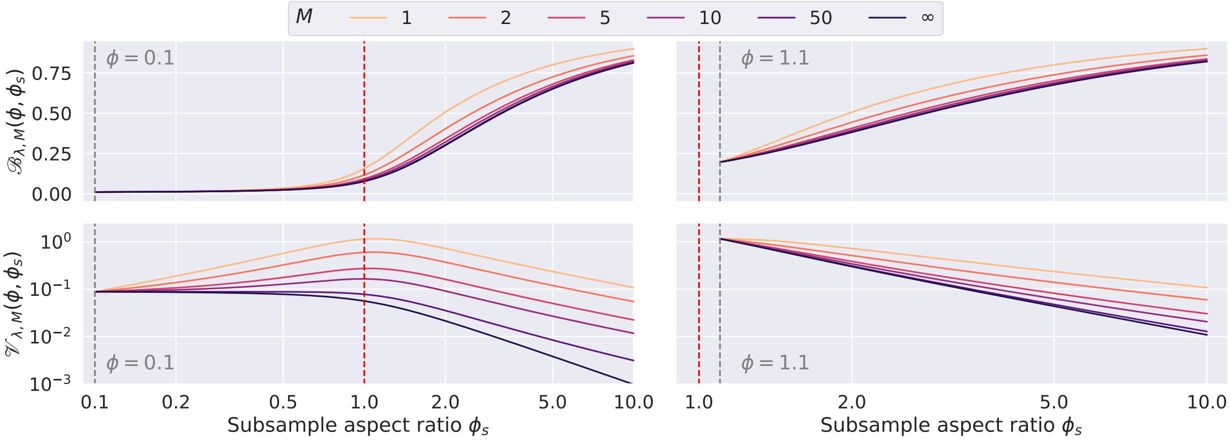


Figure D.6: Asymptotic bias and variance curves in (4.26) for subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathsf{sub}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

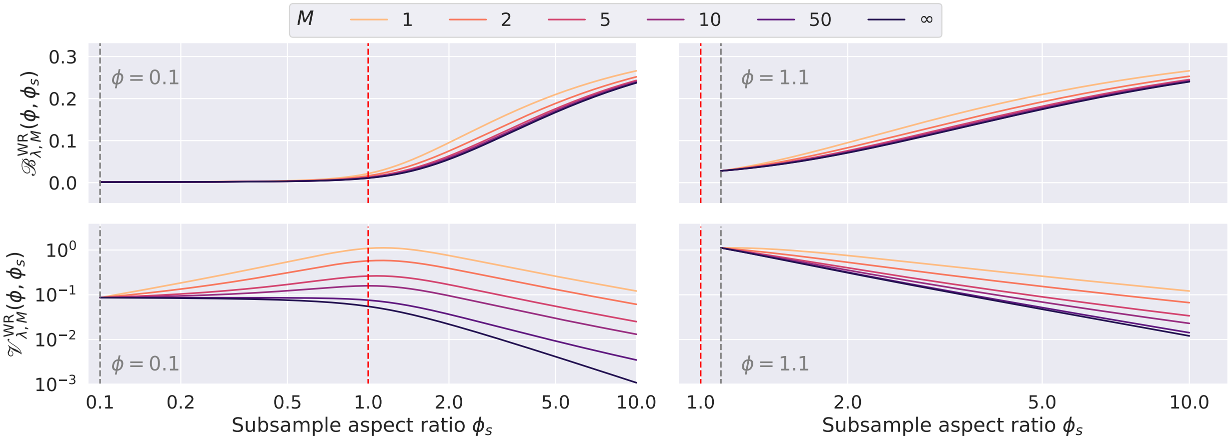


Figure D.7: Asymptotic bias and variance curves in (4.26) for subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

D.10.1.3 Bias-variance curves for subagged ridge predictors with varying λ $(M = 1)$

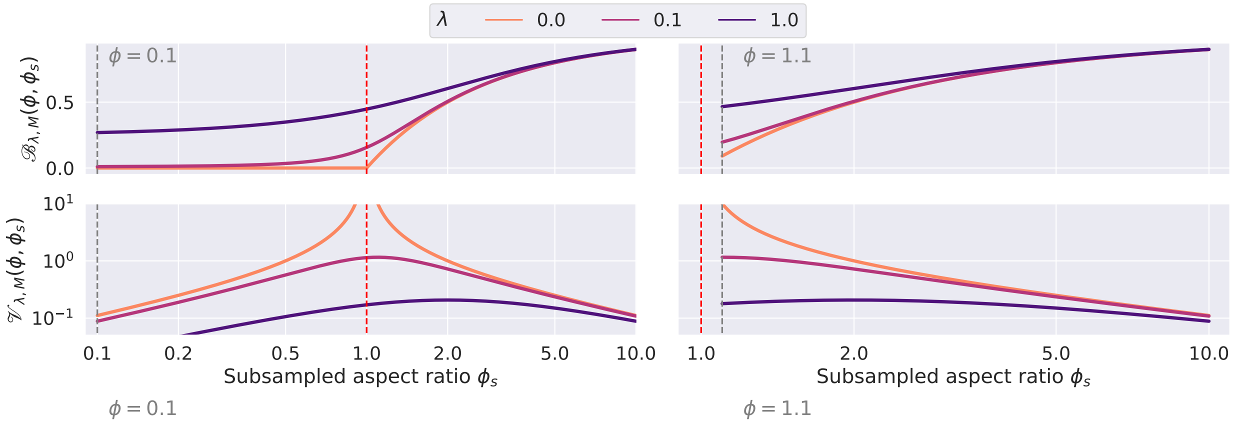


Figure D.8: Asymptotic bias and variance curves in (4.26) for subagged ridge and ridgeless predictors with number of bags $M = 1$, under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying regularization parameter λ. The left and the right panels correspond to the cases when $p < n$ $(\phi = 0.1)$ and $p > n$ $(\phi = 1.1)$, respectively. The values of $\mathscr{V}_{0,M}^{\mathsf{sub}}(\phi, \phi_s)$ are shown in log-10 scale.

D.10.1.4 Bias-variance curves for subagged ridge predictors with varying λ $(M = \infty)$

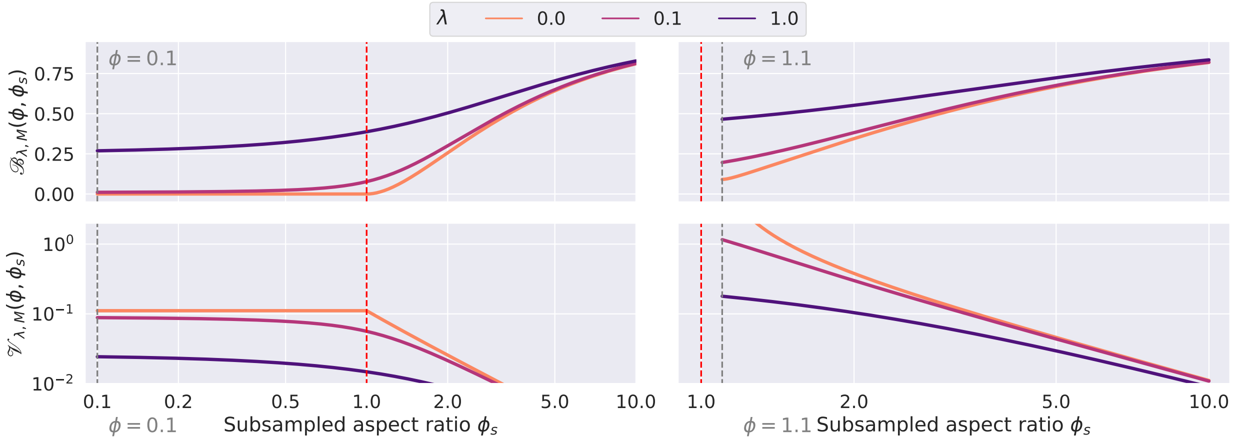


Figure D.9: Asymptotic bias and variance curves in (4.26) for subagged ridge and ridgeless predictors with number of bags $M = \infty$, under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying regularization parameter λ. The left and the right panels correspond to the cases when $p < n$ $(\phi = 0.1)$ and $p > n$ $(\phi = 1.1)$, respectively. The values of $\mathscr{V}_{0,M}^{\mathsf{sub}}(\phi, \phi_s)$ are shown in log-10 scale.

D.10.2 Additional illustrations for Theorem 4.4.6

D.10.2.1 Prediction risk curves for splagged ridgeless and ridge predictors with varying M

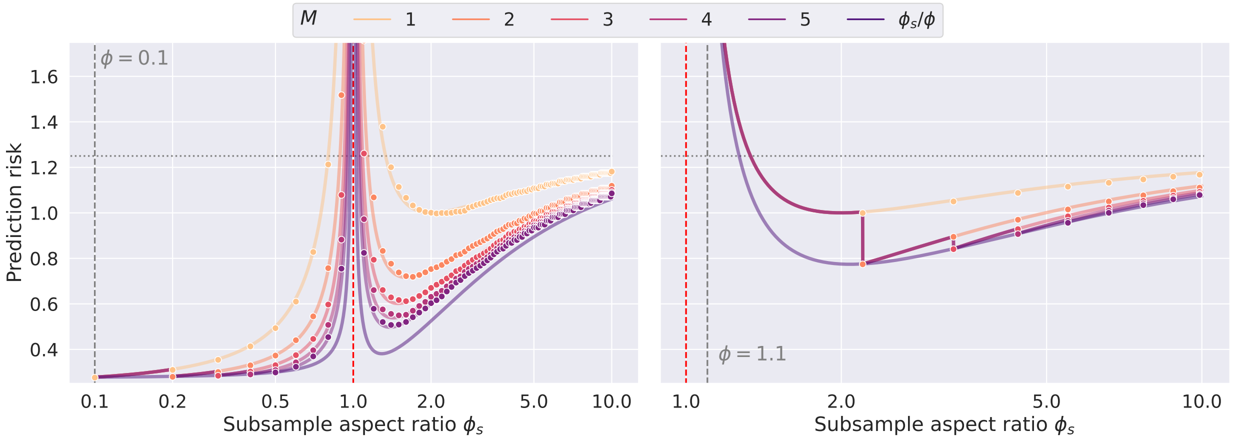


Figure D.10: Asymptotic prediction risk curves in (4.32) for splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$.

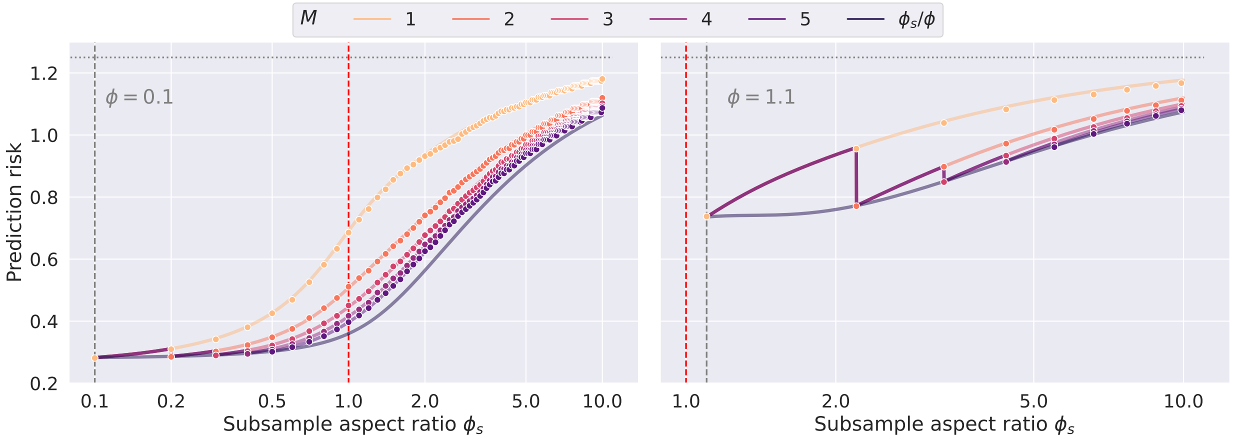


Figure D.11: Asymptotic prediction risk curves in (4.32) for splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$.

D.10.2.2 Bias-variance curves for ridgeless and ridge predictors with varying M

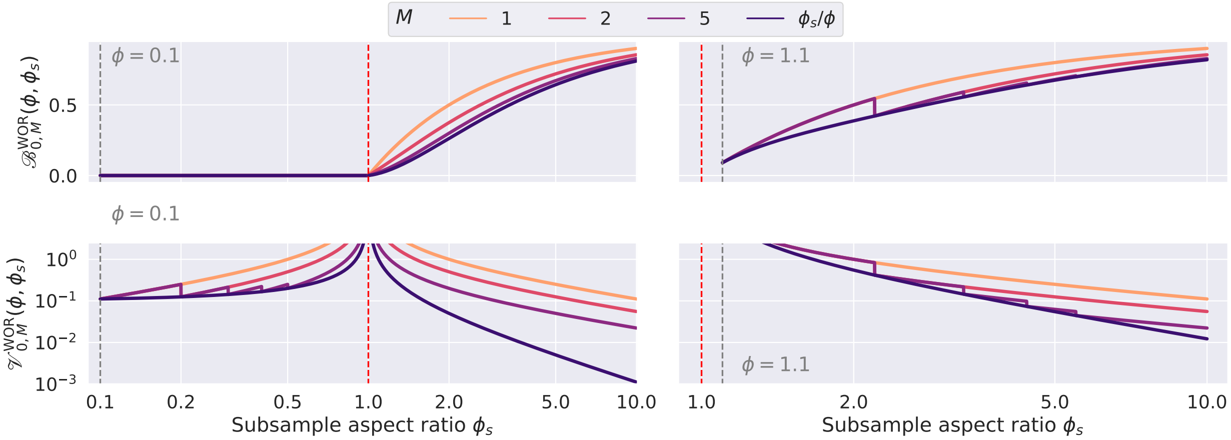


Figure D.12: Asymptotic bias and variance curves in (4.26) for splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{spl}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

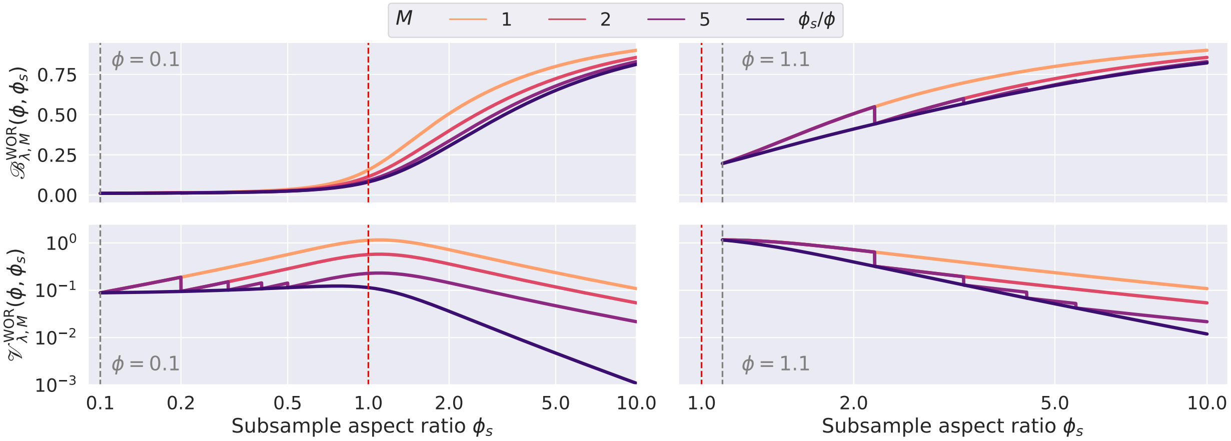


Figure D.13: Asymptotic bias and variance curves in (4.26) for splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ are shown in log-10 scale.

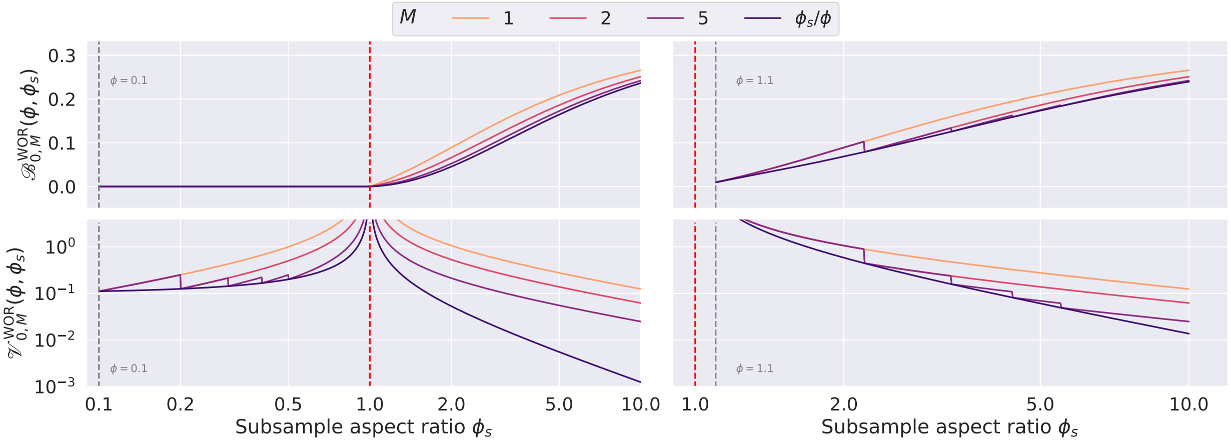


Figure D.14: Asymptotic bias and variance curves in (4.26) for splagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\texttt{spl}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

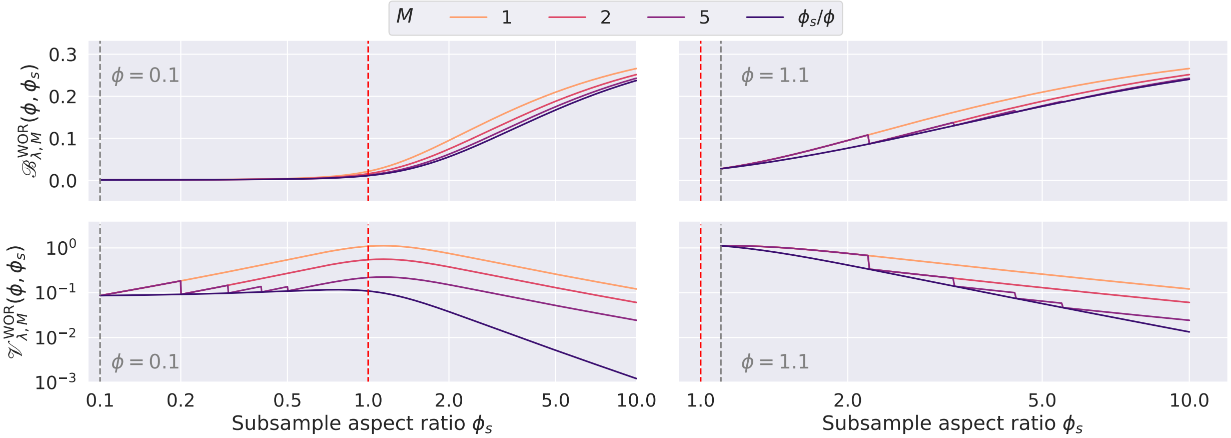


Figure D.15: Asymptotic bias and variance curves in (4.26) for splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\texttt{spl}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

D.10.3 Additional illustrations for Theorem 4.5.1

D.10.3.1 Risk monotonization for subagged ridgeless and ridge predictors

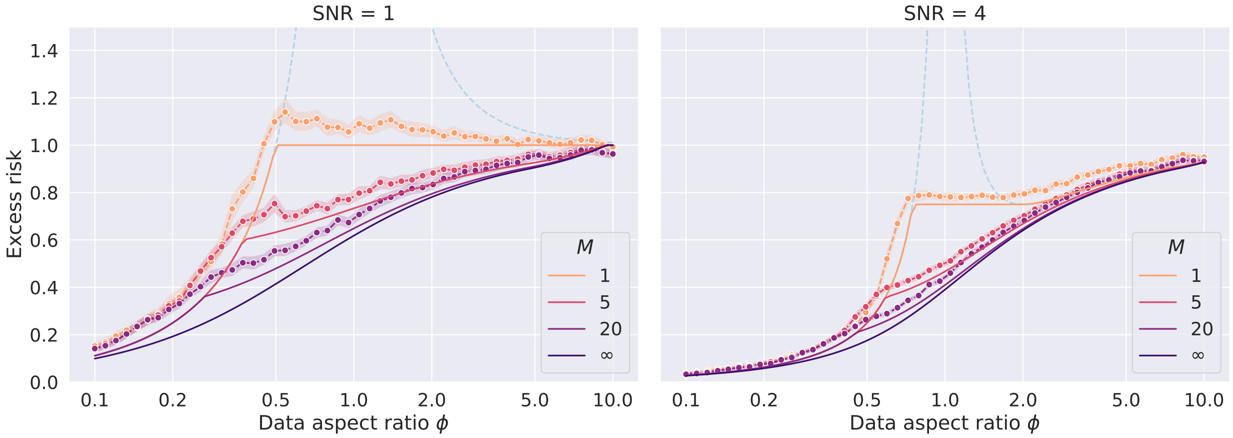


Figure D.16: Asymptotic excess risk curves for cross-validated subagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M with replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 4 respectively. The null risk is marked as a dotted line, and risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\mathrm{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

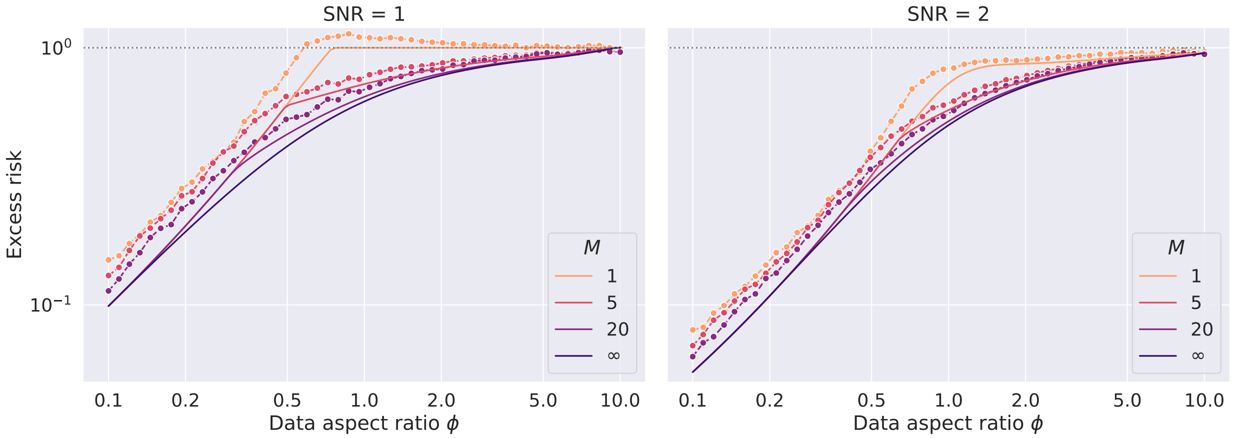


Figure D.17: Asymptotic prediction risk curves for cross-validated subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M with replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 2 respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$, $n_{\mathrm{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

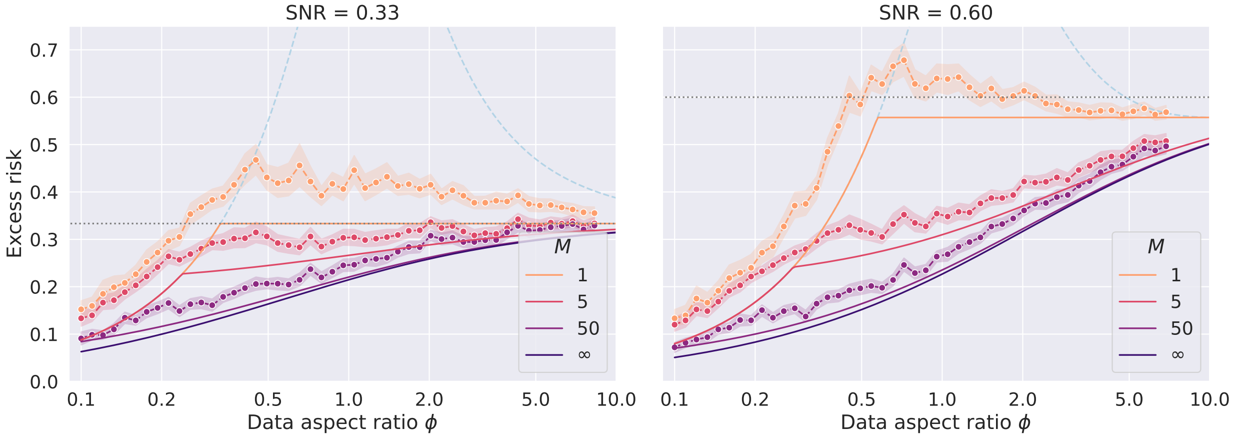


Figure D.18: Asymptotic excess risk curves for cross-validated subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\sigma^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $\texttt{SNR} = 0.33$ ($\rho_{\text{ar1}} = 0.25$) and 0.6 ($\rho_{\text{ar1}} = 0.5$) respectively. The excess null risk is marked as a dotted line, and risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

D.10.3.2 Risk monotonization for splagged ridgeless and ridge predictors

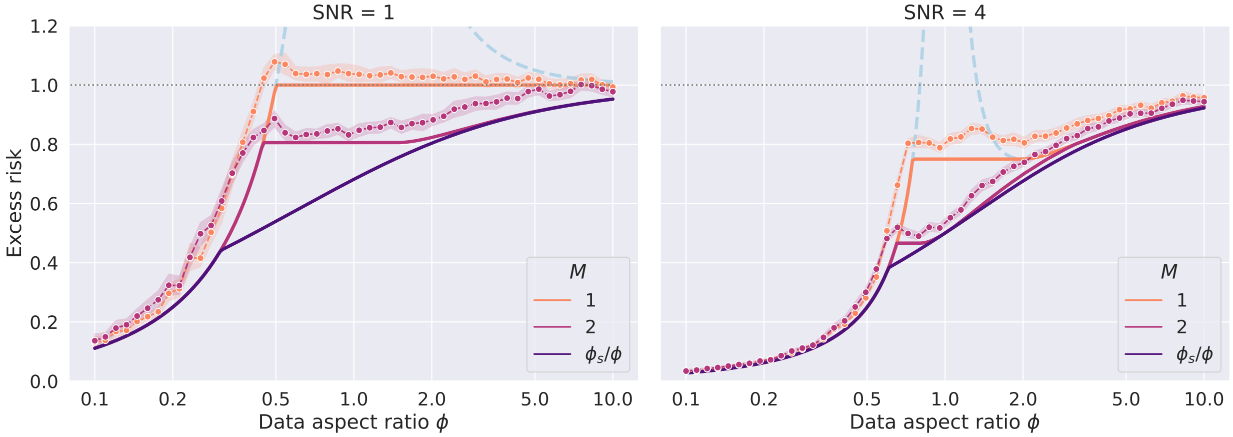


Figure D.19: Asymptotic excess risk curves for cross-validated splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M without replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 4 respectively. The null risk is marked as a dotted line, and risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

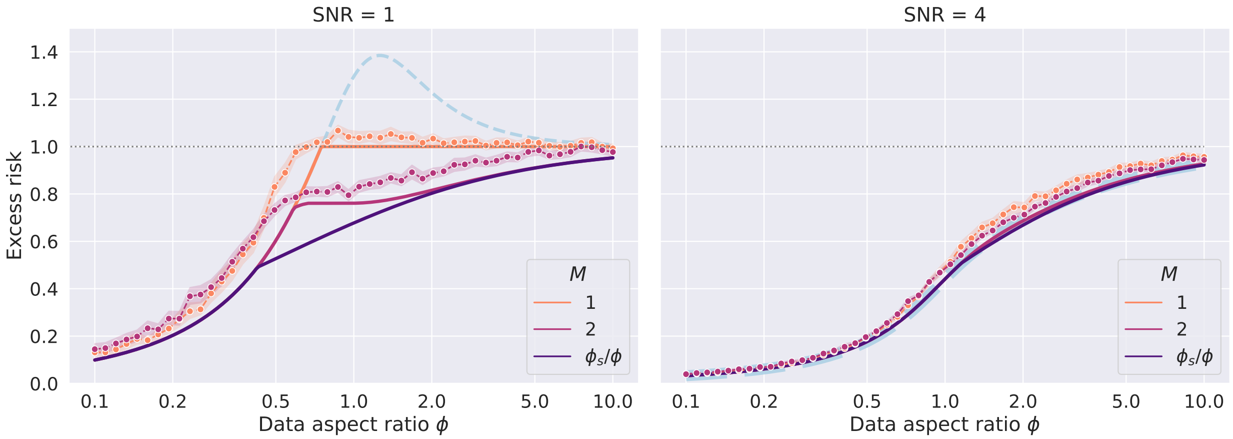


Figure D.20: Asymptotic prediction risk curves for cross-validated splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M without replacement. The left and the right panels correspond to the cases when SNR $= 1$ and 4 respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

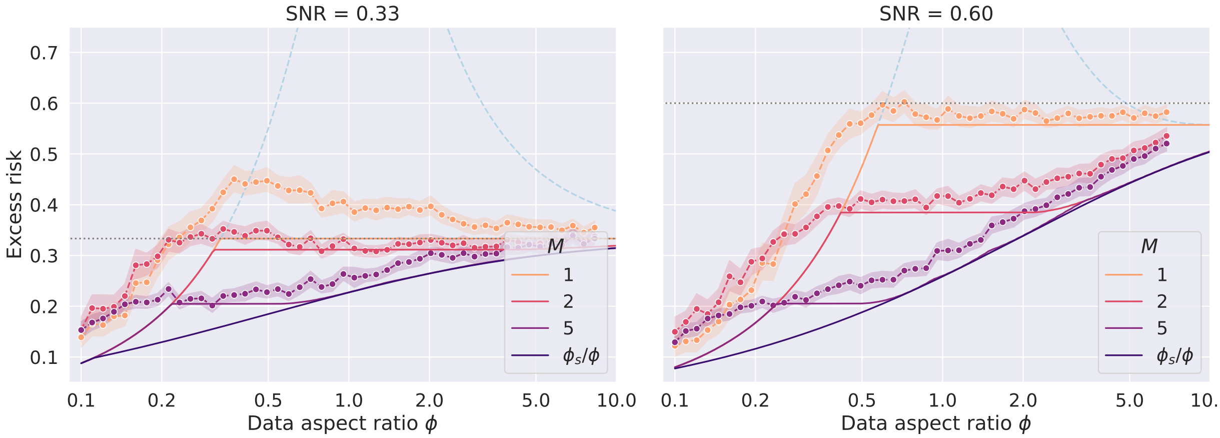


Figure D.21: Asymptotic excess risk curves for cross-validated splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\sigma^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when SNR $= 0.33$ ($\rho_{\text{ar1}} = 0.25$) and 0.6 ($\rho_{\text{ar1}} = 0.5$) respectively. The excess null risk is marked as a dotted line, and risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

D.10.4 Additional illustrations in Section 4.6

D.10.4.1 Subagging with replacement and splagging without replacement

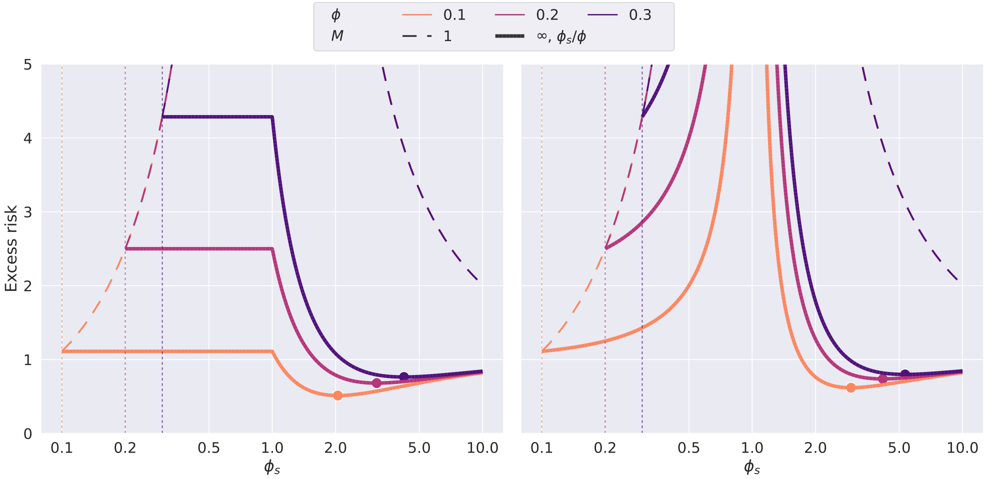


Figure D.22: Asymptotic excess risk (the difference between the prediction risk and the noise level σ^2) curves of bagged ridgeless predictors ($\lambda = 0$) for subagging (left panel) and splagging (right panel), under model (M-ISO-LI) when $\rho^2 = 1$ and $\mathtt{SNR} = 0.1$, for varying ϕ ($p < n$), bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The solid lines represent the optimal risks with respect to M for either with replacement ($M = \infty$) or without replacement ($M = \phi_s/\phi$); the dashed lines represent the risks for $M = 1$; the dotted lines indicates the aspect ratio ϕ of the full dataset; the solid dots represent the optimal risk with respect to both M and ϕ_s.

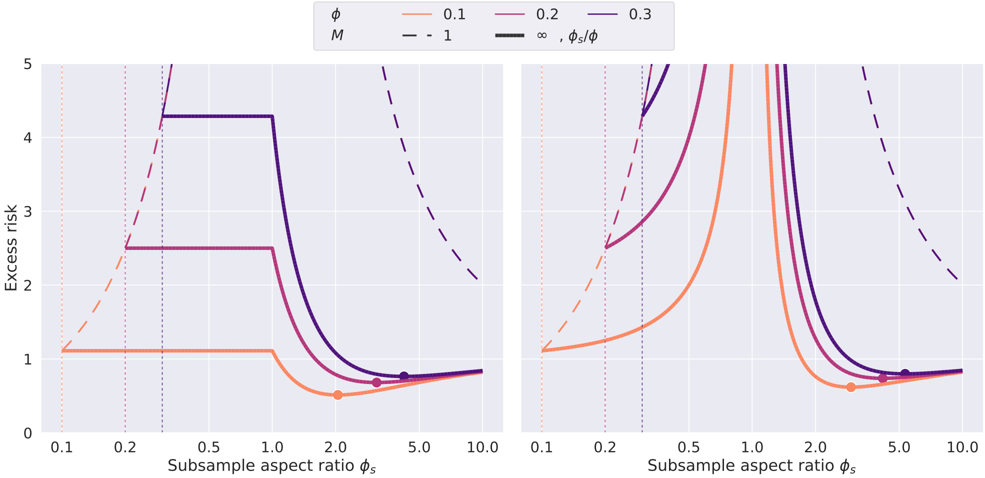


Figure D.23: Asymptotic excess risk (the difference between the prediction risk and the noise level σ^2) curves of bagged ridgeless predictors ($\lambda = 0$) for subagging (left panel) and splagging (right panel), under model (M-ISO-LI) when $\rho^2 = 1$ and $\mathtt{SNR} = 0.5$, for varying ϕ ($p \geq n$), bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The solid lines represent the optimal risks with respect to M for either with replacement ($M = \infty$) or without replacement ($M = \phi_s/\phi$); the dashed lines represent the risks for $M = 1$; the dotted lines indicates the aspect ratio ϕ of the full dataset; the solid dots represent the optimal risk with respect to both M and ϕ_s.

References

Adamczak, R., Litvak, A., Pajor, A., and Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561.

Adlam, B. and Pennington, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Austern, M. and Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111.*

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer.

Bai, Z.-D. and Silverstein, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345.

Banerjee, M., Durot, C., and Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720–757.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177.*

Bayati, M., Lelarge, M., and Montanari, A. (2015). Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822.

Bayati, M. and Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *arXiv preprint arXiv:2007.12671.*

Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. (2017). On optimal generalizability in parametric learning. In *Advances in Neural Information Processing Systems.*

Belkin, M. (2021). Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019a). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180.

Belkin, M., Hsu, D. J., and Mitra, P. (2018a). Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31.

Belkin, M., Ma, S., and Mandal, S. (2018b). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2019b). Does data interpolation contradict statistical optimality? In *International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR.

Bhatia, R. (1997). *Matrix Analysis.* Springer Graduate Texts in Mathematics.

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. (2016). On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1):459–552.

Bogachev, V. I. (2007). *Measure Theory.* Springer. First edition.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics*, 30(4):927–961.

Buja, A. and Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica*, pages 323–351.

Burkholder, D. L. (1973). Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on Information Theory*, 52(12):5406–5425.

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185.

Celentano, M., Montanari, A., and Wei, Y. (2020). The lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716.*

Celisse, A. and Guedj, B. (2016). Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412.*

Chaudhuri, A. (2014). *Modern Survey Sampling.* CRC Press. First edition.

Chen, L., Min, Y., Belkin, M., and Karbasi, A. (2020). Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036.*

Chen, W.-K. and Lam, W.-K. (2021). Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1–44.

Couillet, R. and Debbah, M. (2011). *Random Matrix Methods for Wireless Communications.* Cambridge University Press.

Couillet, R. and Hachem, W. (2014). Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403.

Craven, P. and Wahba, G. (1979). Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403.

Dar, Y., Muthukumar, V., and Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. OUP Oxford.

Deng, Z., Kammoun, A., and Thrampoulidis, C. (2019). A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*.

Deng, Z., Kammoun, A., and Thrampoulidis, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495.

Derezinski, M., Liang, F. T., and Mahoney, M. W. (2020). Exact expressions for double descent and implicit regularization via surrogate random design. *Advances in neural information processing systems*, 33:5152–5164.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive Approximation*, volume 303. Springer Science & Business Media.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37.

Dobriban, E. and Sheng, Y. (2020). Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52.

Dobriban, E. and Sheng, Y. (2021). Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943.

Dobriban, E. and Wager, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279.

Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Donoho, D. L., Maleki, A., and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, 2(2):131–154.

Duin, R. P. (1995). Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964.

Durrett, R. (2019). *Probability: Theory and Examples*, volume 49. Cambridge university press.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association*, 81(394):461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632.

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.

Erdos, L. and Yau, H.-T. (2017). *A Dynamical Approach to Random Matrix Theory.* Courant Lecture Notes in Mathematics.

Fan, J., Li, Q., and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

Frei, S., Chatterji, N. S., and Bartlett, P. L. (2022). Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. *arXiv preprint arXiv:2202.05928*.

Friedman, J. H. and Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. (2019). Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019). A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Greene, E. and Wellner, J. A. (2017). Exponential bounds for the hypergeometric distribution. *Bernoulli*, 23(3):1911.

Grenander, U. and Szegö, G. (1958). *Toeplitz Forms and Their Applications.* University of California Press. First edition.

Gribkova, N. V. (2020). Bounds for absolute moments of order statistics. In *Exploring Stochastic Laws*, pages 129–134. De Gruyter.

Gut, A. (2005). *Probability: A Graduate Course.* Springer, New York.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-free Theory of Nonparametric Regression.* Springer Science & Business Media.

Hachem, W., Loubaton, P., and Najim, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930.

Hall, P. and Samworth, R. J. (2005). Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):363–379.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer Series in Statistics. Second edition.

Hiriart-Urruty, J.-B. and Martınez-Legaz, J.-E. (2003). New formulas for the Legendre–Fenchel transform. *Journal of mathematical analysis and applications*, 288(2):544–555.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30.

Hoerl, A. E. and Kennard, R. W. (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Hoerl, A. E. and Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. In *Conference on Learning Theory.*

Hu, H. and Lu, Y. M. (2020). Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669.*

Janson, L., Fithian, W., and Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485.

Kale, S., Kumar, R., and Vassilvitskii, S. (2011). Cross-validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science.*

Karoui, N. E. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445.*

Karoui, N. E. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175.

Karoui, N. E. and Kösters, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404.*

Kaufman, S. and Rosset, S. (2014). When does more regularization imply fewer degrees of freedom? sufficient conditions and counterexamples. *Biometrika*, 101(4):771–784.

Kini, G. R. and Thrampoulidis, C. (2020). Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532. IEEE.

Knowles, A. and Yin, J. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352.

Kobak, D., Lomond, J., and Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21(169):1–16.

Koenker, R. and Bassett Jr., G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. (2013). Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning.*

Latała, R. (1999). On the equivalence between geometric and arithmetic means for log-concave measures. *Convex geometric analysis*, 34:123–127.

Lecué, G. and Mendelson, S. (2012). General nonexact oracle inequalities for classes with a subexponential envelope. *The Annals of Statistics*, 40(2):832–860.

LeCun, Y., Kanter, I., and Solla, S. (1990). Second order properties of error surfaces: Learning time and generalization. *Advances in neural information processing systems*.

Ledoit, O. and Péché, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264.

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Routledge. First edition.

LeJeune, D., Javadi, H., and Baraniuk, R. (2020). The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, pages 1352–1377.

Li, K.-C. (1986). Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112.

Li, K.-C. (1987). Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347.

Liang, T. and Sur, P. (2020a). A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*.

Liang, T. and Sur, P. (2020b). A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*.

Liu, F., Liao, Z., and Suykens, J. (2021). Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*.

Liu, S. and Dobriban, E. (2019). Ridge regression: Structure, cross-validation, and sketching. *arXiv preprint arXiv:1910.02373*.

Liu, Y., Jiang, S., and Liao, S. (2014). Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, pages 324–332. PMLR.

Loeve, M. (2017). *Probability Theory*. Courier Dover Publications.

Loog, M., Viering, T., Mey, A., Krijthe, J. H., and Tax, D. M. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626.

Lopes, M., Jacob, L., and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Mei, S. and Montanari, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766.

Meijer, R. J. and Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155.

Mendelson, S. and Zhivotovskiy, N. (2020). Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648–1664.

Meyer, Jr., C. D. (1973). Generalized inversion of modified matrices. *SIAM Journal on Applied Mathematics*, 24(3):315–323.

Mhammedi, Z. (2021). Risk monotonicity in statistical learning. *Advances in Neural Information Processing Systems*.

Minsker, S. (2015). Geometric median and robust estimation in banach spaces. *Bernoulli*, 21(4):2308–2335.

Minsker, S. (2018). Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903.

Minsker, S. and Wei, X. (2020). Robust modifications of U-statistics and applications to covariance estimation problems. *Bernoulli*, 26(1):694–727.

Miolane, L. and Montanari, A. (2021). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335.

Mitra, P. P. (2019). Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*.

Montanari, A. and Nguyen, P.-M. (2017). Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2019a). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2019b). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*.

Mücke, N., Reiss, E., Rungenhagen, J., and Klein, M. (2022). Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*, pages 10322–10350. PMLR.

Munkres, J. R. (2000). *Topology*. Pearson Prentice Hall. Second Edition.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

Nakkiran, P. (2019). More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*.

Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. (2021). Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*.

Nayar, P. and Oleszkiewicz, K. (2012). Khinchine type inequalities with optimal constants via ultra log-concavity. *Positivity*, 16(2):359–371.

Neyshabur, B., Tomioka, R., and Srebro, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*.

Obuchi, T. and Kabashima, Y. (2016). Cross validation in LASSO and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment.*

Opper, M. and Kinzel, W. (1996). Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer.

Patil, P., Kuchibhotla, A. K., Wei, Y., and Rinaldo, A. (2022a). Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937.*

Patil, P., Rinaldo, A., and Tibshirani, R. (2022b). Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 6087–6120. PMLR.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. J. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR.

Paul, D. and Silverstein, J. W. (2009). No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57.

Pedersen, G. K. (2012). *Analysis Now.* Springer Graduate Texts in Mathematics.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Pugh, C. C. (2002). *Real Mathematical Analysis.* Springer Undergraduate Texts in Mathematics.

Rad, K. R. and Maleki, A. (2020). A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996.

Rad, K. R., Zhou, W., and Maleki, A. (2020). Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. *arXiv preprint arXiv:2003.01770.*

Richards, D., Mourtada, J., and Rosasco, L. (2020). Asymptotics of ridge(less) regression under general source condition. *arXiv preprint arXiv:2006.06386.*

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis.* Springer Series of Comprehensive Studies in Mathematics.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404.

Rosset, S. and Tibshirani, R. J. (2019). From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association.*

Royden, H. L. (1988). *Real Analysis.* Macmillan New York. Third Edition.

Rubio, F. and Mestre, X. (2011). Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602.

Rudin, W. (1976). *Principles of Mathematical Analysis.* McGraw-Hill New York.

Samworth, R. J. (2012). Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763.

Schapire, R. E. and Freund, Y. (2013). *Boosting: Foundations and Algorithms.* MIT Press.

Sen, P. K. (1970). The Hájek-Rényi inequality for sampling from a finite population. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 181–188.

Serdobolskii, A. V. (1983). On minimal error probability in discriminant analysis. *Reports of the Academy of Sciences of the USSR*, 270:1066–1070.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, volume 162. Wiley Series in Probability and Statistics.

Shiryaev, A. N. (2016). *Probability-1*. Springer. Third edition.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339.

Silverstein, J. W. and Choi, S.-I. (1995). Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309.

Srivastava, R., Li, P., and Ruppert, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151.

Steinberger, L. and Leeb, H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*.

Steinberger, L. and Leeb, H. (2018). Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412*.

Stephenson, W. and Broderick, T. (2020a). Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*.

Stephenson, W. and Broderick, T. (2020b). Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2424–2434. PMLR.

Stojnic, M. (2013). A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–133.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14:323–48.

Sur, P., Chen, Y., and Candès, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1):487–558.

Tao, T. (2010). *An Epsilon of Room, I: Real Analysis*, volume 1. American Mathematical Soc.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628.

Thrampoulidis, C., Oymak, S., and Hassibi, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR.

Tibshirani, R. J. (2015). Degrees of freedom and model search. *Statistica Sinica*, pages 1265–1296.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Doklady Akademii Nauk SSSR*, volume 39, pages 195–198.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk SSSR*, volume 151, pages 501–504.

Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on pattern analysis and machine intelligence*, pages 306–307.

Tsigler, A. and Bartlett, P. L. (2020). Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press.

Van der Vaart, A. W., Dudoit, S., and van der Laan, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Viering, T., Mey, A., and Loog, M. (2019). Open problem: Monotonicity of learning. In *Conference on Learning Theory*, pages 3198–3201.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation Theory III*.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. (2018a). Approximate leave-one-out for fast parameter tuning in high dimensions. *arXiv preprint arXiv:1807.02694*.

Wang, S., Zhou, W., Maleki, A., Lu, H., and Mirrokni, V. (2018b). Approximate leave-one-out for high-dimensional non-differentiable learning problems. *arXiv preprint arXiv:1810.02716*.

Warsaw (2003). Notes on isotropic convex bodies. http://users.uoa.gr/~apgiannop/isotropic-bodies.pdf. [Online; accessed 2022-05-24].

Wellner, J. and van der Vaart, A. (2013). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Series in Statistics.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR.

Wu, D. and Xu, J. (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*.

Wu, Y.-S., Masegosa, A., Lorenzen, S., Igel, C., and Seldin, Y. (2021). Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote. *Advances in Neural Information Processing Systems*.

Xing, Y., Song, Q., and Cheng, G. (2018). Statistical optimality of interpolated nearest neighbor algorithms. *arXiv preprint arXiv:1810.02814*.

Xing, Y., Song, Q., and Cheng, G. (2022). Benefit of interpolation in nearest neighbor algorithms. *arXiv preprint arXiv:2202.11817*.

Xu, J., Maleki, A., and Rad, K. R. (2019). Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*.

Xu, J., Maleki, A., Rad, K. R., and Hsu, D. (2021). Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030.

Yang, Y. (2007). Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192.