

Advancing Bandit Optimization for Generalized Structures: Algorithms and Theory

Yusha Liu

August 2025
CMU-ML-25-113

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee

Aarti Singh, Chair
Aaditya Ramdas
Barnabás Póczos
Akshay Krishnamurthy (Microsoft Research)

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

Copyright © 2025 Yusha Liu

This research was sponsored by: Defense Advanced Research Projects Agency award FA870215D0002; National Institute of Food and Agriculture award 2021-67021-35329; National Science Foundation award CCF1763734; U.S. Army award W911NF2020175; Lockheed Martin Corp. grants PRPS001 and MRA19001RPS005; and Simons Foundation grant 888970.

Keywords: Bandit Optimization, Continuum-Armed Bandit, Sequential Decision-Making, Nonparametric Statistics, Hölder Space, Reproducing Kernel Hilbert Space, M -estimation, Influence Function

In loving memory of my grandmother

Abstract

Sequential decision-making under uncertainty is an important part of machine learning, where optimal algorithms are able to make intelligent decisions based on data acquired in a sequential feedback loop. Optimization with bandit feedback, where only noisy feedback corresponding to the decision choice made (rather than all possible choices) is available, is a pivotal component in the sequential decision-making under uncertainty paradigm. Bandit optimization problems motivate design of algorithms that can handle partial feedback, sequential data that are not independent and identically distributed (i.i.d.), and achieve good performances on-the-fly. This framework has shown significant impact in interactive applications including hyperparameter tuning, recommendation systems, adaptive control trials, and other scientific disciplines. However, difficulty of analyses in bandits often limits the broadness of problem settings for which theoretical understanding and algorithms are developed, especially compared to the variety of problem structures present across these domains. This thesis identifies and addresses open problems in bandit optimization when leveraging general problem structures to design optimal algorithms. This thesis presents results that broaden the bandit framework beyond traditional assumptions by developing novel theoretical guarantees and algorithms that work in more generalized settings, expanding the applicability of bandit algorithms and relevance of theoretical results.

The first part of this thesis focuses on addressing inefficiencies of existing algorithms and limitations in the problem assumptions. Firstly, we present results that unify bandit optimization across reward functions with different levels of smoothness by filling the gap between the non-parametric Lipschitz continuous functions and the parametric linear functions. We present a novel algorithmic framework with optimal statistical guarantees for functions with the general notion of Hölder smoothness which encapsulates both Lipschitz and parametric functions as special cases. Secondly, we investigate the sample complexity without the knowledge of the reward function smoothness. We develop matching upper and lower bounds on sample complexity of algorithms that are adaptive to the smoothness of the reward function, and conclude with a comprehensive overview across multiple function spaces including reproducing kernel Hilbert spaces and Hölder spaces. In the second part of this thesis, we take inspiration from classical results on M -estimators based on influence (score) functions and show that the same idea can be used to derive a unifying analytical framework that naturally fits M -estimators with optimism-in-face-of-uncertainty type of bandit algorithms in a “plug-and-play” fashion.

Together, this thesis advances optimization with bandit feedback by enabling novel algorithms and statistical guarantees pertaining to richer and more general landscapes.

Acknowledgments

My deepest gratitude goes to my advisor, Aarti. I also extend my sincere thanks to my brilliant thesis committee – Aaditya, Barnabás and Akshay. I am grateful for my mentors and collaborators along this journey, the people in the Machine Learning Department at CMU, and last but not least: my dearest friends and family.

Contents

1	Introduction	1
1.1	Summary of Contributions	2
2	Generalization to Hölder Space	4
2.1	Introduction	4
2.1.1	Related Works	5
2.1.2	Our Contributions	6
2.2	Problem Setting	7
2.3	The Meta-Algorithm	8
2.3.1	Algorithm Overview	8
2.3.2	The Misspecified Linear Bandit Algorithm	9
2.3.3	The UCB-Meta-Algorithm	10
2.3.4	The Corral-Meta-Algorithm	12
2.3.5	Discussion	12
2.3.6	Comparison with Existing Lower Bound	12
2.4	Adaptation to Unknown Smoothness	13
2.4.1	Corral Applied with Meta-Algorithms	13
2.4.2	Comparison with Existing Lower Bound for Adaptation	14
2.5	Discussion	15
2.6	Proofs of Results	16
2.6.1	Proof of Lemma 2	16
2.6.2	Proof of Theorem 3	17

2.6.3	Proof of Theorem 4	22
2.6.4	Proof of Theorem 6	26
2.6.5	Proof of Lemma 7	27
2.6.6	Proof of Theorem 8	27
2.7	The Doubling Procedure for Algorithm 2	28
3	Unknown Kernel Regularity in Kernelised Bandits	30
3.1	Introduction	30
3.2	Related Work	33
3.3	Problem Setting	34
3.4	Main Result: Adaptivity Lower Bound	35
3.4.1	Norm Equivalency Between RKHS and Sobolev Space	35
3.4.2	Lower Bound on Adaptivity to Kernel Regularity	36
3.5	Upper Bounds of Adaptive Algorithms	39
3.5.1	Overview: Non-adaptive Minimax Algorithms	39
3.5.2	CORRAL as Adaptive Algorithm	40
3.5.3	RBBE as Adaptive Algorithm	41
3.6	Connection with Adaptivity to Hölder Exponents	42
3.7	Discussion	42
3.8	Auxiliary Derivations	43
3.8.1	Proof of Norm Equivalency Between RKHS Norm and Sobolev Seminorm	43
3.8.2	The Full Statement of Theorem 14	43
3.8.3	Adaptivity Upper Bound of RBBE	44
3.9	Proofs of Results	45
3.9.1	Proof of Theorem 20	45
3.9.2	Proof of Corollary 16	52
3.9.3	Proof of Theorem 17	52
3.9.4	Proof of Theorem 18	53
3.9.5	Proof of Theorem 21	55

4	A General Framework for M-estimators in Bandits	58
4.1	Introduction	58
4.2	Problem Setting	59
4.2.1	Preliminary: Influence Function in Risk Minimization Problems	60
4.3	Methodology: The Plug-and-Play Framework	61
4.3.1	Overview	61
4.3.2	Model Assumptions	62
4.3.3	Step I: Deviation of Empirical Risk Minimizers from Influence Function	64
4.3.4	Step II: Bounding Deviation of Influence Function with Sub- ψ Condition	65
4.3.5	Summary	67
4.4	Case Study: Linear and Kernelised Model	67
4.4.1	Ridge Linear Least-Square	68
4.4.2	Kernelised Model	69
4.5	Case Study: Generalized Linear Models	72
4.5.1	Connection with Deviation of Score Function	73
4.5.2	Bounding the Deviation of the Score Function	74
4.5.3	Conclusion	75
4.6	Case Study: Heavy-tail Noise	75
4.6.1	Data-dependent Loss Function Parameter Satisfies Design Principle 1	76
4.6.2	Bounding Martingale Vector Process with Heavy-Tail Noise	77
4.7	Discussion	78
4.8	Proof of Theorem 27	79
4.9	Auxiliary Derivations	81
4.9.1	Comparison of Sub- ψ_N Concentration Results	81
4.9.2	GP-UCB Uses Confidence Ellipsoid in RKHS	82
4.9.3	Auxiliary Derivations for Section 4.6.2	83
4.9.4	Proof of Lemma 29	83
5	Conclusion and Future Directions	85

Bibliography

88

Chapter 1

Introduction

Sequential decision-making under uncertainty is an essential framework in machine learning. This broad framework encompasses many important areas of study, including bandit optimization, active learning and reinforcement learning, that have received long and evolving attention for their statistical importance and wide range of applications. In passive settings, algorithms typically receive data in one single batch or are given data generated independent and identically distributed (i.i.d.) from some fixed distributions. In contrast, in sequential settings, the algorithm interacts with the unknown environment sequentially in a feedback loop. In this case, the algorithm's decision at the current round can affect data received in the future, and the data observed by the algorithm are often non-i.i.d. The sequential nature of the problem and the feedback loop motivate design of smarter, more efficient algorithms than those designed for passive settings, that can achieve good performance in fewer tries. Optimal algorithms in the sequential setting need to handle sequentially acquired data and make informed decisions (actions) to optimize their performance. The analysis and development of optimal methods for the sequential decision-making tasks are therefore more challenging and theoretically interesting.

Bandit optimization is a classical sequential decision-making problem. At each round, the algorithm interacts with the environment by selecting an action (arm), then receiving feedback (reward) pertaining to that action. The bandit optimization framework is equipped with the following characteristics,

1. Noisy observations: to better align with practical situations, the feedback is typically modeled as a *noisy* instantiation of an underlying true reward (function), making the optimization more difficult.
2. Partial feedback: only the (noisy) reward of the selected arm is revealed, as oppose to the reward of the full set of arms. This differentiates bandit optimization from certain online optimization settings.
3. Cumulative regret minimization: the algorithm's performance is evaluated by a non-decreasing metric known as the cumulative regret. The goal is to minimize this notion of regret. As a result, the algorithm has to control its performance across all the steps.

Notably, the bandit framework incorporates the interesting challenge of the exploration-exploitation trade-off through cumulative regret minimization (point 3). On one end of the spectrum is exploration,

which means that the algorithm should sample unfamiliar actions to learn about the environment so that it can potentially discover better actions for the future. Exploitation, on the other end of the spectrum, means that the algorithm should take advantage of its estimation of the environment and select already-promising actions to keep its performance good.

As a result, optimal algorithms designed under the bandit framework have the ability to make intelligent decisions for more efficient optimization, striking the right balance between both exploration and exploitation, based on non-i.i.d. data. These factors ensure good behavior for learning and optimizing in unknown environments on-the-fly. Thus, the bandit framework has been used in a wide range of applications for interactive scenarios, such as hyperparameter optimization [Li et al., 2018], recommendation system [Li et al., 2010], economics [Chen et al., 2023], and in other scientific disciplines such as material science and astrophysics.

The other side of the coin is that, designing algorithms with sound theoretical guarantees and understanding the learning complexity in the bandit framework are challenging tasks. The difficulty limits the extent and breadth of relevant theoretical studies. This thesis addresses fundamental open problems in the bandit framework by designing novel algorithms with optimal performance and deriving theoretical guarantees, under more generalized problem structures. More specifically, the contributions in this thesis are summarized below. We consider optimization of reward functions that are smooth functions defined over a continuous action space, unless specified otherwise.

1.1 Summary of Contributions

Part I (Chapter 2 ~ 3)

The first part of this thesis focuses on addressing limitations in the problem assumptions and consequently, inefficiency in existing algorithms designed under the limiting assumptions. Chapter 2 describes our contributions in generalizing bandit optimization across different levels of smoothness of the reward function. Previous works either focus on parametric cases such as linear functions, or Hölder (including Lipschitz) continuity for nonparametric reward functions. However, there are many other types of function in between. Our results in Chapter 2 expand bandit optimization to the general Hölder function space, where functions have higher degrees of smoothness than a Lipschitz function, but are not infinitely-differentiable like a linear function. In other words, the statistical difficulty to optimize on this function space lies between Lipschitz and linear function spaces. In such cases as these, traditional methods developed for linear functions would altogether fail to obtain acceptable, sublinear regret bounds, while methods developed for Lipschitz functions would be inefficient in the sense that they are unable to achieve the best possible regret. To address this problem, we develop algorithms that can fully exploit the broad class of Hölder smoothness and achieve minimax optimality if given the knowledge of the smoothness. We derive bound on the cumulative regret (Section 2.2) to demonstrate the optimality of our algorithms. Chapter 2 effectively closes the gap between the infinitely-differentiable linear functions and the non-differentiable Lipschitz functions and offers a complete view for the varying structures in between.

In bandit optimization literature, including Chapter 2, algorithms are typically designed under the standard assumption that is a priori knowledge of the smoothness of the reward function. This as-

sumption, although crucial for those literature, is not entirely realistic because a practitioner cannot always know the exact smoothness of the underlying reward function. To investigate this limitation, Chapter 3 explores a more challenging setting, where the function smoothness is unknown to the algorithm. Specifically, we ask the following question: can algorithms achieve the same cumulative regret rate as if knowing the smoothness? And if not, what is the best performance they can achieve? To answer this question, we analyze the statistical difficulty of bandit optimization when key smoothness parameters of the reward functions are unknown. We develop upper and lower cumulative regret bounds for adaptive algorithms to fully capture this statistical difficulty. Our main focus is on the reproducing kernel Hilbert space (Section 3.3). We provide results for Sobolev space and Hölder space as well that draw connections between the three function spaces, offering insights on the relationship between types of smoothness and adaptation difficulty in the bandit framework.

Part II (Chapter 4)

The second part of this thesis investigates an alternative yet unifying way to interpret bandit optimization problems. Delving deeper into the mechanisms of bandit algorithms, it is well-known that one prominent way to achieve exploration-exploitation trade-off is through the “optimism in the face of uncertainty” principle. A classical example of algorithms following this principle is upper confidence bound(UCB)-based algorithms [Lai and Robbins, 1985], which maintain an upper (“optimistic”) confidence bound on the uncertain environment and select actions with the highest optimistic reward estimate. Because the analyses of such algorithms are usually strictly in line with, and guided by, the specific reward structure and the estimators used by the algorithm, they are typically developed in a case-by-case fashion, with a specific model and estimator in mind. In Chapter 4, we present a unifying analytical framework for M -estimators (empirical risk minimizers) to be integrated in the bandit setting with UCB-based algorithms. We show that this framework subsumes and ties together a wide range of problem structures. Inspired by the use of influence functions for deviation analyses of M -estimators in both asymptotic and non-asymptotic settings, we extend the analyses to regression setting with non-i.i.d. data, and demonstrate its use for bandit optimization. Coupled with a general statistical tool for time-uniform martingale concentration, we show that there is a natural and a versatile approach to analyze finite-sample time-uniform confidence sequences for many M -estimators used in bandit problems, providing a unifying lens for M -estimators in UCB-based bandit algorithms.

Chapter 2

Bridging the Gap between Lipschitz and Linear: Generalization to Hölder Space

This chapter is based on [Liu et al. \[2021\]](#).

2.1 Introduction

In this and the following chapter, we consider the continuum-armed bandit optimization problem, which can be seen as a problem of black-box optimization of a continuous reward function $f : \mathcal{X} \rightarrow \mathcal{R}$ using active queries. The reward function is assumed to be bounded and defined on a compact d -dimensional domain \mathcal{X} . At each round, the algorithm chooses an action $x_t \in \mathcal{X}$ by leveraging the previously collected data and observes a noisy and zeroth order feedback of the function value $f(x_t)$. In the bandit setting, the goal is to minimize the cumulative regret with respect to global maxima. The bandit framework is different from standard global zeroth order optimization because of its unique exploration-exploitation dilemma. While in zeroth order optimization problems, pure exploration will often suffice since the performance is measured by simple regret (i.e. difference between the optimized function value and true function maxima), in bandit optimization, the queried function values need to be controlled through the entire optimization process to minimize the cumulative regret. Therefore, the algorithms require different and often more careful design.

Most existing works on continuum-armed bandit optimization either assume parametric models such as linear bandits ([Dani et al. \[2008\]](#), [Abbasi-Yadkori et al. \[2011\]](#), [Rusmevichientong and Tsitsiklis \[2010\]](#)) for the reward function, or a black-box model where the reward function is assumed to be α -Hölder continuous (including Lipschitz) with $0 < \alpha \leq 1$ with respect to some known metric ([Kleinberg \[2005\]](#), [Auer et al. \[2007\]](#), [Kleinberg et al. \[2008\]](#), [Bubeck et al. \[2010, 2011\]](#), [Locatelli and Carpentier \[2018\]](#)). The main purpose of this chapter is to extend this assumption to the more general Hölder function space (definition 1) with exponent $\alpha > 1$ and exploit the higher order of function smoothness.

Approaches based on fitting an appropriate function using random samples in bins of a discretization of the domain (i.e., exploration) suffice as optimal for controlling cumulative regret for Hölder continuous reward functions with $\alpha \leq 1$, as well as controlling simple regret of Hölder smooth reward functions with any $\alpha > 0$. In contrast, controlling cumulative regret for Hölder smooth reward functions with $\alpha > 1$ requires finer control in bins over the queried values via a local exploration-exploitation tradeoff. Thus, instead of using a single layer algorithm that randomly samples from selected bins, we propose a class of algorithms that use two layers of bandit algorithms - one multi-armed bandit algorithm operating over the bins, and another set of misspecified linear/polynomial bandit algorithms operating in each bin to govern the local exploration-exploitation tradeoff. We derive regret bounds for this class of two-layer bandit algorithms and show that they match the existing lower bounds apart from log factors.

Additionally, we study the problem of adaptation to smoothness exponent α for a continuous scale of Hölder spaces. Unlike the simple regret minimization setting where this adaptation comes at no cost in terms of the minimax rates, it was shown by [Locatelli and Carpentier \[2018\]](#) that it is generally impossible to achieve minimax adaptation under cumulative regret. We propose a procedure with regret bound that matches the existing adaptive lower bound with only access to the range of the unknown parameter α . We start by describing related works, followed by a summary of our contributions.

2.1.1 Related Works

Continuum-Armed Bandit.

In continuum-armed bandit problems, the domain \mathcal{X} is allowed to be a measurable space, and the set of arms is therefore infinite. Previous works in continuum-armed bandit usually assumes global smoothness ([Kleinberg \[2005\]](#)) of the reward function or local smoothness (e.g. [Auer et al. \[2007\]](#)) around the global maxima. The smoothness condition, in particular, is defined as Lipschitz continuity with respect to some metrics ([Kleinberg \[2005\]](#), [Kleinberg et al. \[2008\]](#)) or dissimilarity functions ([Kleinberg et al. \[2008\]](#), [Bubeck et al. \[2010\]](#)), or α -Hölder continuity with $0 < \alpha \leq 1$ ([Kleinberg \[2005\]](#), [Auer et al. \[2007\]](#)). Worst-case lower bound under the Lipschitz assumption is presented in [Kleinberg et al. \[2008\]](#) and that under the Hölder continuity assumption in [Locatelli and Carpentier \[2018\]](#).

Existing works rarely consider the generalization to Hölder space. Recently [Hu et al. \[2020\]](#) studied contextual bandit with reward functions in Hölder spaces, however, the reward function is assumed to be smooth with respect to the observed contexts and the action set is finite. For non-contextual continuum-armed bandits, [Akhavan et al. \[2020\]](#) focus on the strongly convex subset of functions in Hölder spaces with $\alpha \geq 2$ by using projected gradient-like algorithms. [Grant and Leslie \[2020\]](#) analyze Thompson sampling (TS), a Bayesian method, on Hölder spaces with integer-valued exponents and derive a suboptimal upper bound based on the complexity of the function space¹.

¹They comment that the reason could be either the analysis being suboptimal or the nature of TS. They also derive lower bounds under one-dimension setting, but as we later remark, the same lower bound has already been implied by [Wang et al. \[2018\]](#) under a more general setting.

Adaptivity to Smoothness of the Reward Functions.

An intriguing problem is whether an algorithm that is oblivious to the Hölder exponent α can simultaneously achieve minimax rates for a range of values for α . For non-contextual continuum-armed bandits, this has been discussed only under the Hölder continuous ($\alpha \leq 1$) setting. In particular, [Locatelli and Carpentier \[2018\]](#) state that generally, such minimax adaptation to α is impossible by providing a worst-case lower bound for adaptation between two Hölder-continuous function spaces. Additionally, they propose conditions under which it would become possible. (For the contextual finite-armed bandit studied in [Hu et al. \[2020\]](#), [Gur et al. \[2019\]](#) provide lower bounds with similar rates and the extra conditions as well.) However, it remains unclear that, without the extra conditions, whether an algorithm can achieve the lower bound when adapting to a continuous scale of general Hölder spaces.

Model Selection for Bandits.

Another relevant line of work is more broadly model selection in bandit settings, which we will leverage in bandit optimization of Hölder-smooth functions as well as adaptation to the smoothness. In this problem, given a set of base algorithms on possibly different domains, the learner needs to adapt to the best one in an online fashion. The goal is to achieve cumulative regret comparable to the best base algorithm if it were run solely. [Bubeck et al. \[2011\]](#) study the model selection problem for adapting to the unknown Lipschitz constant of functions. [Foster et al. \[2019\]](#) study adapting to the unknown policy dimension in contextual linear bandits by estimating the gap between two policy classes. [Agarwal et al. \[2016\]](#) develop a general algorithm named Corral for bandit model selection under adversarial feedback. It uses online mirror descent to balance between base algorithms. For stochastic feedback particularly, [Pacchiano et al. \[2020b\]](#) modify the Corral algorithm to relax requirements on base-algorithms and improve the result on some problem instances (including the one in [Foster et al. \[2019\]](#)). Another relevant issue addressed in [Krishnamurthy et al. \[2019\]](#) which study contextual continuum-armed bandits with Lipschitz continuous reward functions, is their use of the original Corral algorithm applied with EXP4 for adaptation to unknown Lipschitz constant. UCB-type algorithm for corraling base-algorithms is used in [Arora et al. \[2020\]](#) under the assumption that the base-algorithms are finite-armed, and only one of them has access to the best arm.

2.1.2 Our Contributions

We study bandit optimization of functions in general Hölder spaces. This chapter furthers the previous works in the following two main aspects:

1. We propose a novel class of two-layer bandit algorithms, where a carefully-chosen Meta-algorithm deploys misspecified bandit algorithms as arms. Our algorithms show explicitly how to exploit higher-order smoothness in achieving optimal exploration-exploitation tradeoff. We derive worst-case regret bound for this algorithm that matches the existing lower bound except for log factors, for functions in Hölder space including when $\alpha > 1$. Our results bridges the gap between Lipschitz smooth bandits where the Hölder exponent is $\alpha = 1$ and infinitely-differentiable problems such as linear bandits where the Hölder exponent is $\alpha = \infty$.

2. We study adaptation to a sequence of Hölder spaces indexed by a continuous but unknown variable of exponent α . We propose a strategy with theoretical guarantee, which uses the bandit model selection algorithm Corral from [Pacchiano et al. \[2020b\]](#) applied with versions of our proposed two-layer algorithms. The derived regret bound is to our knowledge the first result on upper bounds when adapting to a continuous scale of Hölder spaces in continuum-armed bandit optimization.

The rest of this chapter is organized as follows: In section 2 we introduce the problem formulation and assumptions. We present the two-layer Meta-algorithms and theoretical guarantees in section 3. In section 4 we study the adaptation to unknown smoothness and conclude the chapter in section 5 with some open questions.

2.2 Problem Setting

In this chapter, we consider bandit optimization of smooth functions in Hölder space $\sum(\alpha, L)$ with $\alpha > 1$. The Hölder space is defined formally in definition 1. Some works also study benign problem instances with additional “growth” conditions than the smoothness to characterize the difficulty of finding global maxima, for improvements in regret bounds. For example, [Auer et al. \[2007\]](#) use a parameter to model the growth rate of Lebesgue measure of the near-optimal arms set as a function of the threshold. The near-optimality dimension in [Bubeck et al. \[2010\]](#) uses packing number but has similar meaning. We focus solely on worst-case regret to preserve simplicity and leave adaptation to benign cases as a future direction. The performance of the learner is measured by cumulative pseudo-regret as stated below where $x^* \in \arg \max_{x \in \mathcal{X}} f(x)$. Throughout this chapter, we simply refer to the pseudo-regret as regret.

$$R(T) = \sum_{t=1}^T [f(x^*) - f(x_t)]. \quad (2.1)$$

Definition 1 ([Tsybakov \[2008\]](#)). The Hölder space $\sum(\alpha, L)$ on domain $\mathcal{X} \in \mathcal{R}^d$ is defined as the set of functions $f : \mathcal{X} \rightarrow \mathbf{R}$ that are $l = \lfloor \alpha \rfloor$ times differentiable and have continuous derivatives². l is the largest integer that is strictly smaller than α . Define the following notions for a vector $s = (s_1 \dots s_d)$: let $|s| = s_1 + \dots + s_d$, $s! = s_1! \dots s_d!$ and $x^s = x_1^{s_1} \dots x_d^{s_d}$. A function f in $\sum(\alpha, L)$ satisfies the following inequality³ for $\forall x, y \in \mathcal{X}$. Here $D^s = \frac{\partial^{|s|}}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$.

$$D^s f(x) - D^s f(y) \leq L \|x - y\|_\infty^{\alpha - l}, \quad \forall s \text{ s.t. } |s| = l.$$

In particular, a function in $\sum(\alpha, L)$ is close to its Taylor approximation:

$$|f(x) - T_y^l(x)| \leq L \|x - y\|_\infty^\alpha, \quad \forall x, y \in \mathcal{X}.$$

We use T_y^l to denote the l -degree Taylor polynomial around y , $T_y^l(x) = \sum_{|s| \leq l} \frac{(x-y)^s}{s!} D^s f(y)$.

²Only when referring to the order of Hölder smooth functions’ derivatives do we denote $\lfloor \cdot \rfloor$ as the largest integer *strictly* less than input. In other places inside this chapter, it denotes less or equal to input.

³We use l_∞ norm as in some works on adaptive confidence bands and optimization ([Low et al. \[1997\]](#), [Tsybakov \[2008\]](#), [Hoffmann et al. \[2011\]](#), [Wang et al. \[2018\]](#)).

Assumptions

We specify the assumptions that are used throughout this chapter.

- G1.** The input domain \mathcal{X} is a hypercube $[0, 1]^d$. For simplicity assume the reward function is bounded: $\|f\|_\infty \leq 1$.
- G2.** The function f belongs to Hölder space $\Sigma(\alpha, L)$ with some constant $L > 0$ ⁴.
- G3.** The observations are noisy: $y = f(x) + \eta$ where the noise η is drawn from i.i.d zero mean sub-gaussian distribution with parameter σ .

2.3 The Meta-Algorithm

A commonly used method for continuum-armed bandits is fixed discretization, which divides the continuous input domain into finite number of bins, to transform the problem into finite-armed bandit. Previous works mostly consider Hölder-continuous ($\alpha \leq 1$) functions. For example Auer et al. [2007] study the α -Hölder continuous functions with $\alpha \leq 1$ for one-dimension domain, followed by Bubeck et al. [2010] who generalize it to d -dimensional domain and propose the HOO algorithm with adaptive discretization⁵. In these works, it suffices to perform random sampling (Auer et al. [2007], Bubeck et al. [2010]) or midpoint sampling (Kleinberg [2005]) inside each bin. The worst-case regret bound for Lipschitz space of $\tilde{\mathcal{O}}(T^{\frac{d+1}{d+2}})$ are matched by the general lower bound of $\Omega(T^{\frac{d+\alpha}{d+2\alpha}})$ (Auer et al. [2007], Bubeck et al. [2010], Locatelli and Carpentier [2018], Bubeck et al. [2011]) apart from log factors. However, if we apply the same methods of random sampling on fixed discretization (Auer et al. [2007]) on functions with Hölder exponent $\alpha > 1$, the regret incurred is $\tilde{\mathcal{O}}(T^{\frac{d+1}{d+2}})$ since the Hölder space with exponent $\alpha > 1$ is a subset of the Lipschitz function space. It prompts us to ask the question of whether a better rate that matches the dependence on α can be achieved for functions that are smoother than Lipschitz. An extreme is when α reaches infinity, where the reward model will be infinitely-differentiable, for example the stochastic linear bandit which enjoys $\tilde{\mathcal{O}}(T^{\frac{1}{2}})$ regret even on continuous domain (Dani et al. [2008], Abbasi-Yadkori et al. [2011]).

2.3.1 Algorithm Overview

We keep to fixed discretization of the domain since we consider only the worst-case regret. We divide $\mathcal{X} = [0, 1]^d$ into n equal-sized hypercubes, leaving n as a parameter of the algorithm. As shown in definition 1, the function is locally well-approximated by Taylor polynomial which reduces to a linear model of a feature map of x with dimension $d(\alpha)$. It is equivalent to observing a misspecified linear model inside each bin, the equivalence formally quantified in Lemma 2. Therefore, local exploration-exploitation tradeoff can be achieved by a base algorithm with sublinear regret on such misspecified models, with a Meta-algorithm to balance the budgets between the base algorithms in the bins.

⁴For simplicity, we assume L is some constant that satisfies assumption G1.

⁵The adaptive discretization does not change worst-case regret but has improvements on benign problems, as introduced in section 2.

Lemma 2. *Let hypercube \mathcal{B}_Δ be a subset of the input space with volume Δ . If a function satisfies assumption G1 \sim 2, there exists a linear parameter $\theta^* \in R^{d(\alpha)}$ and feature map $\phi : x \mapsto \phi(x) \in R^{d(\alpha)}$, such that f can be approximated by the linear function: $\|f - \langle \theta^*, \phi(x) \rangle\|_\infty \leq \epsilon = L\Delta^{\frac{\alpha}{d}}$ for $x \in \mathcal{B}_\Delta$. When $\alpha \leq 2$, $d(\alpha) = d$; when $\alpha > 2$, $d(\alpha) = \mathcal{O}(d^l)$ with l (definition 1). Note that the linear parameter may not be unique.*

The proof is in Appendix section 2.6.1. In the following parts of this section we first present the misspecified bandit algorithm to run inside a bin, and then the Meta-algorithms to control these local algorithms.

2.3.2 The Misspecified Linear Bandit Algorithm

In this subsection we escape from the big picture briefly in order to present the misspecified linear bandit algorithm, modified from the *ConfidenceBall*₂ algorithm in Dani et al. [2008] to serve as “arms” of the Meta algorithm. The algorithm, as shown in its name, is based on construction of confidence ellipsoid of the unobserved linear parameter in dimension d . We prove that the proposed modification can accommodate bias in the function feedback by deriving an upper bound on the cumulative regret⁶ of $\tilde{\mathcal{O}}(d\sqrt{T} + dT\epsilon)$. Here ϵ is the upper bound on bias value and known by the algorithm. We recently discovered that a similar result with proof sketch already appeared in recent work of Lattimore and Szepesvari [2019] (appendix E) who used modification of the algorithm in Abbasi-Yadkori et al. [2011], and hence enjoys the improvement of a multiplicative factor $\sqrt{\log(T)}$. For completeness and to provide necessary intermediate results for Meta-algorithms in later sections, we present our algorithm and full proof as complementary. It is worth mentioning that without the modification, the original algorithm incurs suboptimal regret under misspecification.

Assumptions

We make the following assumptions for the misspecified model. Note that they are consistent with the aforementioned global assumptions.

A1. The feedback model is $y = \langle x, \theta^* \rangle + b(x) + \eta$ with $|b(x)| \leq \epsilon, \forall x \in \mathcal{X} \in R^d$.

A2. The mean reward $\mathbb{E}[y]$ is bounded by $[-1, 1]$.

A3. The noise η is drawn from zero-mean sub-gaussian with parameter σ^7 .

The pseudo-code of the modified algorithm is shown in Algorithm 1. The goal is to minimize the cumulative pseudo-regret of the linear model:

$$R(T) = \sum_{t=1}^T r_t = \sum_{t=1}^T (\langle x^*, \theta^* \rangle - \langle x_t, \theta^* \rangle). \quad (2.2)$$

We prove that this regret is $\mathcal{O}\left(d \ln(T) \sqrt{\ln\left(\frac{T^2}{\delta}\right)T} + \epsilon T d \sqrt{2 \ln(T)}\right)$ with probability $1 - \delta$. This is formally stated in Theorem 3.

⁶For clarity this use of $\tilde{\mathcal{O}}$ omits $\ln(T)$ and δ dependence.

⁷Different from Dani et al. [2008] who assumes bounded noise. This reflects in the difference in β_t .

Algorithm 1 Misspecified linear UCB algorithm (\mathcal{A}^{local})

Require: Misspecification error ϵ , input domain \mathcal{X} and its dimension d , fail probability δ , upper bound on $\|x\|_2^2$: $\kappa^2 = d$.

1: Initialize $A_1 = I_d$ and $x_1 \in \mathcal{X}$ randomly.

2: **for** $t = 1 \dots$ **do**

3: Execute action x_t and observe reward y_t

4: $A_{t+1} = A_t + x_t x_t^T$

5: $\hat{\theta}_{t+1} = A_{t+1}^{-1} (\sum_{\tau=1}^t y_\tau x_\tau)$

6: $\beta_{t+1} = 128\sigma^2 d \ln(1+t) \ln(\frac{4(t+1)^2}{\delta})$

7: Define function $UCB_{t+1}(x) = (\langle x, \hat{\theta}_{t+1} \rangle + \sqrt{\beta_{t+1}} \|A_{t+1}^{-1/2} x\| + \epsilon \sum_{s=1}^t |x^T A_{t+1}^{-1} x_s|)$

8: Compute action $x_{t+1} = \arg \max_{x \in \mathcal{X}} UCB_{t+1}(x)$

9: Return x_{t+1} and $UCB_{t+1}(x_{t+1})$

10: **end for**

Theorem 3. *If assumptions A1~A3 hold, then with probability $1 - \delta$, the cumulative regret of Algorithm 1 is upper bounded by:*

$$R(T) \leq \sqrt{8d\beta_T T \ln(1+T)} + 2\epsilon T d \sqrt{2 \ln(1+T)} + 2\epsilon T. \quad (2.3)$$

The first term is the standard stochastic linear bandit regret rate same as in Dani et al. [2008]. We defer the proof to Appendix section 2.6.2. The increment of a multiplicative factor \sqrt{d} in the second term compared to that in Lattimore and Szepesvari [2019] is due to difference in assumption on $\|x\|^2$. Their assumption is $\|x\|^2 \leq 1$ whereas ours is $\|x\|^2 \leq d$.

2.3.3 The UCB-Meta-Algorithm

We now present the first structure of our Meta-algorithms. We consider the most straightforward structure: UCB-Meta, the pseudo-code is shown in Algorithm 2 (define $\lfloor \cdot \rfloor$ as the action of rounding to nearest integer.) . The confidence estimates of the local linear models are used as the UCB of arms of the Meta-algorithm with adjustment of ϵ , the bias quantity. For adjusting to different values of $l = \lfloor \alpha \rfloor$, we need only to change the space that the linear model is in, specifically the feature mapping $\phi : x \mapsto \phi(x) \in R^{d(\alpha)}$ as defined in proof of Lemma 2. For example, when $\alpha \leq 2$, the sub-algorithms are misspecified linear bandits whose actions spaces are simply bins $B \in \mathcal{X}$.

2.3.3.1 Regret Analysis of Algorithm 2

Theorem 4. *Let $d(\alpha)$ be the dimension of polynomial of x , as defined in Lemma 2. If the reward function satisfies G1~G3 in section 2.2, then with probability $1 - \delta$, the cumulative regret (equation 2.1) of UCB-Meta-Algorithm is upper bounded by⁸*

$$R(T) \leq \mathcal{O} \left(d(\alpha) \ln(T) \sqrt{T n \ln(T^2 n / \delta)} + d(\alpha) \epsilon T \sqrt{\ln(T)} \right). \quad (2.4)$$

⁸The d -dependence of the second term is propagated from Theorem 3

Algorithm 2 UCB-Meta-algorithm (\mathcal{A}^{global})

Require: smoothness parameter α , Hölder constant L , dimension of domain d , time horizon T and fail probability δ , action space \mathcal{X} .

- 1: Initialize $n = \lfloor T^{\frac{d}{d+2\alpha}} / \ln(T)^{\frac{2d}{d+2\alpha}} \rfloor$ and divide the action space \mathcal{X} into same-sized bins $B_{1\dots n}$ with volume $\Delta = 1/n$.
- 2: **for** $k = 1, \dots, n$ **do**
- 3: On bin B_k , start a version of local misspecified base-algorithm \mathcal{A}_k using misspecification error $\epsilon = Ln^{-\frac{\alpha}{d}}$, input domain $\mathcal{X}^* = \{\phi(x), x \in \mathcal{X}\}$ and its dimension $d(\alpha)$, fail probability δ/n .
- 4: Initialize counter $s_k = 1$ to indicate how many times \mathcal{A}_k is queried.
- 5: Query \mathcal{A}_k once by running steps 3-9 of Algorithm 1 with $t = s_k$ and obtain upper confidence bound UCB_k .
- 6: $s_k \leftarrow s_k + 1$
- 7: Update $UCB_k = UCB_k + \epsilon$.
- 8: **end for**
- 9: **for** $\tau = 1 \dots T$ **do**
- 10: Select the bin with index $k(\tau) = \arg \max_k UCB_k$.
- 11: Execute $\mathcal{A}_{k(\tau)}$ once by running steps 3-9 (of Algorithm 1) with $t = s_{k(\tau)}$
- 12: Receive updated recommendation $\phi_\tau \in \{\phi(x), x \in B_{k(\tau)}\}$ and $UCB_{k(\tau)}$.
- 13: Advance counter for $\mathcal{A}_{k(\tau)}$: $s_{k(\tau)} \leftarrow s_{k(\tau)} + 1$
- 14: Update $UCB_{k(\tau)} = UCB_{k(\tau)} + \epsilon$.
- 15: **end for**

The core of the proof is the distribution-independent analysis of UCB, which relies on the honesty of the confidence bands as well as their lengths. In particular, if the function value $f(x)$ at time t is contained in an honest confidence band $[UCB_t(x) - 2l_t(x), UCB_t(x)]$, then we can use the length $l_t(x)$ to bound instantaneous regret incurred by the selected action at this step. The confidence ellipsoids for the piecewise linear parameters $\hat{\theta}_{k,t}$ that are constructed by local misspecified linear bandits offer a convenient confidence estimation of function value, with the additional adjustment factor ϵ , the approximation error. The full proof is deferred to Appendix section 2.6.3. The algorithm defines each bin to be a hypercube with volume $\Delta = 1/n$, according to Lemma 2 we have $\epsilon = Ln^{-\frac{\alpha}{d}}$. Therefore, setting $n = \mathcal{O}(T^{\frac{d}{d+2\alpha}} / \ln(T)^{\frac{2d}{d+2\alpha}})$ will minimize the upper bound and yield the following cumulative regret bound.⁹

$$R(T) \leq \tilde{\mathcal{O}}(d(\alpha)T^{\frac{d+\alpha}{d+2\alpha}}). \quad (2.5)$$

2.3.3.2 Anytime Regret Guarantee for Algorithm 2

To achieve the rate in bound 2.5, Algorithm 2 needs to know the time horizon T in advance to set n and ϵ correspondingly. Here we prove that, with the doubling trick (Auer et al. [1995]), the UCB-Meta-algorithm can get regret that is of the same rate as in bound 2.5 up to constant factors without knowing T . This result is needed in the adaptation problem studied in section 2.4.

Theorem 5. *If Algorithm 2 with access to the time horizon T achieves regret of $\tilde{\mathcal{O}}(T^a)$ with probability $1 - \delta$, then the procedure described in Algorithm 3 can achieve regret rate $\tilde{\mathcal{O}}(T^a)$ with probability $1 - \delta$ without the knowledge of T .*

⁹ δ -dependence absorbed in \mathcal{O} since they are inside log terms.

The pseudo-code for Algorithm 3 is in Appendix section 2.7 and the proof of Theorem 5 in Appendix 2.6.3.4.

2.3.4 The Corral-Meta-Algorithm

Another choice for Meta-algorithm is bandit model selection methods. Here we use the Corral algorithm defined in Pacchiano et al. [2020b], which will be introduced more formally in section 4. An example of corraling misspecified linear bandit algorithms without corruption to the regret rate apart from log factors has already been given in Pacchiano et al. [2020b], but for adaptation to the misspecification error ϵ . Here we demonstrate that it can also be used to corral misspecified bandit base-algorithms on different bins in a discretized domain. We derive the following regret bound that is the same as UCB-Meta-algorithm.

Theorem 6. *First perform the smoothing transformation (Algorithm 3 in Pacchiano et al. [2020b]) to our misspecified linear bandits in Algorithm 1, denote the smoothed misspecified linear bandits as \mathcal{A}_s^{local} . Then, the Meta-algorithm (Algorithm 5 (Corral-Update) reproduced in Pacchiano et al. [2020b]) applied with a set of \mathcal{A}_s^{local} that are initialized in the same way as in Algorithm 2 has expected regret upper bounded by:*

$$\mathbb{E}[R(T)] \leq \tilde{O}(d(\alpha)T^{\frac{d+\alpha}{d+2\alpha}}). \quad (2.6)$$

The proof of this theorem is in Appendix section 2.6.4.

2.3.5 Discussion

The role of the Meta-algorithm is essentially model selection and adaptation to the base-algorithms. It is not a trivial task since the rewards incurred by the base-algorithms are not i.i.d as in standard stochastic settings. However, UCB as a stochastic multi-armed bandit algorithm, is applicable as Meta-algorithm because the local parametric (linear) function approximations provide honest upper confidence bounds for each bin even under the misspecifications, thus enabling the distribution-independent analysis for UCB. The advantage of Corral-Meta is that it potentially allows relaxation of the Hölder smoothness to hold only around the global maxima (Auer et al. [2007], Bubeck et al. [2010]), while the same relaxation is not straightforward for UCB-Meta. The advantage of UCB is that under standard stochastic settings where each arm has i.i.d rewards, it achieves the gap-dependent bound of $\mathcal{O}(\log(T)/\Delta)$. Thus an interesting question for the future is whether similar gap-dependent bounds for the UCB-Meta is available. Such bounds would enable exploitation of the growth conditions (section 2) for potential rate improvements.

2.3.6 Comparison with Existing Lower Bound

We compare the derived upper bounds of $\tilde{O}(d(\alpha)T^{\frac{d+\alpha}{d+2\alpha}})$ to the existing lower bound from Wang et al. [2018], which study global optimization. In their work, the performance of optimization algorithms with output \hat{x}_T is measured by simple regret $\mathcal{L}(\hat{x}_T; f) \triangleq f(x^*) - f(\hat{x}_T)$, for f in Hölder spaces including $\alpha \geq 1$. Theorem 2 (coupled with Proposition 3) in Wang et al. [2018] implies that

$\sup_{f \in \Sigma(\alpha)} \mathbb{E}[\mathcal{L}(\hat{x}_T; f)] = \Omega(T^{\frac{-\alpha}{2\alpha+d}})$. We argue that this lower bound can be directly used to lower bound the worst-case cumulative regret, by making the following observation (remark 3 in [Bubeck et al. \[2010\]](#)): If a strategy achieves expected cumulative regret $\mathbb{E}[R_T]$, then by uniformly selecting a past action as the final output \hat{x}_T , it can also achieve expected simple regret $\mathbb{E}[\mathcal{L}(\hat{x}_T; f)] = \mathbb{E}[R_T]/T$. Therefore, any strategy with cumulative regret $\tilde{o}(T\mathbb{E}[\mathcal{L}(\hat{x}_T; f)])$ will violate the lower bound. Through proof by contradiction, we take the result from [Wang et al. \[2018\]](#) as an $\Omega(T^{\frac{d+\alpha}{d+2\alpha}})$ lower bound on expected cumulative regret, and argue that our results match this bound up to log factors. Our results show that proposed algorithms are minimax optimal in dependence of T and effectively exploit the function smoothness.

2.4 Adaptation to Unknown Smoothness

In this section, we study adaptation to the smoothness exponent α of the reward function. Minimax adaptation, which means a learner can simultaneously achieve the minimax optimal rates ([Hoffmann et al. \[2011\]](#), [Locatelli and Carpentier \[2018\]](#)) under a nested set of Hölder spaces, has been proven to be impossible for cumulative regret minimization without additional assumptions. [Locatelli and Carpentier \[2018\]](#) provide a lower bound for adaptation between two Hölder continuous functions spaces. Assume $\alpha < \gamma \leq 1$, for any strategy with a good expected regret $\mathbb{E}[R_\gamma(T)]$ in $\Sigma(\gamma, L)$, they show that its expected regret in the superset $\Sigma(\alpha, L)$ will depend inversely on $\mathbb{E}[R_\gamma(T)]$, and therefore be suboptimal for $\Sigma(\alpha, L)$. They propose a strategy to match that lower bound that requires values of α and γ , thereby also proving that the lower bound is tight.

However, when adapting to a continuous scale of Hölder spaces (possibly $\alpha \geq 1$), it remains unclear what strategy can generalize and achieve this lower bound for some Hölder spaces. We aim to answer that question by proposing a new strategy that uses a recently developed bandit model selection algorithm (Corral with smooth wrapper in [Pacchiano et al. \[2020b\]](#)) applied with a set of Meta-algorithms (section 2.3). We will present this strategy and its theoretical guarantees next. Throughout the following sections, we refer to minimax optimal in dependence of T as minimax unless otherwise specified.

2.4.1 Corral Applied with Meta-Algorithms

The bandit model selection method Corral is first developed by [Agarwal et al. \[2016\]](#) and based on an instance of online mirror descent with mirror map derived from [Foster et al. \[2016\]](#). Corral with smooth wrapper proposed by [Pacchiano et al. \[2020b\]](#) for stochastic feedback problems is different from the original Corral algorithm in the following aspects. The smoothed version no longer needs to send importance-weighted feedback to base-algorithm, therefore no longer requires the base-algorithms themselves to be modified for stability guarantee (definition 3 in [Agarwal et al. \[2016\]](#)). In the following parts, we will use Corral with smooth wrapper to adapt to the smoothness and refer to it as Corral for simplicity¹⁰. We use a set of M Meta-algorithms $\mathcal{A}^{global}(\alpha_i), i \in [M]$ in Algorithm 2 as bases. The input values α_i are from a grid \mathcal{G} defined later. Therefore, we first specify the regret of a Meta-algorithm with input smoothness parameter α' that is ran on functions with actual Hölder smoothness α .

¹⁰Since the core of online mirror descent in Corral is not changed.

Lemma 7. *For function f that satisfies global assumptions $G1 \sim G3$ with parameter α , the regret of Algorithm 2 with input parameter $\alpha' \leq \alpha$ is bounded with probability $1 - \delta$ by*

$$R(T) \leq \tilde{\mathcal{O}}(d(\alpha')T^{\frac{d+\alpha'}{d+2\alpha'}}). \quad (2.7)$$

The bound does not hold for $\alpha' > \alpha$.

The proof is deferred to Appendix section 2.6.5. Having established the performance of base algorithms with misspecified smoothness exponents, we present the adaptation strategy and its regret bound in Theorem 8. Since it is impossible to achieve minimax optimal rates for multiple values of the smoothness parameter simultaneously, we introduce a user-specified parameter R that controls the Hölder space over which minimax optimality is desired. We show that conditioned on achieving minimax rate for the space $\sum(R, L)$, our adaptation strategy provides best possible regret bound on all supersets $\sum(\alpha, L)$ where $\alpha \leq R$. The results are stated in Theorem 8.

Theorem 8. *Consider adapting to a continuous scale of nested Hölder spaces indexed by α whose value is bounded in a given interval, for simplicity we assume $0 < \alpha \leq 2$, where $d(\alpha) = d$. Define $R \leq 2$ as a parameter set by the decision-maker that specifies the index of Hölder space for which minimax optimal regret is achieved. Define linear grid $\mathcal{G} = \{\alpha_i = \frac{R}{\lfloor \log(T) \rfloor} i, i = 0, 1 \dots \lfloor \log(T) \rfloor\}$ so that the total number of base algorithms is $M = |\mathcal{G}| = \lfloor \log(T) \rfloor$. Consider using Corral with bases that are Meta-algorithms (algorithm 3 in Appendix section 2.7) with input $\alpha_i \in \mathcal{G}, i \in [M]$. Then by setting the learning rate of Corral to be $\eta = d^{-1}T^{-\frac{d+R}{d+2R}}$, the regret rates achieved for any Hölder exponent $\alpha \in (0, 2]$ are:*

$$\sup_{f \in \sum(\alpha, L)} \mathbb{E}[R(T)] \leq \tilde{\mathcal{O}}(dT^{\frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)}}) \text{ for } \alpha \in (0, R], \quad (2.8)$$

$$\sup_{f \in \sum(\alpha, L)} \mathbb{E}[R(T)] \leq \tilde{\mathcal{O}}(dT^{\frac{d+R}{d+2R}}) \text{ for } \alpha \in [R, 2]. \quad (2.9)$$

A straightforward example is shown in Figure 2.1. There are two sources of cost of adaptation, first the cost of adapting to M grid points. Since $M = \mathcal{O}(\log(T))$, this has the same difficulty as the adaptation to two values in Locatelli and Carpentier [2018]. The second one, however, is a consequence of adapting to a continuous scale of α . The cost is the rate difference between the exponent α and the closest value to it on \mathcal{G} , denoted $\hat{\alpha} \in \mathcal{G}$, s.t. $\hat{\alpha} \leq \alpha \leq \hat{\alpha} + \frac{R}{\lfloor \log(T) \rfloor}$. This cost can be alleviated by the design of the linear grid. We defer the full proof to Appendix section 2.6.6.

2.4.2 Comparison with Existing Lower Bound for Adaptation

In this subsection, we compare the results in Theorem 8 to the existing lower bound in Locatelli and Carpentier [2018]. Theorem 3 of Locatelli and Carpentier [2018] state that given two smoothness values $\alpha_1 < \alpha_2 \leq 1$, if a strategy has expected regret $\mathbb{E}[R_{\alpha_2}(T)]$ under exponent α_2 that is $\tilde{\mathcal{O}}(T^{\frac{d+\alpha}{d+2\alpha}})$, then the regret of this strategy under the superset characterized by α_1 is lower bounded by $\sup_{f \in \sum(\alpha_1, L)} \mathbb{E}[R(T)] \geq \tilde{\Omega}(TR_{\alpha_2}(T)^{\frac{-\alpha_1}{\alpha_1+d}})$, even if the strategy has access to both α_1 and α_2 .

We make the following remark: for any pair of exponent values (α_1, α_2) where $\alpha_1 < R$ and $R \leq \alpha_2 \leq 2$, the strategy proposed in Theorem 8 matches the lower bound except for log factors. We verify this by plugging in $\mathbb{E}[R_{\alpha_2}(T)] = \tilde{\mathcal{O}}(T^{\frac{d+R}{d+2R}})$, omitting dependence on d , to yield the lower bound on $\sum(\alpha_1, L)$

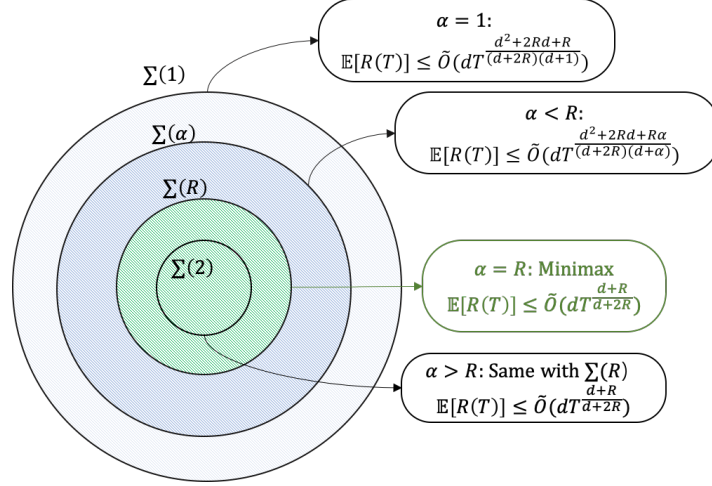


Figure 2.1: Illustration of adaptation to smoothness for continuous scale of Hölder spaces.

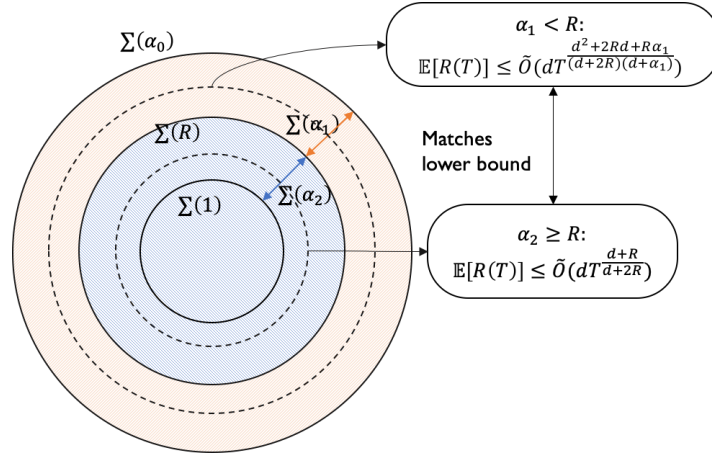


Figure 2.2: Illustration of values of exponents α_1, α_2 on which our proposed strategy matches the lower bound in [Locatelli and Carpentier \[2018\]](#).

which is $\tilde{O}(T^{\frac{d^2+2Rd+R\alpha_1}{d+\alpha_1}})$. This is matched by our upper bound in equation (2.8), apart from log factors and d . An illustration is shown in Figure 2.2. In other words, the proposed algorithm can perform under unknown smoothness exponent and match the lower bound (available only for exponent values within $(0, 1]$) on a subset of Hölder spaces.

2.5 Discussion

The core of this paper is extending the assumption on function space from Lipschitz to Hölder spaces with higher-order smoothness in bandit optimization of black-box functions. We also study adaptation to the smoothness under this scope. The class of two-layer algorithms that we proposed consists of a Meta-algorithm with the choice of UCB ([Auer et al. \[2002\]](#)) or Corral ([Agarwal et al. \[2016\]](#), [Pacchiano et al. \[2020b\]](#)) and a set of misspecified bandit base-algorithms as arms. We derive regret upper bounds for α -Hölder smooth functions with $\alpha > 1$ that matches existing lower bounds in their dependence

on T , the number of active queries, with straightforward generalization to larger α . Our framework provides useful insights in exploiting higher-order smoothness of reward functions for cumulative regret minimization, because our two-layer structure allows base-algorithms to perform local exploration-exploitation tradeoff as opposed to the local pure exploration done for bandit optimization of α -Hölder continuous functions. For adaptation to the smoothness exponent, we further previous works by deriving regret upper bound for adaptation to a continuous scale of Hölder spaces with exponent α in a given range. We show that by using bandit model selection algorithms, it can achieve the existing lower bound between two Hölder spaces, even if the algorithm does not know both exponent values.

Our work inspires several directions for the future. An intriguing direction is to study whether there exist gap-dependent bounds for the UCB-Meta algorithm, whose arms have non i.i.d rewards because they are bandit algorithms themselves. Such bounds could enable better rates for benign problem instances, for example with the growth conditions (mentioned in section 2). Another direction is the relaxation of the Hölder smooth assumption, to hold only around the maxima instead of everywhere on \mathcal{X} , which is considered by prior works such as Auer et al. [2007], Kleinberg et al. [2008], Bubeck et al. [2010]. Finally, it remains an open problem to establish the lower bound for adaptation when the smoothness exponents are larger than 1.

2.6 Proofs of Results

2.6.1 Proof of Lemma 2

Proof. Recall the definition of Hölder smoothness: $|f(x) - T_y^l(x)| \leq L\|x - y\|_\infty^\alpha$. For a hypercube B , $\|x - y\|_\infty \leq \Delta^{\frac{1}{\alpha}}, \forall x, y \in B$. By definition, when the function smoothness exponent $\alpha \in (1, 2]$, $l = 1$. Notice that the Taylor polynomial of degree $l = 1$ around y is a linear ¹¹ function of x : $T_y^{(l=1)}(x) = f(y) + \frac{\partial f}{\partial x_1}(y)(x_1 - y_1) + \frac{\partial f}{\partial x_2}(y)(x_2 - y_2) + \dots + \frac{\partial f}{\partial x_d}(y)(x_d - y_d) = \langle \theta, x \rangle$. When $\alpha > 2$, the Taylor polynomial can still be written as a linear function but of higher-dimensional feature map of x : $\phi : [0, 1]^d \rightarrow [0, 1]^{d(\alpha)}$ which contains exponentiations of elements in x , using the operations defined for definition 1, $\phi(x) = \{x^s, \forall s, s.t. |s| \leq l\}$. So:

$$d(\alpha) = |\{s : 1 \leq |s| \leq l\}| = \sum_{1 \leq j \leq l} \binom{j+d-1}{d-1} = \mathcal{O}(d^l) \quad (2.10)$$

When $l = 1$, it is equivalent to defining $\phi(x) = x$. The parameter θ is determined by the derivatives of f at y and the value of y . Therefore, we know locally there exists an unknown linear parameter in dimension $\theta^* = \arg \min_{\theta} \|f - \phi(x)^T \theta\|_\infty, x \in B$, such that $\|f - \langle \theta^*, \phi(x) \rangle\|_\infty \leq \epsilon = L\Delta^{\frac{\alpha}{2}}, \forall x \in B$. Also, note that $\|\phi(x)\|_2^2 \leq d(\alpha)^2$ according to definition. When the exponent $\alpha \in (0, 1]$, l is 0 and the Taylor polynomial is simply a constant. Therefore the same argument holds for θ^* for example when $\theta_1^*, \dots, \theta_d^* = 0$ (a constant function). \square

¹¹We slightly abuse the notation and define short-hand notation $\langle \theta, x \rangle := \theta_0 + \sum_{i=1}^{d(\alpha)} \theta_i x_i$.

2.6.2 Proof of Theorem 3

Proof. Throughout this proof, we assume that the assumptions A1~3 hold. This proof is modified from that in Dani et al. [2008]. Some techniques are from Abbasi-Yadkori et al. [2011]. We only present the parts which we change. First we proof the following bound on simple regret at each step:

$$r_t \leq 2\sqrt{\beta_t}\|A_t^{-1/2}x\| + 2\epsilon \sum_{\tau=1}^{t-1} \|x^T A_t^{-1}x_\tau\|. \quad (2.11)$$

And then we will bound the sum of these two terms separately. In order to proof inequality 2.11, we start from an important auxiliary theorem of confidence bound on θ^* , Theorem 9.

Theorem 9. Let $\beta_t = C\sigma^2 d \ln(1+t\kappa^2/d) \ln(\frac{2t^2}{\delta})$ ($= \mathcal{O}(d \ln(t) \ln(\frac{t^2}{\delta}))$) for a sufficiently large constant C , then with probability $1 - \delta$, θ^* is contained in the confidence set:

$$\tilde{C}_t = \{\hat{\theta}_t + \sqrt{\beta_t}A_t^{-1/2}z_d - A_t^{-1}(\sum_{s=1}^{t-1} b_s x_s)\},$$

and as a result,

$$\langle x, \theta^* \rangle \leq \langle x, \hat{\theta}_t \rangle + \sqrt{\beta_t}\|A_t^{-1/2}x\| + \epsilon \sum_{s=1}^{t-1} |x^T A_t^{-1}x_s|.$$

The proof of Theorem 9 is in Appendix 2.6.2.1. Now, if $\theta^* \in \tilde{C}_t$, we have

$$\begin{aligned} r_t &= \langle x^*, \theta^* \rangle - \langle x_t, \theta^* \rangle \\ &\leq \langle x^*, \theta^* \rangle - UCB_t(x^*) + UCB_t(x_t) - \langle x_t, \theta^* \rangle \\ &\leq UCB_t(x_t) - \langle x_t, \theta^* \rangle \\ &\leq 2\sqrt{\beta_t}\|A_t^{-1/2}x_t\| + 2\epsilon \sum_{s=1}^{t-1} |x_t^T A_t^{-1}x_s|. \end{aligned}$$

The first inequality is because our algorithm will only choose x_t when $UCB_t(x_t) \geq UCB_t(x^*)$. The last inequality holds because

$$\begin{aligned} \langle x, \theta^* \rangle &\geq \langle x, \hat{\theta}_t \rangle + \min_{z_d \in B_2^d} \sqrt{\beta_t} \langle x, A_t^{-1/2}z_d \rangle - \sum_{s=1}^{t-1} b_s x^T A_t^{-1}x_s \\ &\geq \langle x, \hat{\theta}_t \rangle - \sqrt{\beta_t}\|A_t^{-1/2}x\| - \sum_{s=1}^{t-1} b_s x^T A_t^{-1}x_s \\ &\geq UCB_t(x) - 2\sqrt{\beta_t}\|A_t^{-1/2}x\| - 2\epsilon \sum_{s=1}^{t-1} |x^T A_t^{-1}x_s|. \end{aligned}$$

By assumption on the mean reward function value, the absolute value of instant pseudo-regret $|r_t|$ is bounded by $1 + \epsilon$. Therefore, combining inequality (2.11) and $r_t \leq 2 + 2\epsilon$, we have that¹²

$$\begin{aligned} r_t &\leq (2 + 2\epsilon) \wedge \left(2\sqrt{\beta_t}\|A_t^{-1/2}x_t\| + 2\epsilon \sum_{\tau=1}^{t-1} \|x_t^T A_t^{-1}x_\tau\| \right) \\ &\leq 2 \underbrace{\left(1 \wedge \sqrt{\beta_t}\|A_t^{-1/2}x_t\| \right)}_{\#1} + 2\epsilon \underbrace{\sum_{\tau=1}^{t-1} \|x_t^T A_t^{-1}x_\tau\|}_{\#2} + 2\epsilon. \end{aligned} \quad (2.12)$$

¹² $a \wedge b = \min(a, b)$

Sum of term #1 is bounded using bound (2.28) and Cauchy Schwartz inequality:

$$2 \sum_{t=1}^T (1 \wedge \sqrt{\beta_t} \|A_t^{-1/2} x_t\|) \leq 2 \sqrt{T \beta_T \sum_{t=1}^T (1 \wedge \|x_t^T A_t^{-1} x_t\|)} = \sqrt{8d\beta_T T \ln(1 + T\kappa^2/d)}. \quad (2.13)$$

For sum of term #2, we first have

$$\begin{aligned} \sum_{\tau=1}^{t-1} x_\tau^T A_t^{-1} x_\tau &\leq \sqrt{t \sum_{\tau=1}^{t-1} x_\tau^T A_t^{-1} x_\tau x_\tau^T A_t^{-1} x_t} \\ &= \sqrt{t x_t^T A_t^{-1} \left(\sum_{\tau=1}^{t-1} x_\tau x_\tau^T \right) A_t^{-1} x_t} \\ &\leq \sqrt{t x_t^T A_t^{-1} \left(\sum_{\tau=1}^{t-1} x_\tau x_\tau^T \right) A_t^{-1} x_t + x_t^T A_t^{-1} A_t^{-1} x_t} \\ &= \sqrt{t x_t^T A_t^{-1} \left(\sum_{\tau=1}^{t-1} x_\tau x_\tau^T + I_d \right) A_t^{-1} x_t} = \sqrt{t x_t^T A_t^{-1} x_t}. \end{aligned}$$

Then the sum $\sum_{t=1}^T \left(\sum_{\tau=1}^{t-1} x_\tau^T A_t^{-1} x_\tau \right)$ can be bounded by:

$$\begin{aligned} \sum_{t=1}^T \left(\sum_{\tau=1}^{t-1} x_\tau^T A_t^{-1} x_\tau \right) &\leq \sum_{t=1}^T \left(\sqrt{t x_t^T A_t^{-1} x_t} \right) \\ &\leq \sqrt{\left(\sum_{t=1}^T t \right) \left(\sum_{t=1}^T x_t^T A_t^{-1} x_t \right)}. \end{aligned}$$

Now, we need to bound $\sum_{t=1}^T x_t^T A_t^{-1} x_t$ with inequality (2.28). We know that A_t^{-1} is a full-rank matrix. Therefore, denote its eigenvalues and eigenvectors as $\lambda_1 \dots \lambda_d, v_1 \dots v_d$. Then¹³

$$\begin{aligned} x_t^T A_t^{-1} x_t &= (c_1 v_1 + \dots + c_d v_d)^T A_t^{-1} (c_1 v_1 + \dots + c_d v_d) \\ &= c_1^2 \lambda_1 + \dots + c_d^2 \lambda_d \\ &\leq \lambda_{\max}(A_t^{-1}) \|x_t\|_2^2 = \frac{\kappa^2}{\lambda_{\min}(A_t)} \\ &\leq \frac{\kappa^2}{\lambda_{\min}(I_d) + \lambda_{\min}(X_t^T X_t)} \leq \kappa^2. \end{aligned}$$

The second last inequality holds due to Weyl's inequality. Therefore,

$$\begin{aligned} \sum_{t=1}^T x_t^T A_t^{-1} x_t &\leq \kappa^2 \sum_{t=1}^T (x_t^T A_t^{-1} x_t \wedge 1) \\ &\leq \kappa^2 (2d \ln(1 + T\kappa^2/d)). \end{aligned}$$

Putting the above together,

$$\begin{aligned} \sum_{t=1}^T \left(2\epsilon \sum_{\tau=1}^{t-1} x_\tau^T A_t^{-1} x_\tau \right) &\leq 2\epsilon \sqrt{\left(\sum_{t=1}^T t \right) \left(\sum_{t=1}^T x_t^T A_t^{-1} x_t \right)} \\ &\leq 2\epsilon T \kappa \sqrt{2d \ln(1 + T\kappa^2/d)}. \end{aligned} \quad (2.14)$$

Finally, plugging in $\kappa^2 = d$ gives the final results. \square

¹³This proof is extracted from a remark in proof of Theorem 3 in Abbasi-Yadkori et al. [2011]

2.6.2.1 Proof of Theorem 9

Proof. Let $\hat{\theta}_t = A_t^{-1} X_t^T y$ denote the regularized least square estimator at time t . Matrix X_t has dimension $(t-1) \times d$, where each row is a past action (until time t). We first define an unobserved variable $\tilde{\theta}_t$:

$$\tilde{\theta}_t = A_t^{-1} X_t^T (X_t \theta^* + \eta_t) = \hat{\theta}_t - A_t^{-1} X_t^T b_t, \quad (2.15)$$

here we abuse the notations and let η_t and b_t be the $(t-1) \times 1$ vector containing noise and bias of each time. Then we define the following confidence ellipsoid centered at $\tilde{\theta}_t$:

$$C_t = \{\theta : (\theta - \tilde{\theta}_t)^T A_t (\theta - \tilde{\theta}_t) \leq \beta_t\}, \quad (2.16)$$

and prove the following lemma as an analog to Theorem 5 of Dani et al. [2008]:

Lemma 10. *The true linear parameter θ^* is contained in ellipsoid C_t , specifically, $\mathbb{P}(\forall t, \theta^* \in C_t) \geq 1 - \delta$.*

The proof is in Appendix section 2.6.2.2. However, we do not observe the vector b_t , so we cannot calculate C_t in our algorithm. So instead, we define a larger \tilde{C}_t that contains C_t , which will naturally contains θ^* with high probability. To construct \tilde{C}_t , we first re-write C_t as

$$C_t = \{\tilde{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z_d, \|z_d\|_2 \leq 1\}, \quad (2.17)$$

then plug in equation (2.15) to yield:

$$\begin{aligned} \tilde{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z &= \tilde{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z - A_t^{-1} X_t^T b_t \\ &= \tilde{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z - A_t^{-1} \left(\sum_{s=1}^{t-1} b_s x_s \right). \end{aligned} \quad (2.18)$$

Therefore, we know that with high probability,

$$\theta^* \in \tilde{C}_t = \{\hat{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z_d - A_t^{-1} \left(\sum_{s=1}^{t-1} b_s x_s \right)\}. \quad (2.19)$$

Therefore, we have a computable confidence bound for x :

$$\begin{aligned} UCB_t(x) &= \max_{\theta \in \tilde{C}_t} \langle x, \theta \rangle \\ &= \langle x, \hat{\theta}_t \rangle + \max_{z_d \in B_2^d} \sqrt{\beta_t} \langle x, A_t^{-1/2} z_d \rangle - \sum_{s=1}^{t-1} b_s x^T A_t^{-1} x_s \\ &\leq \langle x, \hat{\theta}_t \rangle + \sqrt{\beta_t} \|A_t^{-1/2} x\| - \sum_{s=1}^{t-1} b_s x^T A_t^{-1} x_s \\ &\leq \langle x, \hat{\theta}_t \rangle + \sqrt{\beta_t} \|A_t^{-1/2} x\| + \epsilon \sum_{s=1}^{t-1} |x^T A_t^{-1} x_s|. \end{aligned} \quad (2.20)$$

The first inequality is derived by Cauchy Schwartz inequality and the fact that z_d is in unit ball. \square

2.6.2.2 Proof of Lemma 10

Proof. Lemma 10 is a parallel to Theorem 5 in Dani et al. [2008], with the difference of sub-gaussian noise, ellipsoid centre $\tilde{\theta}_t$ and misspecification in observation. The key idea is the same, namely to use induction to bound the growth of $Z_t = (\theta^* - \tilde{\theta}_t)^T A_t (\theta^* - \tilde{\theta}_t)$ and proof that $Z_t \leq \beta_t$, i.e. the θ^* is contained in C_t , at each time step t . The following analysis used the same notations and definitions as section 5.2 in Dani et al. [2008] unless otherwise specified. Under Lemma 10's definition of confidence set C_t , we have that:

$$H_t = A_t(\tilde{\theta}_t - \theta^*) = X_t^T \eta_t - \theta^*, \quad (2.21)$$

$$Z_t = (\theta^* - \tilde{\theta}_t)^T A_t (\theta^* - \tilde{\theta}_t) = H_t^T A_t^{-1} H_t. \quad (2.22)$$

Equation 2.21 holds because of this key property:

$$\tilde{\theta}_t : A_t \tilde{\theta}_t = X_t^T X_t \theta^* + X_t^T \eta_t. \quad (2.23)$$

And the rest of the proof in Dani et al. [2008] should go through by substituting Y_t with H_t (defined above) and $\hat{\mu}$ with our definition of $\tilde{\theta}$ (centre of the confidence ellipsoid). Except, to accommodate the sub-gaussian noise assumption that replaces their bounded noise assumption, we have to make two changes in the proof. Both are in analyzing the growth of Z_t in the induction. Recall that Dani et al. [2008] proved this relation:

$$Z_t \leq Z_1 + 2 \sum_{\tau=1}^{t-1} \eta_\tau \frac{x_\tau^T (\tilde{\theta}_t - \theta^*)}{1 + w_\tau^2} + \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1 + w_\tau^2}. \quad (2.24)$$

We first look at the concentration of the sum of martingale difference sequence that makes up Z_t : same with Dani et al. [2008], define $M_t = 2\eta_t \frac{x_t^T (\tilde{\theta}_t - \theta^*)}{1 + w_t^2}$ where $w_t \triangleq \sqrt{x_t^T A_t^{-1} x_t}$. According to our assumption, the noise sequence is a sub-gaussian martingale difference sequence with parameter σ^2 . Therefore, M_t is a sub-gaussian martingale difference sequence. Specifically, we know that the square of subgaussian parameter is $4\sigma^2 \left(\frac{|x_t^T (\tilde{\theta}_t - \theta^*)|}{1 + w_t^2} \right)^2$. By definitions we know that $M_t | \mathcal{H}_t$ is $(\nu_t^2 = 4\sigma^2 \left(\frac{|x_t^T (\tilde{\theta}_t - \theta^*)|}{1 + w_t^2} \right)^2, a_t = 0)$ sub-exponential (definition 2.7 in Wainwright [2019]) and therefore the sum $\sum_{\tau=1}^t M_\tau$ is also sub-exponential, with parameters $(\sqrt{\sum_{\tau=1}^t \nu_\tau^2}, a = \max_\tau a_\tau = 0)$ (Theorem 2.19 (1) in Wainwright [2019]). The following inequality is conditioned on the fact that from time $\tau = 1 \dots t$, θ^* is contained in C_τ (by the induction).

$$\begin{aligned} \sum_{\tau} \nu_\tau^2 &= 4\sigma^2 \sum_{\tau=1}^t \left(\frac{|x_\tau^T (\tilde{\theta}_\tau - \theta^*)|}{1 + w_\tau^2} \right)^2 \\ &\leq 4\sigma^2 \sum_{\tau=1}^t \left(\frac{\sqrt{\beta_\tau} w_\tau}{1 + w_\tau^2} \right)^2 \\ &\leq 4\sigma^2 \sum_{\tau=1}^t \beta_\tau (\min(1/2, w_\tau))^2 \\ &\leq 4\sigma^2 \sum_{\tau=1}^t \beta_\tau \min(1/4, w_\tau^2) \\ &\leq 4\sigma^2 \beta_t \sum_{\tau=1}^t \min(1, w_\tau^2) \\ &\leq 4\sigma^2 \beta_t (2d \ln(1 + t\kappa^2/d)) \text{ See bound 2.28} \\ &= 8\sigma^2 d \beta_t \ln(1 + t\kappa^2/d). \end{aligned}$$

The proof for the first three inequalities is the same as Lemma 7 and section 5.2.1 in Dani et al. [2008]. Then we apply a Bernstein-type concentration bound for sub-exponential martingale difference sequence (Theorem 2.19 (2) in Wainwright [2019]). Plugging in the values of a and $\sum_{\tau=1}^t \nu_\tau^2$, we have that

$$\begin{aligned} \mathbb{P}\left(\left|\sum_{\tau=1}^{t-1} M_\tau\right| \geq s\right) &\leq 2 \exp\left(\frac{-s^2}{2 \sum_{\tau=1}^{t-1} \nu_\tau^2}\right) \\ &\leq 2 \exp\left(\frac{-s^2}{16\sigma^2 d \beta_t \ln(1 + (t-1)\kappa^2/d)}\right) \\ &\stackrel{s=\frac{\beta_t}{2}}{=} 2 \exp\left(\frac{-\beta_t}{64\sigma^2 d \ln(1 + (t-1)\kappa^2/d)}\right) \\ &\leq \frac{\delta}{2t^2} \text{ (Needed for union bound over all times).} \end{aligned} \tag{2.25}$$

Therefore, as long as β_t is larger or equal to $64\sigma^2 d \ln(1 + (t-1)\kappa^2/d) \ln(\frac{4t^2}{\delta})$, $\sum_{\tau=1}^{t-1} M_\tau \leq \frac{\beta_t}{2}$ with probability larger or equal to $1 - \frac{\delta}{2t^2}$.

The second change is for the third quantity that makes up Z_t : $\sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1+w_\tau^2}$. We need to bound $\max_{\tau \leq t-1} \eta_\tau^2$ with high probability. By algebra calculations, we know that η_τ^2 is sub-exponential with parameters $(\nu = 32\sigma^4, a = 4\sigma^2)$ ¹⁴. We can apply union bound with the tail bound of sub-exponential variables:

$$\begin{aligned} \mathbb{P}\left(\max_{\tau \leq t-1} (\eta_\tau^2 - \mathbb{E}[\eta^2]) \geq z\right) &\leq \sum_{\tau=1}^{t-1} \mathbb{P}((\eta_\tau^2 - \mathbb{E}[\eta^2]) \geq z) \\ &\leq (t-1) \exp\left(-\frac{z}{2a}\right) \text{ (Proposition 2.9 in Wainwright [2019])} \\ &\leq \frac{\delta}{2t^2} \text{ (Needed for union bound over all times).} \end{aligned}$$

Set $z = 8\sigma^2 \ln(\frac{2t^3}{\delta})$ so that $\mathbb{P}(\max_{\tau \leq t-1} \eta_\tau^2 - \mathbb{E}[\eta^2] \leq z) = \mathbb{P}(\max_{\tau \leq t-1} \eta_\tau^2 \leq z + \mathbb{E}[\eta^2]) \geq 1 - \frac{\delta}{2t^2}$. By the fact that $\mathbb{E}[\eta] = 0$, $\mathbb{E}[\eta^2] = \text{Var}(\eta) \leq \sigma^2$, which is a property of subgaussian variables. So $\mathbb{P}(\max_{\tau \leq t-1} \eta_\tau^2 \leq z + \sigma^2) \geq 1 - \frac{\delta}{2t^2}$. The following holds with probability larger than $1 - \frac{\delta}{2t^2}$:

$$\begin{aligned} \sum_{\tau=1}^{t-1} \eta_\tau^2 \frac{w_\tau^2}{1+w_\tau^2} &\leq \left(\max_{\tau \leq t-1} \eta_\tau^2\right) \sum_{\tau=1}^{t-1} \min(w_\tau^2, 1) \\ &\leq \left(\max_{\tau \leq t-1} \eta_\tau^2\right) 2d \ln(1 + t\kappa^2/d) \\ &= \left(8\sigma^2 \ln\left(\frac{2t^3}{\delta}\right) + \sigma^2\right) 2d \ln(1 + (t-1)\kappa^2/d) \\ &= 8\sigma^2 \left(\ln\left(\frac{2t^3}{\delta}\right) + \frac{1}{8}\right) 2d \ln(1 + (t-1)\kappa^2/d) \\ &= 16\sigma^2 d \ln(1 + (t-1)\kappa^2/d) \left(\ln\left(\frac{2t^3}{\delta}\right) + \frac{1}{8}\right). \end{aligned}$$

Except the two changes above, one last thing to note is the quantity Z_1 analyzed at the end of proof

¹⁴For this part, we used the proof from Example 2.8 in Wainwright [2019] and <http://proceedings.mlr.press/v33/honorio14-suppl.pdf>

of Lemma 12 in Dani et al. [2008]. In our assumption of the reward function value, we conclude that

$$\begin{aligned} Z_1 &= (\theta^* - 0)^T I(\theta^* - 0) = \|\theta^*\|_2^2 \\ &= \sum_{i=1}^d (e_i^T \theta^*)^2 \quad (e_i \text{ is base vector of dimension } i, \text{ note that } e_i \in \mathcal{X}) \\ &\leq d(1 + \epsilon)^2. \end{aligned}$$

As a result, if it is satisfied that $Z_t \leq Z_1 + \beta_t/2 + 16\sigma^2 d \ln(1 + (t-1)\kappa^2/d)(\ln(\frac{2t^3}{\delta}) + \frac{1}{8}) \leq \beta_t$, which enables the induction in Lemma 14 in Dani et al. [2008], then the rest of the proof should go through smoothly. We argue that setting $\beta_t = C\sigma^2 d \ln(t) \ln(\frac{4t^2}{\delta})$ for a large enough constant C suffices. This is under the reasonable assumption that ϵ is $\mathcal{O}(1)$ and σ is a constant¹⁵.

It is worth mentioning¹⁶ that Dani et al. [2008] requires the relationship between t and δ to be approximately $0 < 1.05\delta \leq t^2$, hence their requirement¹⁷ of “for sufficiently large T ” in Theorem 1 and 2. This is because of the last step of their induction proof for Theorem 5 requires: $Z_t \leq d + \beta_2/2 + 2d \ln(t) \leq \beta_t$. In our setting, the requirement in induction translates to this (second) constraint(plugging in $\kappa^2 = d$): $\beta_t \geq 2d(1 + \epsilon)^2 + 32\sigma^2 d \ln(t)(\ln(\frac{2t^3}{\delta}) + \frac{1}{8})$. Recall the first constraint on β_t is $\beta_t \geq 64\sigma^2 d \ln(t) \ln(\frac{4t^2}{\delta})$, from bound (2.25). Therefore, C should first satisfy $C \geq 64$ and for the second constraint we need¹⁸: $C \geq \frac{3(1+\epsilon)^2}{4(\ln(2))^2\sigma^2} + \frac{3}{2\ln(2)} + 48$. Therefore, the lower bound of C should depend on values of ϵ and σ^2 . The choice of $C = 128$ in the main theorem is an example that requires approximately $\frac{1+\epsilon}{\sigma} \leq 7$. \square

2.6.3 Proof of Theorem 4

Let us treat the number of bins/local algorithms n as the input parameter to the algorithm. The regret bound of UCB-Meta (equation 2.4) should be independent of the input dimension d , given the dimension of the linear model $d(\alpha)$. Therefore, throughout this proof we will abuse the notations and let d denote the linear model dimension for simplicity.

Proof. First, we define the “good event” E_{good} as an event where all confidence bound holds for all bins at all times. For a fixed bin, if $\mathbb{P}(\theta^* \notin \tilde{C}_t, \exists t) \leq \delta/n$, as set in the algorithm, where $\tilde{C}_t = \{\hat{\theta}_t + \sqrt{\beta_t} A_t^{-1/2} z_d - A_t^{-1}(\sum_{s=1}^{t-1} b_s x_s)\}$ (Theorem 9), then by union bound, $\mathbb{P}(\theta_k^* \notin \tilde{C}_{k,t}, \exists k) \leq \delta$, where $\tilde{C}_{k,t}$ is the confidence ellipsoid of bin k at time t . The good event is $E_{good} = \{\forall t, \forall k \in [n], \theta_k^* \in \tilde{C}_{k,t}\}$. It happens with probability $\mathbb{P}(E_{good}) \geq 1 - \delta$, and the following proof will condition on it.

Here are some useful notations that make the proof easier to read: let $N^k(t)$ denote the number of times base-algorithm \mathcal{A}_k^{local} has been selected by(including) time t ; let $k(t)$ denote the bin selected at time t ; let x_t denote the action selected at time t ; let $\{\beta_{k,\cdot}\}$, $\{A_{k,\cdot}\}$ and $\{\hat{\theta}_{k,\cdot}\}$ denote the set of parameters kept by that base-algorithm \mathcal{A}_k^{local} .

¹⁵Recall that according to Lemma 2, ϵ is bounded by the Lipschitz constant L and is therefore $\mathcal{O}(1)$

¹⁶This remark is made by Abbasi-Yadkori et al. [2011].

¹⁷However, we believe that this should not translate to a constraint on t , but on δ instead. Because $Z_t \leq \beta_t$ is required for every step t to complete the induction, so if it only holds for large t then the induction will fail as well.

¹⁸This is from the second constraint: $C\sigma^2 d \ln(t) \ln(\frac{4t^2}{\delta}) \geq \frac{2}{3} C\sigma^2 d \ln(t) \ln(\frac{2t^3}{\delta}) \geq 2d(1+\epsilon)^2 + 32\sigma^2 d \ln(t)(\ln(\frac{2t^3}{\delta}) + \frac{1}{8})$.

The upper confidence bound on value of the local linear function achieved by sub-algorithms at round t is defined as $UCB_{k(t),t}(x) = \langle x, \hat{\theta}_{k,N^k(t)} \rangle + \sqrt{\beta_{k,N^k(t)}} \|A_{k,N^k(t)}^{-1/2} x\| + \epsilon \sum_{\tau=1}^{N^k(t)-1} |x^T A_{N^k(t)}^{-1} x_\tau|$ for any action $x \in B_k$. Using the proof of Theorem 3, the good event hence indicates that for the base-algorithm selected at time t and any action $x \in B_{k(t)}$:

$$UCB_{k(t),t}(x) - 2\sqrt{\beta_{k,N^k(t)}} \|A_{k,N^k(t)}^{-1/2} x\| - 2\epsilon \sum_{\tau=1}^{N^k(t)-1} |x^T A_{N^k(t)}^{-1} x_\tau| \leq \langle x, \theta_k^* \rangle \leq UCB_{k(t),t}(x).$$

By Lemma 2, the expected local function value $f(x)$ is bounded by

$$UCB_{k(t),t}(x) - 2\sqrt{\beta_{k,N^k(t)}} \|A_{k,N^k(t)}^{-1/2} x\| - 2\epsilon \sum_{\tau=1}^{N^k(t)-1} |x^T A_{N^k(t)}^{-1} x_\tau| - \epsilon \leq f(x) \leq UCB_{k(t),t}(x) + \epsilon.$$

A common way to bound pseudo regret for stochastic bandit is via Wald's equality: $R_T = \sum_{k=1}^n \Delta_k \mathbb{E}[\tau_k(T)]$ where $\tau_k(T)$ is the number of times arm k gets pulled until time T , and Δ_k is the reward gap. We cannot trivially follow this, because the rewards of each bins are no longer i.i.d. Instead, we use this gap-independent decomposition for each bin k :

$$\begin{aligned} R_k &= \sum_{t:\text{bin}_t=k} (f^* - f_{x_t \in B_k}(x_t)) \\ &= \sum_{t:\text{bin}_t=k} (f^* - UCB_{\mathcal{A}_{k(t)},t} + UCB_{\mathcal{A}_{k(t)},t} - f(x_t)) \\ &= \sum_{t:\text{bin}_t=k} (f^* - UCB_{\mathcal{A}_{k(t)},t} + UCB_{k(t),t}(x_t) + \epsilon - f(x_t)) \\ &\leq \sum_{t:\text{bin}_t=k} (UCB_{k(t),t}(x_t) + \epsilon - f(x_t)) \tag{2.26} \\ &\leq \sum_{t:\text{bin}_t=k} \left(2\sqrt{\beta_{k,N^k(t)}} \|A_{k,N^k(t)}^{-1/2} x_t\| + 2\epsilon \sum_{\tau=1}^{N^k(t)-1} |x_t^T A_{N^k(t)}^{-1} x_\tau| + 2\epsilon \right) \\ &= \sum_{s=1}^{N^k(T)} \left(2\sqrt{\beta_{k,s}} \|A_{k,s}^{-1/2} x_t\| + 2\epsilon \sum_{\tau=1}^{s-1} |x_t^T A_s^{-1} x_\tau| + 2\epsilon \right). \end{aligned}$$

The first inequality holds because of the algorithm's bin selection rule: if bin B_k is chosen then $f^* \leq UCB_{k^*,t} \leq UCB_{k(t)}$. By the bounded function value assumption, $f^* - f_{x_t \in B_k}(x_t) \leq 2$, therefore:

$$\begin{aligned} R_k &\leq \sum_{s=1}^{N^k(T)} \left(2\sqrt{\beta_{k,s}} \|A_{k,s}^{-1/2} x_t\| + 2\epsilon \sum_{\tau=1}^{s-1} |x_t^T A_s^{-1} x_\tau| + 2\epsilon \right) \wedge 2 \\ &\leq \sum_{s=1}^{N^k(T)} \left(\underbrace{2 \left(\sqrt{\beta_{k,s}} \|A_{k,s}^{-1/2} x_t\| \wedge 1 \right)}_{\#1} + \underbrace{2\epsilon \sum_{\tau=1}^{s-1} |x_t^T A_s^{-1} x_\tau|}_{\#2} \right) + 2\epsilon N^k(T). \tag{2.27} \end{aligned}$$

2.6.3.1 High probability regret bound part I (term #1)

First we establish this bound the same way as Dani et al. [2008]. Namely, for any local misspecified linear bandit algorithm that is ran T times with data $(x_t, y_t)_{t=1\dots T}$,

$$\begin{aligned}
\sum_{t=1}^T \|x_t^T A_t^{-1} x_t\| \wedge 1 &\leq 2 \ln \left(\prod_{t=1}^T (1 + x_t^T A_t^{-1} x_t) \right) \\
&= 2 \ln \left(\prod_{t=1}^T \frac{\det(A_{t+1})}{\det(A_t)} \right) \\
&= 2 \ln \left(\frac{\det A_{T+1}}{\det A_1} \right) \leq 2 \ln((1 + T\kappa^2/d)^d) \\
&= 2d \ln(1 + T\kappa^2/d),
\end{aligned} \tag{2.28}$$

where we used Lemma 11. Now we can bound term #1 using bound (2.28).

$$\begin{aligned}
&\sum_{s=1}^{N^k(T)} 2(\sqrt{\beta_{k,s}} \|A_{k,s}^{-1/2} x_t\| \wedge 1) \\
&\leq \sqrt{N^k(T) \sum_{s=1}^{N^k(T)} 4(\beta_{k,s} \|x_{k,s}^T A_{k,s}^{-1} x_{k,s}\| \wedge 1)} \\
&\leq \sqrt{4\beta_{k,N^k(T)} N^k(T) \sum_{s=1}^{N^k(T)} \|x_{k,s}^T A_{k,s}^{-1} x_{k,s}\| \wedge 1} \\
&= \sqrt{4\beta_{k,N^k(T)} N^k(T) 2 \ln \left(\prod_{s=1}^{N^k(T)} (1 + x_{k,s}^T A_{k,s}^{-1} x_{k,s}) \right)} \\
&= \sqrt{4\beta_{k,N^k(T)} N^k(T) 2 \ln \left(\frac{\det(A_{N^k(T)+1})}{\det(A_1)} \right)} \\
&= \sqrt{8d\beta_{k,N^k(T)} N^k(T) \ln(1 + N^k(T)\kappa^2/d)} \\
&\stackrel{\kappa^2 \equiv d}{=} \sqrt{8d\beta_{k,N^k(T)} N^k(T) \ln(1 + N^k(T))}.
\end{aligned}$$

Lemma 11. For $t \geq 1$, $1 + x_t^T A_t^{-1} x_t = \det(A_{t+1})/\det(A_t)$. Also, $\det(A_t) \leq (1 + (t-1)\kappa^2/d)^d$.

Proof of Lemma 11.

$$\begin{aligned}
\det(A_{t+1}) &= \det(A_t(I_d + A_t^{-1} x_t x_t^T)) = \det(A_t) \det(I_d + A_t^{-1} x_t x_t^T) \\
&= \det(A_t) \det(I_1 + x_t^T A_t^{-1} x_t) = \det(A_t)(1 + x_t^T A_t^{-1} x_t).
\end{aligned}$$

The third equation uses Sylvester's determinant theorem: $\det(I_m + A_{m \times n} B_{n \times m}) = \det(I_n + B_{n \times m} A_{m \times n})$. The trace of a matrix is the product of its eigenvalues and the determinant is the sum of eigenvalues, and for the trace of the positive definite matrix A_t we have,

$$\text{tr}(A_t) = \text{tr}\left(I + \sum_{\tau}^{t-1} x_{\tau} x_{\tau}^T\right) = d + \sum_{\tau}^{t-1} \|x_{\tau}\|_2^2 \leq d + (t-1)\kappa^2.$$

Therefore, using the inequality of arithmetic and geometric mean, $\det(A_t) \leq (1 + (t-1)\kappa^2/d)^d$. \square

Summing over all the suboptimal bins, we have that

$$\begin{aligned}
& \sum_{k=1}^{n-1} \sum_{s=1}^{N^k(T)} 2(\sqrt{\beta_{k,s}} \|A_{k,s}^{-1/2} x_{k,s}\| \wedge 1) \leq \sum_{k=1}^n \sqrt{8d\beta_{k,N^k(T)} N^k(T) \ln(1 + N^k(T))} \\
& \leq \sqrt{\sum_{k=1}^n N^k(T) \sum_{k=1}^n 8d\beta_{k,N^k(T)} \ln(1 + N^k(T))} \\
& = \sqrt{T \sum_{k=1}^n 8d\beta_{k,N^k(T)} \ln(1 + N^k(T))} \\
& \stackrel{N^k(T) \leq T}{\leq} \sqrt{8dTn\beta_T \ln(1 + T)}.
\end{aligned} \tag{2.29}$$

2.6.3.2 High probability regret bound part II (term #2)

Here we directly call previous result in bound (2.14), but replace the total number of step with $N^k(T)$, the number of pulls for one fixed bin k . We have for term #2,

$$\sum_{s=1}^{N^k(T)} 2\epsilon \sum_{\tau=1}^{s-1} |x_{k,s}^T A_{k,s}^{-1} x_{k,\tau}| \leq 2\epsilon N^k(T) d \sqrt{2 \ln(1 + N^k(T))}.$$

Summing over all suboptimal bins, we have that

$$\begin{aligned}
& \sum_{k=1}^n 2\epsilon N^k(T) d \sqrt{2 \ln(1 + N^k(T))} \\
& \stackrel{N^k(T) \leq T}{\leq} 2\epsilon d \sqrt{2 \ln(1 + T)} \sum_{k=1}^n N^k(T) \\
& = 2\epsilon d T \sqrt{2 \ln(1 + T)}.
\end{aligned} \tag{2.30}$$

2.6.3.3 Putting it together

Combining the decomposition in equation (2.27) and the results in subsections 2.6.3.1 and 2.6.3.2, we have a high probability regret bound for the UCB-Meta-algorithm:

$$\begin{aligned}
R_T &= \sum_{k=1}^n R_k \\
&\leq \sqrt{8dTn\beta_T \ln(1 + T)} + 2\epsilon d T \sqrt{2 \ln(1 + T)} + 2\epsilon T \\
&= \mathcal{O}(d \ln(T) \sqrt{Tn \ln(T^2 n / \delta)}) + \epsilon d T \sqrt{\ln(T)} + \epsilon T.
\end{aligned} \tag{2.31}$$

The last step plugs in $\beta_T = \mathcal{O}(d \ln(T) \ln(T^2 n / \delta))$.

□

2.6.3.4 Proof of Theorem 5

Proof. Algorithm 3 executes Algorithm 2 for a sequence of pre-defined time periods, $\{T_i = 2^i, i = 0, 1, \dots, N\}$. At the beginning of each period, the update history is cleared and the number of arms n is

reset with respect to the current horizon T_i . However, since we would like to acquire a high-probability regret bound after applying the doubling trick, we need to set the fail probability of Meta-algorithms during period i to $\delta_i = 6\delta/\pi^2 i^2$. Using a union bound, we can conclude the following ($R_i(T_i)$ denotes the regret incurred in time period i of length T_i only).

$$\begin{aligned} & \mathbb{P}(\forall i, \text{ the bound hold for } R_i(T_i)) \\ &= 1 - \sum_i \mathbb{P}(\text{the bound does not hold for } R_i(T_i)) \\ &= 1 - \sum_i \frac{6\delta}{\pi^2 i^2} \approx 1 - \delta. \end{aligned}$$

In the last step we use the fact that the sum of sequence $\sum_i^\infty \frac{1}{i^2}$ converges to $\frac{\pi^2}{6}$.

Now, the total regret is simply a summation over i . The following holds with probability $1 - \delta$,

$$\begin{aligned} R(T) &\leq \sum_{i=1}^N R_i(T_i) \\ &\leq \sum_{i=1}^N \tilde{\mathcal{O}}(dT_i^a) = \tilde{\mathcal{O}}\left(d \sum_{i=1}^N 2^{ia}\right) \\ &\leq \tilde{\mathcal{O}}\left(d2^{a(N-1)}\right) \\ &= \tilde{\mathcal{O}}(dT^a). \end{aligned} \tag{2.32}$$

At step 4, the number of time periods N is the smallest integer such that $\sum_{i=0}^N 2^i \geq T$, so $N = 1 + \lceil \log_2(T) \rceil$. The sum of geometric sequence is $2^{a \lceil \log_2(T) \rceil} = (2^{\log_2(T)+c})^a = T^a 2^{ca}$ for some constant c smaller than 1. Also, note that step 2 holds even though the fail probability is changed to $\delta_i = 6\delta/\pi^2 i$ is because as specified in Theorem 4, the term δ appears in a log term and the maximum value of $1/\delta$ is $1/\delta_N = \pi^2 \log_2(T)/6\delta$, therefore the extra factor caused by smaller δ to the regret is still a log term of T_i and omitted in the proof here.

Bound (2.32) suffices to say that meta-algorithm with doubling trick has the same regret rate as meta-algorithm with known horizon, with some additional constant factors suffered from restarting. \square

2.6.4 Proof of Theorem 6

Proof. Here we prove that Corral with smooth-wrapper is applicable to this task and achieves minimax expected regret rate apart from log factors. We directly use the proof of Theorem 5.3 in [Pacchiano et al. \[2020b\]](#) and their notations. δ is the fail probability, M is the number of base-algorithms, ρ is the reciprocal of the smallest possibility for base-algorithms over the T rounds and η is the learning rate. $U(T, \delta)$ is the high probability bound of the selected base-algorithm. The regret of Corral with smooth wrapper is bounded by:

$$R(T) \leq \mathcal{O}\left(\frac{M \ln(T)}{\eta} + T\eta\right) + \delta T + 8\sqrt{MT \log\left(\frac{4TM}{\delta}\right)} - \mathbb{E}\left[\frac{\rho}{40\eta \ln(T)} - 2\rho U(T/\rho, \delta) \log(T)\right], \tag{2.33}$$

and we know from Theorem 3 in our paper that the base algorithm (Algorithm 1) that locates in the global maximum's bin has anytime high probability regret bound $U(T, \delta) = \tilde{\mathcal{O}}(\epsilon T d(\alpha) + c(\delta)d(\alpha)\sqrt{T})$,

note that this is because the dimension of the local linear parameter is $d(\alpha)$. Therefore,

$$\begin{aligned} R(T) &= \tilde{\mathcal{O}}(\sqrt{MT} + \frac{M}{\eta} + T\eta) + \delta T - \mathbb{E}[\frac{\rho}{40\eta \ln(T)} - 2\rho\tilde{\mathcal{O}}(d(\alpha)\sqrt{T/\rho} + \frac{\epsilon d(\alpha)T}{\rho})] \\ &= \tilde{\mathcal{O}}(\sqrt{MT} + \frac{M}{\eta} + T\eta) + \delta T + \tilde{\mathcal{O}}(\epsilon T d(\alpha)) + \mathbb{E}[\tilde{\mathcal{O}}(d(\alpha)\sqrt{T\rho} - \frac{3\rho}{\eta})]. \end{aligned} \quad (2.34)$$

Firstly, we set $\delta = 1/T$ so that $\delta T = \mathcal{O}(1)$. Then we maximize this formulation over ρ by setting $\rho = \tilde{\mathcal{O}}(\eta^2 d(\alpha)^2 T)$, yielding the following bound on expected regret.

$$\begin{aligned} &\tilde{\mathcal{O}}(\sqrt{MT} + \frac{M}{\eta} + T\eta + \epsilon T d(\alpha) + \eta d(\alpha)^2 T) \\ &\stackrel{\substack{M=n, \\ \epsilon=n^{-\frac{\alpha}{d}}}}{=} \tilde{\mathcal{O}}(\sqrt{nT} + \frac{n}{\eta} + n^{-\frac{\alpha}{d}} T d(\alpha) + \eta d(\alpha)^2 T). \end{aligned} \quad (2.35)$$

We minimize this by setting the derivative w.r.t n and η to zero, i.e. $\eta = \frac{1}{d(\alpha)} \sqrt{\frac{n}{T}}$ and $n = \tilde{\mathcal{O}}(T^{\frac{d}{d+2\alpha}})$. As a result the rate comes to $\tilde{\mathcal{O}}(d(\alpha)T^{\frac{d+\alpha}{d+2\alpha}})$. \square

2.6.5 Proof of Lemma 7

Proof. According to Theorem 4, the algorithm sets $n = T^{\frac{d}{d+2\alpha'}} / \ln(T)^{\frac{2d}{d+2\alpha'}}$ and $\epsilon = n^{-\frac{\alpha'}{d}}$. Note that we can only use the result in Theorem 4 if the high probability upper confidence bound defined in line 4 of sub-procedure Algorithm 2 holds honestly. If the input parameter α' is larger than α , then the calculated misspecification error ϵ is smaller than the true $\epsilon^* = \tilde{\mathcal{O}}(T^{\frac{\alpha}{d+2\alpha}})$, causing the confidence bound to be invalid. Therefore, the regret bound does not hold for when $\alpha' > \alpha$. If the input parameter is smaller than α , then we can simply use the fact that functions that are α -Hölder smooth are also α' -Hölder smooth: $H(\alpha, L) \subset H(\alpha', L)$. Therefore, the regret of the algorithm with input parameter $\alpha' \leq \alpha$ is bounded by $R(T) \leq \tilde{\mathcal{O}}(d(\alpha')(\sqrt{Tn} + \epsilon T)) = \tilde{\mathcal{O}}(d(\alpha')T^{\frac{d+\alpha'}{d+2\alpha'}})$. \square

2.6.6 Proof of Theorem 8

Proof. There exists an $\hat{\alpha} \in \mathcal{G}$, s.t. $\hat{\alpha} \leq \alpha \leq \hat{\alpha} + \frac{R}{\log(T)}$, for any true α in $(0, R]$. There are two sources that made up the cost of adaptation when using Corral. The first one is the cost of searching over a grid for the unknown point $\hat{\alpha}$. The second one is the cost of approximation, specifically the difference between the rates achieved for $\hat{\alpha}$ and the true α . We will first derive the cost of grid search.

As specified in the proof of Theorem 5.3 in Pacchiano et al. [2020b], the following bound of regret of the Corral algorithm holds with respect to any of its base-algorithm with high probability regret bound $U(T, \delta)$. The notations were introduced in Appendix section 2.6.4.

$$R(T) \leq \mathcal{O}(\frac{M \ln(T)}{\eta} + T\eta) - \mathbb{E}[\frac{\rho}{40\eta \ln(T)} - 2\rho U(T/\rho, \delta) \log(T)] + \delta T + 8\sqrt{MT \log(\frac{4TM}{\delta})}. \quad (2.36)$$

Plugging the regret rate of base-algorithm in Lemma 7, the expected pseudo-regret of Corral with

smooth wrapper is therefore bounded by:

$$\begin{aligned}
R(T) &\stackrel{\hat{\alpha} \leq \alpha}{\leq} \tilde{\mathcal{O}}\left(\frac{M}{\eta} + T\eta + \sqrt{MT}\right) + \delta T - \mathbb{E}\left[\frac{\rho}{40\eta \ln(T)} - 2\rho\left(\tilde{\mathcal{O}}\left(d\left(\frac{T}{\rho}\right)^{\frac{d+\hat{\alpha}}{d+2\hat{\alpha}}}\right)\right) \log(T)\right] \\
&\stackrel{\delta=1/T}{=} \tilde{\mathcal{O}}\left(\frac{M}{\eta} + T\eta + \sqrt{MT}\right) - \mathbb{E}\left[\tilde{\mathcal{O}}\left(\frac{\rho}{\eta} - \rho d\left(\frac{T}{\rho}\right)^{\frac{d+\hat{\alpha}}{d+2\hat{\alpha}}}\right)\right] \\
&= \tilde{\mathcal{O}}\left(\frac{M}{\eta} + T\eta + \sqrt{MT}\right) - \mathbb{E}\left[\tilde{\mathcal{O}}\left(\frac{\rho}{\eta} - dT^{\frac{d+\hat{\alpha}}{d+2\hat{\alpha}}}\rho^{\frac{\hat{\alpha}}{d+2\hat{\alpha}}}\right)\right].
\end{aligned} \tag{2.37}$$

Similarly, we first maximize over ρ by setting the derivative w.r.t ρ to zero by setting $\rho = \tilde{\mathcal{O}}\left(\eta^{\frac{d+2\hat{\alpha}}{d+\hat{\alpha}}} d^{\frac{d+2\hat{\alpha}}{d+\hat{\alpha}}} T\right)$. Then the above rate comes to

$$R(T) \leq \tilde{\mathcal{O}}\left(\frac{M}{\eta} + T\eta + \sqrt{MT} + d^{\frac{d+2\hat{\alpha}}{d+\hat{\alpha}}} T\eta^{\frac{\hat{\alpha}}{d+\hat{\alpha}}}\right). \tag{2.38}$$

However, since η is a parameter of the Corral algorithm which does not know $\hat{\alpha}$ or α , we will rely on the parameter R specified by the user. Let us set η with respect to $\alpha = R$, i.e. $\eta = \tilde{\mathcal{O}}\left(d^{-1}T^{-\frac{d+R}{d+2R}}\right)$, and plug in the number of grid points (base-algorithms) $M = \lceil \log(T) \rceil$.

$$\begin{aligned}
&\tilde{\mathcal{O}}\left(\frac{M}{\eta} + T\eta + \sqrt{MT} + d^{\frac{d+2\hat{\alpha}}{d+\hat{\alpha}}} T\eta^{\frac{\hat{\alpha}}{d+\hat{\alpha}}}\right) \\
&= \tilde{\mathcal{O}}\left(dT^{\frac{d+R}{d+2R}} + d^{-1}T^{\frac{R}{d+2R}} + dT^{\frac{d^2+2Rd+R\hat{\alpha}}{(d+2R)(d+\hat{\alpha})}}\right) \\
&= \tilde{\mathcal{O}}\left(dT^{\frac{d+R}{d+2R}} + dT^{\frac{d^2+2Rd+R\hat{\alpha}}{(d+2R)(d+\hat{\alpha})}}\right).
\end{aligned} \tag{2.39}$$

It is obvious that this rate is not the minimax optimal rate for class $\sum(\hat{\alpha})$, this gap shows the cost of grid search.

Next, let us consider the cost of approximation and how it is eliminated by using the linear grid (Hoffmann et al. [2011]). Namely, we show that adaptation for $\hat{\alpha}$ is equivalent to adaptation for α :

$$\tilde{\mathcal{O}}\left(dT^{\frac{d+R}{d+2R}} + dT^{\frac{d^2+2Rd+R\hat{\alpha}}{(d+2R)(d+\hat{\alpha})}}\right) = \tilde{\mathcal{O}}\left(dT^{\frac{d+R}{d+2R}} + dT^{\frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)}}\right). \tag{2.40}$$

The equality holds because $|\alpha - \hat{\alpha}| \leq \frac{R}{\log(T)}$. Let $J = \frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)}$ and $Q = \frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)}$, then $W \triangleq \frac{T^J}{T^Q} \leq T^{\frac{(d^2+2Rd+R\alpha) \frac{R}{\log(T)}}{(d+2R)(d+\alpha)(d+\hat{\alpha})}}$. Taking the log of W yields $\log(W) = R \frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)(d+\hat{\alpha})}$. Since both α and $\hat{\alpha}$ are bounded by a constant range $(0, 2]$, the term $\frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)(d+\hat{\alpha})} \leq C$ for some constant C , W is therefore $\mathcal{O}(1)$ as well.

Therefore, for functions with Hölder exponent $\alpha < R$, the second term in equation (2.40) is the dominant term and the expected regret rate is $\tilde{\mathcal{O}}\left(dT^{\frac{d^2+2Rd+R\alpha}{(d+2R)(d+\alpha)}}\right)$. For functions with Hölder exponent $\alpha \geq R$, which essentially belongs to a subset of $\sum(R, L)$, they will all have the same rate which is $\tilde{\mathcal{O}}\left(dT^{\frac{d+R}{d+2R}}\right)$. When $\alpha = R$, this matches the minimax rate for α . \square

2.7 The Doubling Procedure for Algorithm 2

Algorithm 3 Doubling procedure for Algorithm 2

Require: Meta-algorithm \mathcal{A}^{global} (Algorithm 2), fail probability δ

- 1: **for** $i = 0 \dots$ **do**
 - 2: $T_i = 2^i$
 - 3: Restart \mathcal{A}^{global} with initialization parameters $n_i = \lfloor T_i^{\frac{d}{d+2\alpha}} / \ln(T_i)^{\frac{2d}{d+2\alpha}} \rfloor$ and fail probability $\delta_i = 6\delta / \pi^2 i^2$
 - 4: Run \mathcal{A}^{global} for T_i steps
 - 5: **end for**
-

Chapter 3

Adaptation Sample Complexity: Unknown Kernel Regularity in Kernelised Bandits

This chapter is based on [Liu and Singh \[2023\]](#).

3.1 Introduction

Recall that in the continuum-armed bandit setting, the performance of algorithms is measured by the cumulative regret (equation 2.1) which is the sum of differences between the maximum of the underlying function and the reward incurred by the learning algorithm across all the time steps. As mentioned in Chapter 1, optimizing cumulative regret requires a delicate exploration-exploitation tradeoff. The algorithm needs to simultaneously exploit high-reward regions and explore uncertain regions. The exploration-exploitation tradeoff is often dependent on complexity of the function space to which the reward function belongs. In most theoretical analyses of cumulative regret of algorithms, complexity of the function space is assumed to be known. Many studies use this assumption to design algorithms that achieve minimax optimal performance when the function space is known, for example, for linear functions [[Dani et al., 2008](#), [Abbasi-Yadkori et al., 2011](#)], functions residing in reproducing kernel Hilbert spaces (RKHS) [[Valko et al., 2013](#), [Janz et al., 2020](#)] or drawn from Gaussian Processes [[Srinivas et al., 2009](#), [Chowdhury and Gopalan, 2017](#)], as well as neural networks functions [[Zhou et al., 2020](#), [Kassraie and Krause, 2021](#)].

However, despite the theoretical convenience, it is not always realistic to assume access to the underlying function space. For this reason, some recent works in continuum-armed bandits have started to develop adaptive algorithms for when the function space is misspecified (see Section 3.2 for a summary of related works). The best possible performance of adaptive algorithms is equivalent to algorithms that know the parameter. An algorithm that simultaneously achieves minimax cumulative regret rates without access to the parameter is said to achieve minimax adaptivity. While minimax adaptivity is

possible under the simple regret minimization setting, recent works have proved that it is not always achievable for cumulative regret minimization [Locatelli and Carpentier, 2018], due to the exploration-exploitation dilemma.

When the reward function resides in an RKHS induced of some kernel function k , the problem also is referred to as kernelised bandit. In this work, we focus on an important and open problem in adaptivity in kernelised bandits, precisely, adaptivity to unknown kernel regularity. Recently, there has been a line of theoretical works that study adaptivity under the kernelised bandit setting, such as adaptivity to the length scale parameter and the RKHS norm [Berkenkamp et al., 2019] for a given kernel, and adaptivity to ϵ -misspecification, where the underlying function is ϵ -approximated by functions in an RKHS [Bogunovic and Krause, 2021]. To the best of our knowledge, the work of Kassraie et al. [2022] is most closely related to our setting. They consider the setting where the underlying function lies in an RKHS but the kernel is unknown. Kassraie et al. [2022] assume that the kernel is a sparse combination of known base kernels and design algorithms with sublinear regret guarantees under this assumption. A more detailed discussion of the prior works on adaptivity in kernelised bandit is continued in Section 3.2.

Adaptivity of any algorithm with respect to the explicit regularity of the kernel function k , however, remains an unsolved problem. We characterize the regularity of k using a general notion: the decay rate of the Fourier transform of k (Section 3.3). In contrast to, for example, adapting to the RKHS norm which measures the smoothness of a function with respect to a fixed kernel, we adapt to the regularity of kernels which controls the differentiability of functions in the associated RKHSs. The kernel regularity thus determines the statistical complexity of the associated learning problem in a more fundamental way. In estimation, optimization (including simple regret minimization) [Bull, 2011] and cumulative regret minimization tasks [Srinivas et al., 2009, Kandasamy et al., 2019, Janz et al., 2020], the kernel regularity affects the minimax regret rate through exponential dependence on T , as opposed to the RKHS norm which only affects the rate polynomially. We focus on this fundamental problem of how well bandit algorithms can adapt to the unknown kernel regularity.

The contributions of this work are summarized as follows:

1. We derive the first lower bound on adaptivity to kernel regularity, expressed in terms of the kernel Fourier transformation decay rate, for kernelised bandit problems. This lower bound serves as an impossibility result, that no algorithms can simultaneously achieve minimax optimal performance in RKHSs with different regularities.
2. For RKHSs of the Matérn family [Matern et al., 1960] of kernels, we prove that CORRAL [Pacchiano et al., 2020b], an existing model selection algorithm, applied with (non-adaptive) minimax optimal kernelised bandit algorithms, matches the adaptivity lower bound¹ in the dependence on T . In contrast, another model selection algorithm RBBE Pacchiano et al. [2020a] does not match the lower bound.
3. By comparing the upper and lower bounds derived by this work to existing adaptivity results, we draw connections between the statistical difficulty of adaptivity in three types of function spaces: RKHSs, Sobolev spaces, and Hölder spaces.

A summary of our results amongst existing results can be found in Table 3.1. Our main results

¹Except for log factors.

Table 3.1: Summary of Our Results and Comparison to Existing Results

Regret		RKHS of Matern- ν : $\mathcal{H}_{k,\nu}(\mathcal{X}) = \mathcal{W}^{\nu+\frac{d}{2}}(\mathcal{X})$	Hölder Space: $\Sigma^\alpha(\mathcal{X})$
		Relationship: $\mathcal{H}_{k,\nu}(\mathcal{X}) = \mathcal{W}^{\nu+\frac{d}{2}}(\mathcal{X}) \subset \Sigma^{\alpha=\nu}(\mathcal{X})$	
Non-adaptive minimax		$\tilde{\Theta}(T^{\frac{\nu+d}{2\nu+d}})$ Valko et al. [2013], Scarlett et al. [2017]	$\tilde{\Theta}(T^{\frac{d+\alpha}{d+2\alpha}})$ Liu et al. [2021], Wang et al. [2018]
Adaptive ($d = 1$)	Upper bound	$\tilde{O}(T^{\frac{1+2\tilde{\nu}+\nu\nu}{(1+2\tilde{\nu})(1+\nu)}})$, for $\nu < \tilde{\nu}$ $\tilde{\nu}$: Input to adaptive algorithm. This work (Theorem 18)	$\tilde{O}(T^{\frac{1+2R+R\alpha}{(1+2R)(1+\alpha)}})$, for $\alpha < R$ R : Input to adaptive algorithm. Liu et al. [2021, Theorem 8]
	Lower bound	$\tilde{\Omega}(T^{\frac{1^2+2\tilde{\nu}+\tilde{\nu}\nu}{(1+2\tilde{\nu})(1+\nu)}})$, for $\nu < \tilde{\nu}$ This work (Corollary 16)	$\tilde{\Omega}(T^{\frac{1^2+2R+R\alpha}{(1+2R)(1+\alpha)}})$, for $\alpha < R \leq 1$ Locatelli and Carpentier [2018, Theorem 3]

(Section 3.4.2) are stated for more general kernels but in Table 3.1 only results with Matérn- ν (Definition 15) kernels are shown as an example, for clear comparisons. For adaptive results, the values $\tilde{\nu}$ and R are input parameters to the adaptive algorithms, such that they achieve (non-adaptive) minimax regret rates if the true parameter satisfies $\nu = \tilde{\nu}$ (for Matérn RKHS) or $\alpha = R$ (for Hölder spaces). We use \tilde{O} to denote the asymptotic regret rate of T . \tilde{O} omits dependence on other parameters such as the radius of the RKHS ball B (Section 3.3), any constant factors, and \log factors of T unless otherwise specified.

Relationship with Neural Bandits.

The kernelised bandit formulation has implications for optimization of more complex functions under the bandit setting as well, such as neural network functions. The Neural Tangent Kernel (NTK) literature [Jacot et al., 2018, Arora et al., 2019, Lee et al., 2019, Bietti and Bach, 2020, Chen and Xu, 2020] argue that over-parameterized neural networks can be approximated by functions in an RKHS of some composite kernel named the Neural Tangent Kernel, given that the network is sufficiently wide and the training is lazy [Chizat et al., 2019]. Recent advances in this field establish interesting connections between the structure of a neural network and the regularity of its corresponding NTK. For example, Vakili et al. [2021a] consider wide fully-connected neural networks with activation functions with smoothness s . They show that the RKHS of the NTK of such a network is norm equivalent to, or embedded in, the RKHS of a Matérn- ν kernel with $\nu = s - \frac{1}{2}$. The value of ν dictates the differentiability of functions in the RKHS. Hence, the neural network functions considered in Vakili et al. [2021a] are approximated by functions in the RKHS of a Matérn- ν kernel.² These connections imply that adaptivity to the kernel regularity in kernelised bandits can potentially be extended to adaptivity to the structure of neural networks (such as smoothness of the activation functions considered in Vakili et al. [2021a]) in neural network bandits.

The rest of the paper is structured as follows. In Section 3.2, we discuss relevant prior works. In Section 3.3 we state the problem formulation. In Section 3.4 we present the main result of this paper, a lower bound on adaptivity to kernel regularity. In Section 3.5 we discuss upper bounds of existing adaptive algorithms and whether they match the lower bound. In Section 3.6 we connect adaptivity

²The result in Corollary 3 of Bietti and Bach [2020] can be thought as a special case of when $s = 1$, since the activation function considered is ReLU.

to kernel regularity and adaptivity to Hölder exponents. Finally, we discuss the limitations and future directions of our work in Section 3.7.

3.2 Related Work

Kernelised Bandit

In kernelised bandit problems, the reward function lies in a reproducing kernel Hilbert space (RKHS). This problem has been studied by many previous works, under the assumption that the kernel and other parameters (such as the upper bound on the function’s RKHS norm) are known. Valko et al. [2013] take a frequentist approach and design a SuperKernelUCB algorithm, based on applying the kernel trick to the (Sup)LinREL and (Sup)LinUCB algorithms [Auer, 2002, Chu et al., 2011]. The same technique is later used in extension to neural networks by Kassraie and Krause [2021], who propose SupNTKUCB which works with neural networks. SupKernelUCB achieves $\tilde{O}(\sqrt{T\gamma_T})$ regret where γ_T is the maximum information gain between T total observations and the underlying function. For common kernels such as the Matérn- ν kernels, this regret is minimax optimal in its dependence on T (except for log factors), by the lower bound provided later in Scarlett et al. [2017]. However, SupKernelUCB relies on a batching technique that makes the algorithm performs poorly in practice [Calandriello et al., 2019, Janz et al., 2020]. In the (parallel) Bayesian setting (the Gaussian Process bandit problem), the underlying function is assumed to be drawn from a GP. GP-UCB algorithm [Srinivas et al., 2009, Chowdhury and Gopalan, 2017, Janz et al., 2020] achieves the same regret bound as SupKernelUCB $\tilde{O}(\sqrt{T\gamma_T})$ in the GP setting but becomes suboptimal (sometimes with linear regret rate) in the RKHS setting with a $\tilde{O}(\gamma_T\sqrt{T})$ regret [Vakili et al., 2021b].

Adaptivity in Kernelised Bandit

This problem we consider falls under the scope of model misspecification in bandit setting, which has been studied for linear functions and Hölder-smooth functions [Du et al., 2019, Foster et al., 2019, Lattimore et al., 2020, Zhu and Nowak, 2021, Locatelli and Carpentier, 2018, Liu et al., 2021]. For Hölder functions, in particular, Locatelli and Carpentier [2018], Hadiji [2019] provide a lower bound indicating that it is impossible to achieve minimax adaptivity to the Hölder exponent. In this work, we convey a similar message with respect to the regularity of RKHS. For adaptivity in kernelised bandit problems, Berkenkamp et al. [2019] propose an algorithm with sublinear regret for when the lengthscale parameter (Definition 15) and upper bound on the RKHS norm (equation 3.4) are unknown. Neiswanger and Ramdas [2021] develop robust confidence sequence under the Bayesian framework to use in adaptive methods for GP optimization when the prior mean and/or covariance parameters are unknown. They conduct simulations for optimization on functions drawn from GPs but do not provide explicit regret analyses. Bogunovic and Krause [2021] develop methods for ϵ -misspecification, where the underlying function can be arbitrarily non-smooth, but is approximated by functions in a (known) RKHS with an ϵ -error in infinity norm. They prove a $\Omega(\epsilon T)$ lower bound for this setting and derived a matching upper bound. However, note that between two function spaces, the approximation error is a constant value and does not depend on T . Since a constant ϵ means an inevitable linear regret ($\Omega(\epsilon T)$), the ϵ -misspecification setting [Bogunovic and Krause, 2021] does not directly apply

to adaptation to the kernel parameters. In the Meta-learning regime, [Kassraie et al. \[2022\]](#) consider RKHS with unknown kernels that are sparse combinations of known base kernels and proves that a Meta-learned kernel can yield sublinear regret. However, since the kernel is Meta-learned, it relies on offline tasks as training data. We do not assume the availability of offline data in the (fully online) bandit setting.

To summarize, prior works (to the best of our knowledge) only consider parameters that influence the regret rate in polynomial factors while our focus is on the regularity parameter which affects the rate in the exponent of T .

General Model Selection for Bandit

Another line of recent works on model selection in bandit settings makes less stringent assumptions on the underlying function. Model selection algorithms are previously discussed in Section 2.1 as well. A prominent type of such algorithms are based on a “corralling” mechanism, where a master algorithm “corrals” several base algorithms as arms and each base algorithm selects actions with different principles. In this chapter we focus on such corralling algorithms. The base algorithms usually assume different function spaces. [Agarwal et al. \[2017\]](#), [Pacchiano et al. \[2020b\]](#) propose an algorithm named CORRAL where the master algorithm is based on online mirror descent. In certain cases, CORRAL performs comparably to the best base algorithm running standalone.³ [Pacchiano et al. \[2020a\]](#) propose the Regret Bound Balancing and Elimination (RBBE) which uses a (simpler) stochastic master algorithm and an additional base-algorithm-elimination step. We refer readers to Section 3.5 for details about these two methods and their performance in our problem setting.

3.3 Problem Setting

Problem Formulation

Again, consider the stochastic continuum-armed bandit problem. At time step $t \in \{1, \dots, T\}$, the learner chooses an action x_t from the compact domain $\mathcal{X} = [0, 1]^d$, and receives a reward y_t . The reward is a noisy observation of the underlying reward function $f : \mathcal{X} \rightarrow \mathbb{R}$:

$$y_t = f(x_t) + \eta_t, \quad (3.1)$$

where the noise variable η_t follows a zero-mean sub-Gaussian distribution (see Theorem 14). The optimization objective is the cumulative (pseudo-)regret (equation 2.1) is restated as follows.

$$R_T = \sum_{t=1}^T f(x^*) - f(x_t), \quad (3.2)$$

where x^* is the global maximizer of f , unknown to the learner. Results in this paper are expressed in expected cumulative (pseudo-)regret $\mathbb{E}[R_T]$, where the expectation is taken over the randomness of $\{x_t\}_{t=1 \dots T}$.

³[Arora et al. \[2021\]](#) also study the problem of corralling bandit algorithms in the stochastic setting, but only finite-armed case is considered.

Kernelised Bandit

We consider the setting where f is square-integrable and resides in an RKHS \mathcal{H}_k of a symmetric, positive-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. The RKHS is unique given the kernel [Wainwright, 2019, Theorem 12.11]. We denote the RKHS of k on domain \mathcal{X} as $\mathcal{H}_k(\mathcal{X})$. In this work, we restrict our attention to *translation-invariant* kernels, precisely, kernels that satisfy the following: $k(x, x') = \kappa(x - x')$, for some function $\kappa : \mathbb{R}^d \rightarrow \mathbb{R}$. For a translation-invariant kernel, the regularity of functions in the RKHS is captured by the Fourier transform of the kernel. Precisely, we have the following definition when the domain is \mathbb{R}^d . Let $\hat{g}(\omega)$ denote the Fourier transformation [Wendland, 2004, Williams and Rasmussen, 2006] of a function g as $\forall \omega \in \mathbb{R}^d$.

$$\begin{aligned} \mathcal{H}_k(\mathbb{R}^d) &= \{f \in \mathcal{L}_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \\ \|f\|_{\mathcal{H}_k} &:= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\hat{f}(\omega)|^2}{\hat{\kappa}(\omega)} d\omega < \infty\}. \end{aligned} \quad (3.3)$$

When the domain \mathcal{X} is a subset of \mathbb{R}^d , $\hat{\kappa}$ still captures the regularity of $\mathcal{H}_k(\mathcal{X})$, via a norm equivalency result that holds as long as \mathcal{X} has a Lipschitz boundary. Details can be found in Section 3.4.1, Lemma 12. We write $\|f\|_k \triangleq \|f\|_{\mathcal{H}_k(\mathcal{X})}$ for simplicity. We apply the common assumption [Srinivas et al., 2009, Valko et al., 2013] that the RKHS norm of f is upper bounded by a value B , $0 < B < \infty$:

$$f \in \mathcal{H}_k(\mathcal{X}, B) := \{f : f \in \mathcal{H}_k, \|f\|_k \leq B\}. \quad (3.4)$$

We refer to $\mathcal{H}_k(\mathcal{X}, B)$ as a ball in the RKHS with radius B .

3.4 Main Result: Adaptivity Lower Bound

In this section, we present the main result, a lower bound on adaptivity to the regularity of kernel (Theorem 14). The regularity of a translation-invariant kernel is expressed as the decay rate of its Fourier transformation (equation 3.9). We next instantiate this idea with a norm equivalency result between an RKHS and a Sobolev space. The norm equivalency result is dependent on the kernel Fourier decay rate. The proof of Theorem 14, in turn, relies on this norm equivalency as well.

3.4.1 Norm Equivalency Between RKHS and Sobolev Space

Consider integer-order Sobolev space $\mathcal{W}^{m,p}(\mathcal{X})$ where m, p are integers greater or equal to 1. We define the following notions for a multi-index vector $\alpha = (\alpha_1 \dots \alpha_d)$: $|\alpha| = \alpha_1 + \dots + \alpha_d$, $\alpha! = \alpha_1! \dots \alpha_d!$ and $x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$. Let $D^{(\alpha)} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ denote the multivariate mixed partial weak derivative. The Sobolev space and corresponding Sobolev norm ($\|\cdot\|_{m,p,\mathcal{X}}$) are defined as follows.

$$\mathcal{W}^{m,p}(\mathcal{X}) = \{f \in \mathcal{L}_p(\mathcal{X}) : D^{(\alpha)} f \in \mathcal{L}_p(\mathbb{R}^d), \forall |\alpha| \leq m\}, \quad (3.5)$$

$$\|f\|_{m,p,\mathcal{X}} := \left(\sum_{|\alpha| \leq m} \int |D^{(\alpha)} f(x)|^p dx \right)^{\frac{1}{p}}. \quad (3.6)$$

We refer to m as the order of the Sobolev space. Furthermore, define the j -th order seminorm [Adams and Fournier, 2003, Definition 4.11] $|\cdot|_{j,p,\mathcal{X}}$ with integer $j \leq m$, which is the sum of \mathcal{L}_p norm of its j -th

weak derivatives.

$$|f|_{j,p,\mathcal{X}} = \left(\sum_{|\alpha|=j} \int |D^{(\alpha)} f(x)|^p dx \right)^{\frac{1}{p}}. \quad (3.7)$$

In correspondence to the RKHS ball (equation 3.4), we define a Sobolev ball with radius L as the set of functions whose m -th order *seminorm* are upper bounded by L .

$$\mathcal{W}^{m,p}(\mathcal{X}, L) = \{f \in \mathcal{W}^{m,p}(\mathcal{X}) : |f|_{m,p,\mathcal{X}} \leq L\}. \quad (3.8)$$

When $p = 2$, the Sobolev space is equivalent to the RKHS of a translation-invariant kernel k . This connection plays an important role in the analysis. We consider only Sobolev spaces with $p = 2$, and hence abbreviate $\mathcal{W}^m(\mathcal{X}) \triangleq \mathcal{W}^{m,2}(\mathcal{X})$. The precise norm equivalency is introduced in the following lemma.

Lemma 12. *Wendland [2004, Corollary 10.48] Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a translation-invariant kernel function such that $k(\cdot, \cdot) = \kappa(\cdot - \cdot)$ for $\kappa \in \mathcal{L}_1(\mathbb{R}^d)$. Suppose $\Omega \in \mathbb{R}^d$ is a domain with Lipschitz boundary. Suppose $\hat{\kappa}$ has the following polynomial decay rate of s , for $s > d/2$, $s \in \mathbb{N}$,*

$$c_1(1 + \|\omega\|_2^2)^{-s} \leq \hat{\kappa}(\omega) \leq c_2(1 + \|\omega\|_2^2)^{-s}, \forall \omega \in \mathbb{R}^d, \quad (3.9)$$

for some constants $0 < c_1 \leq c_2$. Then, the associated RKHS $\mathcal{H}_k(\Omega)$ is norm equivalent to the Sobolev space $W^m(\Omega)$ with $m = s$.

Having established the equivalency between RKHS and Sobolev spaces, we further introduce some notions to quantify the relationship between Sobolev seminorm (which is the radius of Sobolev balls) and RKHS norm in the following lemma.

Lemma 13. *Suppose that m is a positive integer larger than $d/2$. Let Ω be a finite-width domain with Lipschitz boundary. Let $\mathcal{W}_0^{m,p}(\Omega)$ denote the closure of $C_0^\infty(\Omega)$ (set of functions that have compact support in Ω and, together with their infinite order of partial derivatives, are continuous) in $\mathcal{W}^{m,p}(\Omega)$ Adams and Fournier [2003]. Then, the m -th Sobolev seminorm of f can be bounded by its RKHS norm with respect to a translation-invariant kernel k with Fourier decay rate m . Precisely,*

$$\underline{c}|f|_{m,2} \leq \|f\|_{\mathcal{H}_k} \leq \bar{c}|f|_{m,2}, \quad (3.10)$$

for some constants $0 < \underline{c} < \bar{c}$.

The constants \underline{c}, \bar{c} are used globally in this work and appear in the lower bound in Section 3.4.2. The proof of Lemma 13 can be found in Appendix 3.8.1.

3.4.2 Lower Bound on Adaptivity to Kernel Regularity

Theorem 14 presents our lower bound for adapting between a pair of RKHSs of different (kernel) regularities. An intuitive interpretation of the theorem is as follows. Consider a nested pair of balls in two RKHSs. Suppose both kernels satisfy the conditions in Lemma 12 but with different Fourier decay rates: $m_1 \in \mathbb{N}$ and $m_2 \in \mathbb{N}$ such that $0 < m_1 < m_2$. If an algorithm that is oblivious to the true regularity value somehow achieves a small (for example, minimax optimal) regret on all functions inside the (smoother) RKHS ball with parameter m_2 , this algorithm will suffer a price of larger (suboptimal)

regret on at least one function inside the (rougher) RKSH ball with parameter m_1 . For the lower bound analysis, we consider $d = 1$ and leave the extension of the lower bound to $d > 1$ as a future direction (Section 3.7).

Theorem 14. *Consider the problem setting in Section 3.3 with noises $\{\eta_t\}_{t=1\dots T}$ that are $\frac{1}{4}$ -subgaussian. Let \tilde{R} be a positive number, let m_1, m_2 be two positive integers that satisfy $m_1 < m_2$. There exist two positive values B_1 and B_2 , such that the following statement is true. Consider an algorithm that achieves in the RKHS of a kernel k_{m_2} with Fourier decay rate m_2 the following regret upper bound.*

$$\sup_{f \in \mathcal{H}_{k_{m_2}}(\mathcal{X}, B_2)} \mathbb{E}[R_T] \leq \tilde{R}. \quad (3.11)$$

Then, the regret of this algorithm in a (less smooth) RKHS of another kernel k_{m_1} with Fourier decay rate m_1 is lower bounded by the following. Suppose that functions in the function spaces have bounded \mathcal{L}_2 norm.⁴

$$\sup_{f \in \mathcal{H}_{k_{m_1}}(\mathcal{X}, B_1)} \mathbb{E}[R_T] \geq \frac{1}{8} \left(\frac{C(m_1)}{32} \right)^{\frac{m_1-1/2}{m_1+1/2}} \left(\frac{B_1}{\bar{c}} \right)^{\frac{1}{m_1+1/2}} \tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}} T. \quad (3.12)$$

Here, $C(m_1)$ denotes a constant that depends on m_1 .

It is worth noting that, although the lower bound has a factor of T , the regret is not necessarily linear in T , because \tilde{R} also depends on T and in fact usually ranges from $\tilde{O}(\sqrt{T})$ to $\tilde{O}(T)$. The full version of this theorem is presented as Theorem 20 in Appendix 3.8.2, where we state the full constraints on the radius values B_1 and B_2 . Since B_1 and B_2 are only upper bounds on the RKHS norm and *not* the kernel regularity that we focus on, we present only the concise version here to show the adaptivity difficulty with respect to regularity parameters m_1 and m_2 .

3.4.2.1 Proof Sketch

The proof of Theorem 14 consists of two key parts. The first part is constructing the hypothesis functions, in which we borrow ideas from lower bounds in regression problems [Tsybakov, 2004]. The second part is lower bounding the cumulative regret, given the constructed hypothesis functions, where we follow Hadiji [2019, Section 2.2]. Intuitively, the second part shows that if any player achieves a small regret on all the smoother functions, then it inevitably incurs large regret on the rougher functions in the space, because of its disproportionately small amount of exploration. The method in Hadiji [2019] is itself an improved version of the adaptivity lower bound for Hölder spaces proposed in Locatelli and Carpentier [2018].

3.4.2.2 A Sobolev Version of the Lower Bound

It is convenient to construct functions with compact support and finite Sobolev semi-norms from an infinitely-differentiable base function, such as the bump function Tsybakov [2004]. On the other hand, directly constructing functions with finite RKHS norms [Scarlett et al., 2017, Section III.A] involves

⁴Functions in Sobolev spaces and RKHSs are square-integrable.

inverse Fourier transformation of the bump function and thus leads to wavelet-like functions with non-compact support. Therefore, it is more natural for us to first consider functions in (integer-order) Sobolev spaces as hypothesis functions, and then use the norm equivalency result between Sobolev spaces and RKHSs to prove the lower bound. More precisely, the hypothesis functions constructed in the proof reside in a Sobolev ball $\mathcal{W}^m(\mathcal{X}, L)$, for some (integer) order m and radius L . Via the norm equivalency (Lemma 13), those functions also resides in a RKHS ball of a kernel with Fourier decay rate m .

As a result, there is a Sobolev version of the adaptivity lower bound. Informally, let m_1, m_2 be two positive integers such that $m_2 > m_1$. If an algorithm achieves a \tilde{R} regret upper bound in the smoother Sobolev space $\mathcal{W}^{m_2}(\mathcal{X})$, then its regret over functions in $\mathcal{W}^{m_1}(\mathcal{X})$ is lower bounded by $\Omega(\tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}} T)$. We formally state the Sobolev version of the adaptivity lower bound in Theorem 22 in Appendix 3.9.1. The two lower bounds share the same proof structure, connected via the norm equivalency in Lemma 12.

3.4.2.3 Impossibility Result for Matérn Kernels

For the Matérn- ν family of kernels [Matern et al., 1960], an implication of Theorem 14 is that no algorithm can achieve minimax adaptivity between two RKHSs if they have different regularity. Therefore, we also refer to this lower bound as an impossibility result for adaptivity to the kernel regularity. We formally define Matérn- ν family of kernels in Definition 15.

Definition 15. The Matérn- ν kernel and its Fourier transformation are defined as follows for dimension d .

$$k_{\text{Matérn},\nu}(x, x') \tag{3.13}$$

$$= \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x - x'\|_2}{l} \right)^\nu J_\nu\left(\frac{\sqrt{2\nu}\|x - x'\|_2}{l}\right), \tag{3.14}$$

$$\hat{k}_{\text{Matérn},\nu}(\omega) = c_1 \left(\frac{2\nu}{l^2} + \|\omega\|_2^2 \right)^{-(\nu + \frac{d}{2})}. \tag{3.15}$$

where $c_1 = \frac{2^d \pi^{d/2} \Gamma(\nu + d/2) (2\nu)^\nu}{\Gamma(\nu) l^{2\nu}}$, J_ν is the modified Bessel function of the second kind, l is the length-scale, and $\nu > 0$ is the regularity parameter. In this work, we assume for simplicity that the length-scale is set to $\propto \sqrt{2\nu}$.

The Fourier transformation of a Matérn kernel with regularity parameter ν decays with a rate of $\nu + \frac{d}{2}$ (equation 3.15). Therefore, we can instantiate the impossibility result for Matérn kernels. The result is presented in Corollary 16. Precisely, for $0 < \nu_1 < \nu_2$, if an adaptive algorithm achieves minimax regret rate on a Matérn RKHS with regularity ν_2 , then it has a strictly suboptimal regret rate on the RKHS with ν_1 .

Corollary 16. *Suppose the problem is the same as defined in Theorem 14. Let ν_1, ν_2 be real numbers that satisfy $0 < \nu_1 < \nu_2$ and $\nu_1 + \frac{1}{2} \in \mathbb{N}, \nu_2 + \frac{1}{2} \in \mathbb{N}$. There exist two positive values B_1, B_2 , such that the following statement is true. Suppose an algorithm oblivious to the true regularity parameter value achieves the following minimax optimal regret ⁵ on $\mathcal{H}_{k_{\text{Matérn},\nu_2}}(\mathcal{X}, B_2)$,*

$$\sup_{f \in \mathcal{H}_{k_{\text{Matérn},\nu_2}}(\mathcal{X}, B_2)} \mathbb{E}[R_T] = \tilde{O}\left(T^{\frac{\nu_2+1}{2\nu_2+1}}\right), \tag{3.16}$$

⁵Omitting the dependence on the upper bound on RKHS norm.

then the regret of this algorithm on RKHS $\mathcal{H}_{k_{\text{Matérn},\nu_1}}(\mathcal{X}, B_1)$ is lower bounded by the following.

$$\sup_{f \in \mathcal{H}_{k_{\text{Matérn},\nu_1}}(\mathcal{X}, B_1)} \mathbb{E}[R_T] = \tilde{\Omega} \left(T^{\frac{\nu_1 \nu_2 + 2\nu_2 + 1}{(\nu_1 + 1)(2\nu_2 + 1)}} \right), \quad (3.17)$$

The proof of Corollary 16 is an application of Theorem 14 and can be found in Appendix 3.9.2. The cumulative regret rate in 3.17 is suboptimal compared to the minimax rate which is $\tilde{O}(T^{\frac{\nu_1 + 1}{2\nu_1 + 1}})$ (see Section 3.5.1 for non-adaptive minimax rates). Therefore, Theorem 14 is an impossibility result for adaptivity to kernel regularity with Matérn kernels.

3.5 Upper Bounds of Adaptive Algorithms

We consider two adaptive algorithms particularly: CORRAL from Agarwal et al. [2017], Pacchiano et al. [2020b] and Regret Bound Balancing and Elimination (RBBE) from [Pacchiano et al., 2020a]. The two algorithms (i) can be applied to the problem of adaptation to kernel regularity and (ii) have explicit regret guarantees in this setting.

The adaptive algorithms, however, need base algorithms that are non-adaptive minimax optimal. We first provide an overview of such non-adaptive algorithms for kernelised bandit in Section 3.5.1. Then, we derive adaptivity upper bounds of CORRAL and RBBE in Section 3.5.2 and Section 3.5.3 respectively. For concreteness, we only consider RKHS of Matérn- ν kernel (Definition 15) in this section. To match the lower bound, we set $d = 1$. Comparison of the upper bounds to the lower bound (Theorem 14), shows that CORRAL (coupled with minimax optimal base algorithms) can match the lower bound in dependence on T between certain pairs of values for ν .

3.5.1 Overview: Non-adaptive Minimax Algorithms

We discuss the theoretical performance of algorithms developed for kernelised bandits in Section 3.5.1.1. We show that a recent algorithm that is designed for continuum-armed bandit in Hölder spaces [Liu et al., 2021] is also optimal over functions in RKHS of Matérn kernels in Section 3.5.1.2.

3.5.1.1 SupKernelUCB and GP-UCB for RKHS

Recall that the lower bound (in terms of T) on cumulative regret for kernelised bandit with Matérn- ν kernels $k_{\text{Matérn},\nu}$ is $\Omega(T^{\frac{\nu+1}{2\nu+1}})$, as proved by Scarlett et al. [2017]. There are mainly two types of algorithms applicable for the kernelised bandit problem: (i) GP-UCB [Srinivas et al., 2009] and its variants [Chowdhury and Gopalan, 2017, Janz et al., 2020], and (ii) KernelUCB and its *Sup*-variant Valko et al. [2013]. The GP-UCB-style algorithms display a non-trivial empirical advantage over the impractical SupKernelUCB. That being said, GP-UCB is suboptimal theoretical upper bounds for certain types of kernels under the RKHS assumption, including for Matérn- ν kernels. In the RKHS of a Matérn kernel $k_{\text{Matérn},\nu}$, GP-UCB achieves a regret of $\tilde{O}(T^{\frac{\nu+\frac{3}{2}}{2\nu+1}})$.⁶ On the other hand,

⁶The suboptimality of GP-UCB is discussed more extensively in Vakili et al. [2021b]

SupKernelUCB matches the lower bound with a regret rate of $\tilde{O}(T^{\frac{\nu+1}{2\nu+1}})$.⁷

3.5.1.2 UCB-Meta for Hölder Space

Apart from the kernelised bandit algorithms discussed above, Liu et al. [2021] propose an algorithm for continuum-armed bandits in Hölder space with exponent $\alpha > 1$ with regret upper bound that matches existing lower bounds [Wang et al., 2018, Singh, 2021] except log factors. This algorithm is named UCB-Meta. We show in Theorem 17 that UCB-Meta is naturally minimax optimal in dependence on T over the RKHS of certain kernels.

Theorem 17. *Consider the kernelised bandit problem where $f \in \mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X}, B)$, where $\nu > 0$ and $\nu + \frac{1}{2} \in \mathbb{N}$. Then, UCB-Meta achieves the following regret upper bound,*

$$\sup_{f \in \mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X}, B)} \mathbb{E}[R_T] = \tilde{O}\left(T^{\frac{\nu+1}{2\nu+1}}\right), \quad (3.18)$$

where \tilde{O} omits dependence on radius of the RKHS ball B , constant factors depending on ν , and log factors of T .

The regret rate shown in Theorem 17 is derived from the result that $\mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X})$ is embedded in a Hölder space $\Sigma^\alpha(\mathcal{X})$ with $\alpha = \nu$. The proof can be found in Appendix 3.9.3. Singh [2021] have shown a similar argument while focusing mainly on the connection between Besov and Hölder spaces.

3.5.2 CORRAL as Adaptive Algorithm

The original CORRAL algorithm for model selection in the bandit setting is first proposed by Agarwal et al. [2017]. The original CORRAL requires that modifications be made to each base algorithm for them to satisfy a stability condition (Definition 3 in Agarwal et al. [2017]). These modifications, however, have to be made on a case-by-case basis. Therefore, we use the smoothed version of CORRAL which is proposed by Pacchiano et al. [2020b]. The smoothed CORRAL puts a smoothing operation between the master algorithm and base algorithms and thus does not require modifications be made to the base algorithms. Smoothed CORRAL operates only with stochastic environments, which is satisfied by our assumptions (Section 3.3). For simplicity, we refer to the smoothed version of CORRAL as CORRAL. CORRAL uses an adversarial online mirror descent algorithm as the master algorithm.

Recall that a non-adaptive minimax kernelised bandit algorithm achieves $\tilde{O}(T^{\frac{\nu+1}{2\nu+1}})$ regret (Section 3.5.1), if instantiated with the correct parameter ν . By plugging in the regret of base kernelised bandit algorithms in the general result in Theorem 5.3 in Pacchiano et al. [2020b], we derive a adaptive upper bound for CORRAL in Theorem 18. CORRAL achieves sublinear $\tilde{o}(T)$ regret on all possible values of ν^* (See Theorem 18). Oppositely, a non-adaptive algorithm instantiated with parameter value $\tilde{\nu}$ does not have sublinear regret guarantees if the true parameter $\nu^* < \tilde{\nu}$, because the underlying function space $\mathcal{H}_{k_{\text{Matérn},\nu^*}}$ is not contained in algorithm's hypothesis space. In Theorem 18, $\tilde{\nu} \in \mathbf{u}$ is a parameter that is specified by the user and can be interpreted as the parameter that specifies the space on which the algorithm is configured to achieve minimax regret.

⁷The analysis of SupKernelUCB was originally for finite-armed setting, but Cai and Scarlett [2021, Appendix A.4] state that it can be extended to the continuum-armed setting where $\mathcal{X} = [0, 1]^d$, suffering only a $O(d \log(T))$ term in the regret.

Theorem 18. Consider the kernelised bandit problem where $f \in \mathcal{H}_{k_{\text{Matérn}, \nu^*}}(\mathcal{X}, B^*)$, $\nu^* + \frac{1}{2} \in \mathbb{N}$ and ν^*, B^* unknown to the learner. Let $\mathbf{u} = \{(\nu_1, B_1), (\nu_2, B_2), \dots, (\nu_M, B_M)\}$ be a list of candidate input value pairs such that \mathbf{u} specifies a nested set of RKHS: $\mathcal{H}_{k_{\text{Matérn}, \nu_1}}(\mathcal{X}, B_1) \subset \mathcal{H}_{k_{\text{Matérn}, \nu_2}}(\mathcal{X}, B_2) \subset \dots \mathcal{H}_{k_{\text{Matérn}, \nu_M}}(\mathcal{X}, B_M)$. Suppose that $(\nu^*, B^*) \in \mathbf{u}$. Let $\mathbb{A} = \{\mathcal{A}_i, i \in [M]\}$ be a set of (non-adaptive) minimax optimal kernelised bandit algorithms with anytime regret guarantees, each instantiated with the regularity and radius $(\nu_i, B_i) \in \mathbf{u}$. The regret from running CORRAL with input total time steps T and learning rate $\eta = \tilde{O}(T^{-\frac{1+\tilde{\nu}}{1+2\tilde{\nu}}})$ applied with base algorithms from \mathbb{A} is as follows.⁸

$$\sup_{f \in \mathcal{H}_{k_{\text{Matérn}, \nu^*}}} \mathbb{E}[R_T] = \tilde{O} \left(T^{\max\left(\frac{1+\tilde{\nu}}{1+2\tilde{\nu}}, \frac{1^2+2\tilde{\nu}+\tilde{\nu}\nu^*}{(1+2\tilde{\nu})(1+\nu^*)}\right)} \right). \quad (3.19)$$

The proof of Theorem 18 can be found in Appendix 3.9.4. This result indicates that CORRAL achieves (i) minimax optimal rate $\tilde{O} \left(T^{\frac{1+\nu^*}{1+2\nu^*}} \right)$ in terms of T , if the underlying kernel regularity $\nu^* = \tilde{\nu}$; (ii) suboptimal rate $\tilde{O} \left(T^{\frac{1+\tilde{\nu}}{1+2\tilde{\nu}}} \right)$ if $\nu^* > \tilde{\nu}$ and (iii) suboptimal rate $\tilde{O} \left(T^{\frac{1+2\tilde{\nu}+\tilde{\nu}\nu^*}{(1+2\tilde{\nu})(1+\nu^*)}} \right)$ when $\nu^* < \tilde{\nu}$. Let ν_1^*, ν_2^* satisfying $\nu_1^* < \nu_2^*$ be two possible values of the true regularity that both satisfy the assumptions in Theorem 18. Suppose $\nu_1^* \leq \tilde{\nu} < \nu_2^*$. By Theorem 18, CORRAL achieves regret $\tilde{O} \left(T^{\frac{1^2+2\tilde{\nu}+\tilde{\nu}\nu_1^*}{(1+2\tilde{\nu})(1+\nu_1^*)}} \right)$ if the true parameter is ν_1^* and $\tilde{O} \left(T^{\frac{1+\tilde{\nu}}{1+2\tilde{\nu}}} \right)$ if the true parameter is ν_2^* . By Theorem 14, the lower bound over the rougher RKHS with ν_1^* is $\tilde{\Omega} \left(T^{\frac{1+2\tilde{\nu}+\tilde{\nu}\nu_1^*}{(1+2\tilde{\nu})(1+\nu_1^*)}} \right)$. The lower bound is matched by the upper bound in the exponent of T .

In conclusion, CORRAL matches the adaptivity lower bound in the dependence on T except log factors, between any pair of regularity values (ν_1^*, ν_2^*) , such that $\nu_1^* \geq \tilde{\nu}, \nu_1^* + \frac{1}{2} \in \mathbb{N}$ and $\nu_2^* < \tilde{\nu}, \nu_2^* + \frac{1}{2} \in \mathbb{N}$.

Finally, note that in this subsection, the assumption is that the true parameter(s) are contained in the candidate set \mathbf{u} . Hence, Theorem 18 reflects the cost of adaptation (model selection), which is the difficulty of selecting the best base learner out of all candidates. If, however, the true parameter is not contained in \mathbf{u} , then adaptive algorithms will incur another type of cost, namely the cost of “discretization”. This cost is generated from the difference between the true parameter and the closest value in \mathbf{u} . Using an exponential [Pacchiano et al., 2020b] or linear [Liu et al., 2021] grid for \mathbf{u} can usually incur a small cost of “discretization”.

3.5.3 RBBE as Adaptive Algorithm

The regret bound balancing and elimination (RBBE) algorithm proposed in Pacchiano et al. [2020a] achieves near-optimal regret in several adaptivity problems with linear function spaces. RBBE can be thought of as using a stochastic master algorithm that selects the base algorithm with the smallest candidate cumulative regret at each time. Therefore, it enjoys advantages such as gap-dependent regret bounds and high probability regret bounds. Unlike CORRAL, it does not need a user-specified parameter to control the space over which the algorithm will achieve minimax optimal regret on. Instead, the algorithm achieves simultaneously on all possible values of ν^* the regret upper bound of $\tilde{O} \left(T^{\frac{1+4\nu^*+2\nu^{*2}}{1+4\nu^*+4\nu^{*2}}} \right)$. If we plug this upper bound in Theorem 14 for $\nu^* = \nu_2^*$, then a lower bound of $\tilde{\Omega} \left(T^{\frac{(2\nu_2^*+1)^2+2\nu_1^*\nu_2^{*2}}{(2\nu_2^{*2}+1)^2(\nu_1^*+1)}} \right)$ is incurred for when $\nu^* = \nu_1^*$, given that $0 < \nu_1^* < \nu_2^*$. The upper bound of

⁸ \tilde{O} omits dependence on radius of the RKHS ball B , constant factors depending on ν , and \log factors of T .

RBBE is larger than the lower bound in the exponent of T . A more detailed description of the RBBE algorithm and a formal statement of its adaptivity upper bound can be found in Appendix 3.8.3 and Theorem 21 therein.

To summarize, although both CORRAL and RBBE as adaptive algorithms can achieve sublinear regret simultaneously on different kernel regularity, CORRAL has a better theoretical adaptivity in this problem. While RBBE fails to match the lower bound, CORRAL achieves the adaptivity lower bound for certain pairs of ν values for Matérn- ν kernels.⁹

3.6 Connection with Adaptivity to Hölder Exponents

The adaptivity lower bound in Theorem 14 specifies the difficulty of adapting between two RKHSs of kernels with polynomial Fourier decay rate m_1 and m_2 , where $0 < m_1 < m_2, m_1 \in \mathbb{N}, m_2 \in \mathbb{N}$. Recall that \tilde{R} is the regret upper bound on the smoother RKHS with parameter m_2 . The lower bound on the RKHS specified by m_1 depends inversely on \tilde{R} through an $\Omega(T \cdot \tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}})$ dependence.

Shifting the perspective from RKHS to Hölder spaces, the adaptivity difficulty has been studied by Locatelli and Carpentier [2018], Hadiji [2019], for a subset of values for the Hölder exponent α . Precisely, Theorem 3 in Locatelli and Carpentier [2018] provides an $\Omega(T \cdot \tilde{R}^{-\frac{\alpha_1}{\alpha_1+1}})$ dependence as the lower bound, for adapting between two Hölder spaces with exponents α_1, α_2 satisfying $\alpha_1 < \alpha_2 \leq 1$.¹⁰ Here, \tilde{R} is the upper regret bound on the smoother Hölder space $\Sigma^{\alpha_2}(\mathcal{X})$. We know by Lemma 12 that an RKHS $\mathcal{H}_{k_{m_1}}(\mathcal{X})$ with kernel Fourier decay rate m_1 is norm equivalent to Sobolev space $\mathcal{W}^{m_1}(\mathcal{X})$. Coupled with the Sobolev embedding theorem for integer-order Sobolev spaces [Adams and Fournier, 2003, Theorem 5.4], it is straightforward to see that $\mathcal{H}_{k_{m_1}}(\mathcal{X}) \subset \Sigma^\alpha(\mathcal{X})$, where $\alpha = m_1 - \frac{1}{2}$ (Appendix 3.9.3).

Note that we have the following equivalence between the lower bounds if $\alpha_1 = m_1 - \frac{1}{2}$.

$$T\tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}} \propto T\tilde{R}^{-\frac{\alpha_1}{\alpha_1+1}}. \quad (3.20)$$

Therefore, for continuum-armed bandit problems, the statistical difficulty of adapting to kernel regularity of RKHS is the same as adapting to Hölder exponents, if the Hölder exponents represent the smallest Hölder spaces that the RKHSs embed in.

3.7 Discussion

We discuss several future directions, stemming from the current limitations of our work. Our current theoretical results are for the domain with $d = 1$,¹¹ so it is of interest to extend the current results to $d > 1$. Instead of partitioning the domain $\mathcal{X} = [0, 1]$ into M sub-intervals, one needs to partition the hypercube $[0, 1]^d$ into M sub-cubes and construct the hypothesis functions with appropriate Fourier decay correspondingly. Such an extension is possible akin to Scarlett et al. [2017].

⁹It is our conjecture that the stochastic master used by RBBE (as opposed to the adversarial one in CORRAL) limits its model selection ability in certain cases.

¹⁰Proving the adaptivity rate for when the exponents are larger than 1 remains an open problem.

¹¹Note that this is not equivalent to the dimension of the feature map of a kernel.

Another direction is to derive adaptivity upper bounds in terms of Fourier decay as well and verify the tightness of the lower bound in more cases than Matérn kernels. Since we currently investigate translation-invariant kernels, a more long-term direction is the investigation of adaptivity to rotation-invariant kernels, to connect to NTKs which are usually rotation-invariant dot-product kernels [Bietti and Bach, 2020, Chen and Xu, 2020, Vakili et al., 2021a]. Finally, this study is of theoretical nature, so it remains an open problem to empirically study adaptivity to kernel regularity, based on the insights provided by our lower and upper bounds.

3.8 Auxiliary Derivations

3.8.1 Proof of Norm Equivalency Between RKHS Norm and Sobolev Seminorm

Proof of Lemma 13. It is shown by Wendland [2004, Theorem 10.12, Corollary 10.48] that: if a translation-invariant kernel k with Fourier decay rate s (Lemma 12), then the associated RKHS \mathcal{H}_k defined on a Lipschitz domain Ω is norm equivalent to the Sobolev space $\mathcal{W}^{m=s,2}(\Omega)$. The norm equivalency indicates that there exist two constants c_1, c_2 , $0 < c_1 < c_2$, such that for $f \in \mathcal{H}_k(\Omega)$, the following statement holds.

$$c_1 \|f\|_{m,2,\mathcal{X}} \leq \|f\|_{\mathcal{H}_k} \leq c_2 \|f\|_{m,2,\mathcal{X}}. \quad (3.21)$$

Now, we examine conditions under which the norm equivalency can be extended between the *seminorm* (equation 3.7) of Sobolev spaces and the RKHS norm. As in Lemma 13, let $W_0^{m,p}(\mathcal{X})$ denote the closure of $C_0^\infty(\mathcal{X})$ in $W^{m,p}(\mathcal{X})$.¹² Adams and Fournier [2003, 6.26] give the following result: if \mathcal{X} has finite width, then for $f \in W_0^{m,p}$, the seminorm $|\cdot|_{m,p}$ is equivalent to the standard norm $\|\cdot\|_{m,p}$. The one-dimensional interval domain \mathcal{X} we consider trivially satisfies the Lipschitz boundary condition, hence we have the following result.

Lemma 19. *If a function lies in $W_0^{m,p}(\mathcal{X})$ where $\mathcal{X} = [0, 1]$, then there exists a constant $K < \infty$, such that*

$$|\cdot|_{m,p,\mathcal{X}} \leq \|\cdot\|_{m,p,\mathcal{X}} \leq K |\cdot|_{m,p,\mathcal{X}}. \quad (3.22)$$

Combining Lemma 19 with the norm equivalency in equation 3.21, we recover the inequalities in Lemma 13.

$$c_1 |f|_{m,2,\mathcal{X}} \leq \|f\|_{\mathcal{H}_k} \leq K c_2 |f|_{m,2,\mathcal{X}}. \quad (3.23)$$

□

3.8.2 The Full Statement of Theorem 14

We present the full version of Theorem 14, which fully states the constraints on the radius B_1 and B_2 in Theorem 14. The proof is deferred to Appendix 3.9.1.

Theorem 20. *Consider the bandit problem setting (Section 3.3) with noises $\{\eta_t\}_{t=1\dots T}$ that are $\frac{1}{4}$ -subgaussian. Further assume that the \mathcal{L}_2 norm of functions f we consider is upper bounded by finite*

¹²Here, we borrow the definitions from Adams and Fournier [2003].

value $\gamma_0 < \infty$: $\|f\|_2 \leq \gamma_0$. Let \tilde{R} be a positive number, let $m_2 > m_1 > 0$ be two positive integers, and let B_1, B_2 be two positive variables that satisfy the following conditions.

$$\begin{aligned} \bar{c} \max \left\{ \frac{3^{m_1 + \frac{1}{2}}}{32} C(m_1)^{-m_1 + \frac{1}{2}} \tilde{R}^{-1}, K(m_1, m_2, \gamma_0, \mathcal{X}) \bar{c}^{\frac{m_2 - m_1}{m_2}} B_2^{\frac{m_1}{m_2}} \right\} \\ \leq B_1 \leq C'(m_1, m_2)^{-(m_1 + \frac{1}{2})} \bar{c}^{-(m_1 + \frac{1}{2})} B_2^{m_1 + \frac{1}{2}} \tilde{R}^{m_1 - \frac{1}{2}} \end{aligned} \quad (3.24)$$

where $C(m_1)$ and $C'(m_1, m_2)$ are constants whose exact forms are defined in equation 3.58 and equation 3.64 in the proof. $K(m_1, m_2, \gamma_0, \mathcal{X})$ is a constant depending on m_1, m_2 , the domain and γ_0 .¹³

Consider any algorithm that achieves in RKHS ball $\mathcal{H}_{k_{m_2}}(\mathcal{X}, B_2)$ the following regret upper bound, where the kernel k_{m_2} has Fourier decay rate m_2 .

$$\sup_{f \in \mathcal{H}_{k_{m_2}}(\mathcal{X}, B_2)} \mathbb{E}[R_T] \leq \tilde{R}, \quad (3.25)$$

then, the regret of this algorithm in a (less smooth) RKHS ball induced by another kernel k_{m_1} with Fourier decay rate m_1 is lower bounded by the following.

$$\sup_{f \in \mathcal{H}_{k_{m_1}}(\mathcal{X}, B_1)} \mathbb{E}[R_T] \geq \frac{1}{8} \left(\frac{C(m_1)}{32} \right)^{\frac{m_1 - 1/2}{m_1 + 1/2}} \left(\frac{B_1}{\bar{c}} \right)^{\frac{1}{m_1 + 1/2}} \tilde{R}^{-\frac{m_1 - 1/2}{m_1 + 1/2}} T. \quad (3.26)$$

3.8.3 Adaptivity Upper Bound of RBBE

At each round, RBBE [Pacchiano et al., 2020a] first performs an elimination step to remove misspecified base algorithms, then selects a base algorithm among the remaining ones. The elimination step tests whether each base algorithm is well-specified, that is, whether each base algorithm's hypothesis space contains the underlying function. If a base algorithm fails the test, then it is eliminated. In the selection step, the master algorithm simply chooses the base algorithm with the smallest presumed cumulative pseudo-regret. Therefore, RBBE can be thought of as using a stochastic master algorithm (remarked in Pacchiano et al. [2020a] as well), instead of using an adversarial one as CORRAL [Agarwal et al., 2017, Pacchiano et al., 2020b] does.

The general regret of RBBE is stated in terms of the play ratio, which is the ratio between the number of times a base algorithm is played and the number of times that the best base algorithm is played. To instantiate the play ratio, Pacchiano et al. [2020b] considers only the setting where the regret rates of all base algorithms (if well-specified) have the same exponents on T . That is, the regret rates are T^β with a fixed $\beta \in (0, 1]$ across all base algorithms. However, this setting does not align with our setting where, for base algorithm i with input value ν_i , the exponent of T in its (well-specified) regret bound is $\frac{\nu_i + 1}{2\nu_i + 1}$. Hence, we make changes to the proof in Pacchiano et al. [2020b] to apply it to our problem setting. The result of RBBE is stated in Theorem 21 and the proof is deferred to Appendix 3.9.5.

Theorem 21. *Suppose that the problem setting, the set of candidate values \mathbf{u} and the set of base algorithms \mathbb{A} are the same as defined in Theorem 18. The regret of RBBE applied with base algorithms in \mathbb{A} is as follows, with high probability $1 - \delta$.*

$$\sup_{f \in \mathcal{H}_{k_{\text{Matérn}, \nu^*}}} R_T = \tilde{O}\left(T^{\frac{1 + 4\nu^* + 2\nu^{*2}}{1 + 4\nu^* + 4\nu^{*2}}}\right). \quad (3.27)$$

¹³The exact value of $K(m_2, \gamma_0, \mathcal{X})$ is deferred to the proof of Theorem 4.14 in Adams and Fournier [2003].

3.9 Proofs of Results

3.9.1 Proof of Theorem 20

As explained in Section 3.4.2.1, the proof of Theorem 20 arises from the proof of a parallel Sobolev version of the adaptivity lower bound. We formally state the Sobolev version of adaptivity lower bound below.

Theorem 22. *Consider the bandit problem setting (Section 3.3) with noises $\{\eta_t\}_{t=1\dots T}$ that are $\frac{1}{4}$ -subgaussian. Further assume that the \mathcal{L}_2 norms of functions f we consider are upper bounded by the finite value $\gamma_0 < \infty$: $\|f\|_2 \leq \gamma_0$.¹⁴ Let \tilde{R} be a positive number, let $m_2 > m_1 > 0$ be two positive integers, and let L_1, L_2 be two positive variables that satisfy the following conditions:*

$$\begin{aligned} \max \left\{ \frac{3^{m_1 + \frac{1}{2}}}{32} C(m_1)^{-m_1 + \frac{1}{2}} \tilde{R}^{-1}, K(m_1, m_2, \gamma_0, \mathcal{X}) L_2^{\frac{m_1}{m_2}} \right\} \\ \leq L_1 \leq C'(m_1, m_2)^{-(m_1 + \frac{1}{2})} L_2^{m_1 + \frac{1}{2}} \tilde{R}^{m_1 - \frac{1}{2}} \end{aligned} \quad (3.28)$$

where $C(m_1)$, $C'(m_1, m_2)$ are constants whose exact forms are defined in equation 3.58 and equation 3.64 respectively. $K(m_1, m_2, \gamma_0, \mathcal{X})$ is a constant depending on m_1, m_2 , the domain and γ_0 , the upper bound on the \mathcal{L}_2 norm of functions in the Sobolev ball.¹⁵ Consider an algorithm that achieves in the Sobolev ball $\mathcal{W}^{m_2}(\mathcal{X}, L_2)$ a regret upper bound of \tilde{R} .

$$\sup_{f \in \mathcal{W}^{m_2, 2}(\mathcal{X}, L_2)} \mathbb{E}[R_T] \leq \tilde{R}, \quad (3.29)$$

then, the regret of this algorithm in the less-smooth Sobolev ball $\mathcal{W}^{m_1}(\mathcal{X}, L_1)$ is lower bounded by the following.

$$\sup_{f \in \mathcal{W}^{m_1}(\mathcal{X}, L_1)} \mathbb{E}[R_T] \geq \frac{1}{8} \left(\frac{C(m_1)}{32} \right)^{\frac{m_1 - 1/2}{m_1 + 1/2}} L_1^{\frac{1}{m_1 + 1/2}} \tilde{R}^{-\frac{m_1 - 1/2}{m_1 + 1/2}} T. \quad (3.30)$$

In the next part, we present the proof of Theorem 22, which also leads to Theorem 20. The values B_1, B_2 in Theorem 14 should be set as follows.

$$B_1 = \bar{c}L_1, B_2 = \bar{c}L_2, \quad (3.31)$$

where \bar{c} is the global constant in Lemma 13.

Proof. Consider the Sobolev version of the theorem (Theorem 22). Recall that the adaptivity is between balls in two different spaces, the ‘‘rougher’’ space $\mathcal{W}^{m_1}(\mathcal{X}, L_1)$ and the ‘‘smoother’’ space $\mathcal{W}^{m_2}(\mathcal{X}, L_2)$. First, we consider the constraints between L_1 and L_2 such that $\mathcal{W}^{m_2}(\mathcal{X}, L_2) \subset \mathcal{W}^{m_1}(\mathcal{X}, L_1)$. In other words, $f \in \mathcal{W}^{m_2}(\mathcal{X}, L_2)$ should be sufficient condition for $f \in \mathcal{W}^{m_1}(\mathcal{X}, L_1)$. Theorem 4.14 in Adams and Fournier [2003] and references therein give the following interpolation upper bound between orders of smoothness for a function $f \in \mathcal{W}^{m_1}(\mathcal{X})$,

$$|f|_{m_1, 2} \leq K(m_2, \mathcal{X}) (|f|_{m_2})^{\frac{m_1}{m_2}} \|f\|_2^{\frac{m_2 - m_1}{m_2}}, \quad (3.32)$$

¹⁴By our assumption on the underlying function f in equation 3.5, we know that it has bounded \mathcal{L}_2 norm.

¹⁵The exact value of $K(m_2, \gamma_0, \mathcal{X})$ is deferred to the proof of Theorem 4.14 in Adams and Fournier [2003].

where $K(m_2, \mathcal{X})$ is a constant depending only on m_2 and the domain \mathcal{X} . If $f \in \mathcal{W}^{m_2}(\mathcal{X}, L_2)$, then by definition (equation 3.8) we know that $|f|_{m_2} \leq L_2$. Using equation 3.32, we have that:

$$|f|_{m_1, 2} \leq K(m_2, \mathcal{X}) L_2^{\frac{m_1}{m_2}} \|f\|_2^{\frac{m_2 - m_1}{m_2}} \quad (3.33)$$

To ensure that the two Sobolev balls are nested, L_1 should be larger than the right-hand side of the above inequality. The \mathcal{L}_2 norm of f is upper bounded by $\|f\|_2 \leq \gamma_0$. Plugging it in equation 3.33 incurs an lower bound for L_1 :

$$L_1 \geq K(m_1, m_2, \gamma_0, \mathcal{X}) L_2^{\frac{m_1}{m_2}} = K(m_2, \mathcal{X}) \gamma_0^{\frac{m_2 - m_1}{m_2}} L_2^{\frac{m_1}{m_2}}.$$

Having established $\mathcal{W}^{m_2}(\mathcal{X}, L_2) \subset \mathcal{W}^{m_1}(\mathcal{X}, L_1)$, we start with the formal proof of the adaptivity lower bound.

Function Construction Part I. This part is adapted from the regression lower bounds in [Tsybakov \[2004, Section 2.6\]](#). Let M be a positive integer parameter, which is the number of hypothesis functions we need. The value of M remains to be determined later in the proof. In the following, we shall assume $M \geq 2$ and eventually prove that this assumption holds. Further, define bandwidth $h = \frac{1}{2M}$. Let $\Delta > 0$ be a parameter that represents the maximum of the M hypothesis functions in $\mathcal{W}^{m_1, 2}(\mathcal{X}, L_1)$. The value of Δ remains to be determined later in the proof same as M .

Partition the 1-dimensional domain $\mathcal{X} = [0, 1]$ into $M + 1$ bins: $H_{0 \dots M}$, such that $\cup_{s=0 \dots M} H_s = \mathcal{X}$. Define the bins and their middle points $\bar{x}_{0 \dots M}$ as follows.

$$H_s = \left[\frac{s-1}{2M}, \frac{s}{2M} \right], \quad \bar{x}_s = \frac{s - \frac{1}{2}}{2M}, \quad \text{for } s = 1 \dots M,$$

$$H_0 = \left[\frac{1}{2}, 1 \right], \quad \bar{x}_0 = \frac{3}{4}.$$

We use the bump function as a base function, then we shift the base function to construct the hypothesis functions. The bump function is defined as follows. It has compact support on $(-1, 1)$. Function $K_0(\cdot)$ is infinitely differentiable with continuous derivatives [[Tsybakov, 2004, \(2.34\)](#)].

$$K_0(x) = \exp\left(\frac{-1}{1-x^2}\right) \mathbb{I}(|x| < 1). \quad (3.34)$$

Next, define $M + 1$ functions as follows, each one has support inside one of the $M + 1$ bins.

$$f_s = ah^{m_1 - \frac{1}{2}} K\left(\frac{x - \bar{x}_s}{h}\right), \quad s = 1 \dots M, \quad (3.35)$$

$$f_0 = \tilde{a}h^{m_2 - \frac{1}{2}} \tilde{K}\left(\frac{x - \bar{x}_0}{h}\right), \quad (3.36)$$

where

$$K(u) = K_0(bu), \quad (3.37)$$

$$\tilde{K}(u) = K_0(\tilde{b}u). \quad (3.38)$$

$a, b, \tilde{a}, \tilde{b}$ are non-negative parameters to be defined later. We require that $b \geq 2$ and $\tilde{b} \geq 4h$, so that the support of every function f_s is inside H_s , $\forall s \leq M$. Lemma 24 ensures that the requirements on b, \tilde{b} hold, by posing constraints between Δ and M .

We introduce the following lemma to specify requirements on the variables $a, b, \tilde{a}, \tilde{b}$, with respect to Δ and L_1, L_2 . This is to make sure that values of $a, b, \tilde{a}, \tilde{b}$ guarantee that $f_s \in \mathcal{W}^{m_1}(\mathcal{X}, L_1)$, $\forall 1 \leq s \leq M$ and $f_0 \in \mathcal{W}^{m_2}(\mathcal{X}, L_2)$.

Lemma 23. *Let K_0^* to denote the maximum value of $K_0(\cdot)$, a constant less than 1. Let I_{m_1}, I_{m_2} denote the \mathcal{L}_2 norms of the m_1, m_2 -th order derivatives of $K_0(\cdot)$, respectively. That is, $I_{m_1} = \int_{-1}^1 [K_0^{(m_1)}(u)]^2 du$ and $I_{m_2} = \int_{-1}^1 [K_0^{(m_2)}(u)]^2 du$. Then, if Δ is the maximum of f_s in $\mathcal{W}^{m_1,2}(\mathcal{X}, L_1)$, for all $s = 1 \dots M$ and $\Delta/2$ is the maximum of f_0 in $\mathcal{W}^{m_2,2}(\mathcal{X}, L_2)$, the function parameters $a, b, \tilde{a}, \tilde{b}$ satisfy the following:*

$$a = \Delta(2M)^{m_1 - \frac{1}{2}} / K_0^* \quad (3.39)$$

$$\tilde{a} = \Delta(2M)^{m_2 - \frac{1}{2}} / 2K_0^* \quad (3.40)$$

$$b \leq \left(\frac{L_1^2 K_0^{*2}}{\Delta^2 (2M)^{2m_1 - 1} I_{m_1}} \right)^{\frac{1}{2m_1 - 1}} \quad (3.41)$$

$$\tilde{b} \leq \left(\frac{4L_2^2 K_0^{*2}}{\Delta^2 (2M)^{2m_2 - 1} I_{m_2}} \right)^{\frac{1}{2m_2 - 1}}, \quad (3.42)$$

Proof of Lemma 23. The constraints on a, \tilde{a} follows trivially from the requirement that $f_s^* = \Delta$ for $s = 1 \dots M$, $f_0^* = \Delta/2$, and plugging in $h = 1/2M$.

The constraints on b, \tilde{b} are to ensure that

$$\begin{aligned} \|f_s^{(m_1)}\|_2 &\leq L_1, \quad s = 1 \dots M \\ \|f_0^{(m_2)}\|_2 &\leq L_2 \end{aligned}$$

We first consider requirement for $\|f_s^{(m_1)}\|_2 \leq L_1$, $s = 1 \dots M$. For $s \geq 1$,

$$\begin{aligned} &\|f_s^{(m_1)}\|_2^2 \\ &= \int_0^1 [f^{(m_1)}(x)]^2 dx \\ &= \int_0^1 \left[ah^{m_1 - \frac{1}{2}} \frac{\partial^{m_1}}{\partial x^{m_1}} \left(K \left(\frac{x - \bar{x}_s}{h} \right) \right) \right]^2 dx \\ &= a^2 h^{2m_1 - 1} \int_0^1 \left[\frac{\partial^{m_1}}{\partial x^{m_1}} \left(K_0 \left(\frac{b}{h} (x - \bar{x}_s) \right) \right) \right]^2 dx \\ &= a^2 h^{2m_1 - 1} \int_0^1 \left[\left(\frac{b}{h} \right)^{m_1} K_0^{(m_1)} \left(\frac{b}{h} (x - \bar{x}_s) \right) \right]^2 dx \\ &\stackrel{u = \frac{b}{h}(x - \bar{x}_s)}{=} a^2 h^{-1} b^{2m_1} \int_{\frac{b}{h}(-\bar{x}_s)}^{\frac{b}{h}(1 - \bar{x}_s)} [K_0^{(m_1)}(u)]^2 \frac{h}{b} du \\ &= a^2 b^{2m_1 - 1} \int_{-1}^1 [K_0^{(m_1)}(u)]^2 du = a^2 b^{2m_1 - 1} I_{m_1}. \end{aligned}$$

The second to last step follows because the bump function K_0 has compact support on $(-1, 1)$ and the upper and lower limits of the integral satisfy:

$$\begin{aligned} \frac{b}{h}(1 - \bar{x}_s) &= b \left(\frac{1}{h} - s + \frac{1}{2} \right) > 1, \\ \frac{b}{h}(-\bar{x}_s) &= -b \left(s - \frac{1}{2} \right) \leq -1. \end{aligned}$$

Therefore, for $\|f_s^{(m_1)}\|_2^2 \leq L_1^2$ to hold, we need $a^2 b^{2m_1-1} I_{m_1} \leq L_1^2$. This leads to

$$b \leq \left(\frac{L_1^2}{a^2 I_{m_1}} \right)^{\frac{1}{2m_1-1}} \quad (3.43)$$

$$= \left(\frac{L_1^2 (K_0^*)^2}{\Delta^2 (2M)^{2m_1-1} I_{m_1}} \right)^{\frac{1}{2m_1-1}}. \quad (3.44)$$

Similarly, for $s = 0$, we have the following.

$$\begin{aligned} & \|f_s^{(m_2)}\|_2^2 \\ &= \int_0^1 \left[f^{(m_2)}(x) \right]^2 dx \\ &= \int_0^1 \left[\tilde{a} h^{m_2 - \frac{1}{2}} \frac{\partial^{m_2}}{\partial x^{m_2}} \left(\tilde{K} \left(\frac{x - \bar{x}_0}{h} \right) \right) \right]^2 dx \\ &= \int_0^1 \tilde{a}^2 h^{2m_2-1} \left[\frac{\partial^{m_2}}{\partial x^{m_2}} \left(K_0 \left(\frac{\tilde{b}(x - \bar{x}_0)}{h} \right) \right) \right]^2 dx \\ &= \tilde{a}^2 h^{2m_2-1} \int_0^1 \left[\left(\frac{\tilde{b}}{h} \right)^{m_2} K_0^{(m_2)} \left(\frac{\tilde{b}}{h} (x - \bar{x}_0) \right) \right]^2 dx \\ & \stackrel{u = \tilde{b}(x - \bar{x}_0)/h}{=} \tilde{a}^2 \tilde{b}^{2m_2-1} \int_{-\frac{3\tilde{b}}{4h}}^{\frac{\tilde{b}}{h}(1 - \frac{3}{4})} \left[K_0^{(m_2)}(u) \right]^2 du \\ &= \tilde{a}^2 \tilde{b}^{2m_2-1} \int_{-1}^1 \left[K_0^{(m_2)}(u) \right]^2 du \\ &= \tilde{a}^2 \tilde{b}^{2m_2-1} I_{m_2}. \end{aligned}$$

Note that in the third last equation, the integral upper and lower limit satisfy:

$$\frac{\tilde{b}}{h} \left(1 - \frac{3}{4} \right) > 1, \quad -\frac{3\tilde{b}}{4h} < -1.$$

For the above $\|f_s^{(m_2)}\|_2^2$ to be less or equal to L_2^2 , we need:

$$\tilde{b} \leq \left(\frac{L_2^2}{\tilde{a}^2 I_{m_2}} \right)^{\frac{1}{2m_2-1}} = \left(\frac{4L_2^2 (K_0^*)^2}{\Delta^2 (2M)^{2m_2-1} I_{m_2}} \right)^{\frac{1}{2m_2-1}} \quad (3.45)$$

□

Combining Lemma 23 with what we required of the function parameters: $b \geq 2$ and $\tilde{b} \geq 4h$, we then need the following requirements for the parameter Δ . Intuitively, the following lemma says that the functions cannot be too “wavy”, so that they stay within the corresponding balls in Sobolev spaces.

Lemma 24. *For $b \geq 2, \tilde{b} \geq 4h$ to hold, Δ needs to satisfy the following constraints with respect to M and the smoothness constants L_1, L_2 .*

$$\Delta/L_1 \leq \frac{K_0^*}{2^{2m_1-1} M^{m_1 - \frac{1}{2}} \sqrt{I_{m_1}}}, \quad (3.46)$$

$$\Delta/L_2 \leq \frac{K_0^*}{2^{2m_2-2} \sqrt{I_{m_2}}}. \quad (3.47)$$

Proof of Lemma 24. First, consider function f_s when $s \geq 1$. Using the conclusions in Lemma 23 we need the following,

$$\frac{L_1^2(K_0^*)^2}{\Delta^2(2M)^{2m_1-1}I_{m_1}} \geq b^{2m_1-1} \geq 2^{2m_1-1}.$$

What directly follows is the constraint on Δ :

$$\Delta^2 \leq \frac{L_1^2(K_0^*)^2}{2^{4m_1-2}M^{2m_1-1}I_{m_1}}. \quad (3.48)$$

Similarly, for f_0 , we need

$$\frac{L_2^2(K_0^*)^2}{\Delta^2(2M)^{2m_2-1}I_{m_2}} \geq \tilde{b}^{2m_2-1} \geq (4h)^{2m_2-1} = 2^{2m_2-1}M^{1-2m_2}.$$

This leads to second constraint on Δ :

$$\Delta^2 \leq \frac{L_2^2(K_0^*)^2}{2^{4m_2-4}I_{m_2}}. \quad (3.49)$$

□

Function Construction Part II. We have defined $f_0 \dots f_M$ in Part I, and identified the constraints between the floating parameters M and Δ , with respect to given parameters m_1, m_2, L_1, L_2 and known constants K_0^*, I_{m_1}, I_{m_2} . In this second part, we define $M + 1$ bandit problems by defining their reward functions ϕ_s , $s = 0 \dots M$ in the following way:

$$\phi_0 = f_0, \quad (3.50)$$

$$\phi_s = f_s + f_0, \quad \forall 1 \leq s \leq M. \quad (3.51)$$

It is obvious that the reward functions satisfy the following conditions. The conditions below are the Sobolev version. They are necessary for the latter half of this proof. Similar conditions were required in [Locatelli and Carpentier \[2018\]](#), [Hadiji \[2019\]](#), see below for details.

1. The function ϕ_0 has peak value $\Delta/2$ and functions $\phi_s, 1 \leq s \leq M$ all have peak value Δ .
2. The function $\phi_0 \in \mathcal{W}^{m_2,2}(\mathcal{X}, L_2)$ and functions $\phi_s \in \mathcal{W}^{m_1,2}(\mathcal{X}, L_1), 1 \leq s \leq M$.
3. For $s \geq 1$, $\phi_s(x) = \phi_0(x)$ for $x \notin H_s$. Also, $\phi_s^* - \phi_s(x) \geq \frac{\Delta}{2}$ when $x \notin H_s$. Here $\phi_s^* = \max_{x \in \mathcal{X}} \phi_s(x)$.

RKHS Version of the Proof. We have now defined $M + 1$ hypothesis functions in two balls in two different Sobolev spaces. By (i) the norm equivalency between Sobolev seminorm (Lemma 13) and the RKHS norm; and (ii) the relationships between B_1, L_1 and B_2, L_2 in equation 3.31, the reward functions also satisfy the following conditions. The conditions below are the RKHS version.

1. The function ϕ_0 has peak value $\Delta/2$ and functions $\phi_s, 1 \leq s \leq M$ all have peak value Δ .
2. $\phi_0 \in \mathcal{H}_{k_{m_2}}(\mathcal{X}, B_2)$, $\phi_s \in \mathcal{H}_{k_{m_1}}(\mathcal{X}, B_1)$, for $1 \leq s \leq M$.
3. $\forall s \geq 1$, $\phi_s(x) = \phi_0(x)$ when $x \notin H_s$. Also, $\phi_s^* - \phi_s(x) \geq \frac{\Delta}{2}$ when $x \notin H_s$.

Lower Bounding Cumulative Regret (Proof Sketch). This part shows the cumulative regret of an algorithm on functions $\phi_1 \dots \phi_M$ is lower bounded by a rate that depends reversely on \tilde{R} , if this algorithm has a regret upper bound of \tilde{R} on reward function ϕ_0 . The proof in the following directly follows from [Hadji \[2019\]](#) and relies on Pinsker's inequality. We write down a proof sketch here for completeness, readers interested in the full version can refer to [Hadji \[2019, Section F\]](#). We use their notations in this part unless otherwise specified. Those include $N_{H_s}(T)$ which is the number of times an algorithm selects an action in bin H_s ; $\mathbb{P}_s^T(\cdot)$ which is the probability distribution of trajectory $\{x_t, y_t\}_{t=1 \dots T}$, when the reward function in the bandit setting is defined by ϕ_s , for $0 \leq s \leq M$. Similarly, $\mathbb{E}_s[\cdot]$ is the expectation with respect to probability \mathbb{P}_s .

By definitions of the reward functions, when the underlying function is ϕ_s for some $s \geq 1$, the cumulative regret is lower bounded by

$$R_{T,s} \geq \frac{\Delta}{2}(T - \mathbb{E}_s[N_{H_s}(T)]) \quad (3.52)$$

For $s = 0$, the regret is lower bounded by

$$R_{T,0} \geq \frac{\Delta}{2} \sum_{s'=1}^M \mathbb{E}_0[N_{H_{s'}}(T)]. \quad (3.53)$$

Pinsker's inequality is used to establish a relationship between the two lower bounds defined above. The equation [3.54](#) is a core step of the proof.

$$\frac{1}{T} \mathbb{E}_s[N_{H_s}(T)] - \frac{1}{T} \mathbb{E}_0[N_{H_s}(T)] \leq \sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_0^T, \mathbb{P}_s^T)}. \quad (3.54)$$

Calculation of KL distance $D_{\text{KL}}(\cdot, \cdot)$ relies on condition [3](#) of $\phi_{0 \dots M}$, as well as the assumption that the noise is 1/4-subgaussian. The result is that the KL distance is bounded by the following.

$$D_{\text{KL}}(\mathbb{P}_0^T, \mathbb{P}_s^T) = 2\mathbb{E}_0[N_{H_s}(T)]\Delta^2. \quad (3.55)$$

With the above, a key intermediate result is reiterated below.

$$\frac{1}{M} \sum_{s=1}^M R_{T,s} \geq \frac{T}{2} \Delta \left(1 - \frac{1}{M} - \sqrt{\frac{\Delta \cdot R_{T,0}}{M}} \right). \quad (3.56)$$

Recall that our [Theorem 22](#) assumes that $\sup_{f \in \mathcal{W}^{m_2, 2}(\mathcal{X}, L_2)} R_T \leq \tilde{R}$, and since $\phi_0 \in \mathcal{W}^{m_2, 2}(\mathcal{X}, L_2)$, it follows directly that $R_{T,0} \leq \tilde{R}$. Therefore, the above inequality becomes

$$\begin{aligned} \frac{1}{M} \sum_{s=1}^M R_{T,s} &\geq \frac{T}{2} \Delta \left(1 - \frac{1}{M} - \sqrt{\frac{\Delta \cdot R_{T,0}}{M}} \right) \\ &\geq \frac{T}{2} \Delta \left(\frac{1}{2} - \sqrt{\frac{\Delta \tilde{R}}{M}} \right). \end{aligned}$$

In the last inequality, $M \geq 2$ is used. This assumption is *not* violated, as shown later.

Choosing the Appropriate value for Δ . Following the above lower bound, we need to choose a value for Δ that (i) does not violate any of the requirements ([Lemma 24](#)) and (ii) maximizes/tightens the lower bound. To do so, the value of Δ should satisfy:

1. $\sqrt{\frac{\Delta \tilde{R}}{M}} \leq \frac{1}{4}$, where $\frac{1}{4}$ is a constant less than $\frac{1}{2}$ (chosen in an arbitrary manner).
2. $\Delta/L_1 \leq \frac{(K_0^*)}{2^{2m_1-1} M^{m_1-\frac{1}{2}} I_{m_1}^{\frac{1}{2}}}$. Note that this condition satisfies only half of the requirements in Lemma 24. We later show that the other condition in Lemma 24 is also satisfied with the selected Δ .

When maximizing Δ , we first set $\Delta/L_1 \approx \frac{(K_0^*)}{2^{2m_1-1} M^{m_1-\frac{1}{2}} I_{m_1}^{\frac{1}{2}}}$ to achieve the optimal trade-off between M and Δ . That is, we set

$$M = \left\lceil \left(\frac{L_1 K_0^*}{2^{2m_1-1} I_{m_1}^{\frac{1}{2}} \Delta} \right)^{\frac{1}{m_1-\frac{1}{2}}} \right\rceil, \quad (3.57)$$

since M needs to be an integer. By simplifying the constant term:

$$C(m_1) \triangleq \left(\frac{K_0^*}{2^{2m_1-1} I_{m_1}^{\frac{1}{2}}} \right), \quad (3.58)$$

we get a simpler expression of M :

$$M = \left\lceil C(m_1) L_1^{\frac{2}{2m_1-1}} \Delta^{\frac{-2}{2m_1-1}} \right\rceil. \quad (3.59)$$

If $\Delta \tilde{R} / \left(C(m_1) L_1^{\frac{2}{2m_1-1}} \Delta^{\frac{-2}{2m_1-1}} \right) \leq \frac{1}{32}$, the condition $\sqrt{\frac{\Delta \tilde{R}}{M}} \leq \frac{1}{4}$ would be satisfied, using the fact that $\frac{x}{2} \leq \lfloor x \rfloor, \forall x > 2$. Shuffling some terms, the requirement $\Delta \tilde{R} / \left(C(m_1) L_1^{\frac{2}{2m_1-1}} \Delta^{\frac{-2}{2m_1-1}} \right) \leq \frac{1}{32}$ becomes:

$$\begin{aligned} \Delta &\leq \frac{1}{32} C(m_1) L_1^{\frac{2}{2m_1-1}} \Delta^{\frac{-2}{2m_1-1}} B^{-1} \\ \Delta^{\frac{2m_1+1}{2m_1-1}} &\leq \frac{C(m_1)}{32} L_1^{\frac{2}{2m_1-1}} \tilde{R}^{-1} \\ \Delta &\leq \left(\frac{C(m_1)}{32} \right)^{\frac{m_1-\frac{1}{2}}{m_1+\frac{1}{2}}} L_1^{\frac{1}{m_1+\frac{1}{2}}} \tilde{R}^{-\frac{m_1-\frac{1}{2}}{m_1+\frac{1}{2}}}. \end{aligned}$$

To maximize Δ , we thereby choose

$$\Delta = \left(\frac{C(m_1)}{32} \right)^{\frac{m_1-\frac{1}{2}}{m_1+\frac{1}{2}}} L_1^{\frac{1}{m_1+\frac{1}{2}}} M^{-\frac{m_1-\frac{1}{2}}{m_1+\frac{1}{2}}}. \quad (3.60)$$

This leads to the final lower bound:

$$\begin{aligned} &\frac{1}{M} \sum_{s=1}^M R_{T,s} \\ &\geq \frac{T}{2} \Delta \left(\frac{1}{2} - \sqrt{\frac{\Delta \tilde{R}}{M}} \right) \geq \frac{T \Delta}{8} \\ &= \frac{1}{8} \left(\frac{C(m_1)}{32} \right)^{\frac{m_1-1/2}{m_1+1/2}} T L^{\frac{1}{m_1+1/2}} \tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}}. \end{aligned} \quad (3.61)$$

Verify Assumptions. Last but not least, we have to make sure that the assumptions made throughout the proof are satisfied, by our choice of Δ in equation 3.60 and M in equation 3.59.

1. $M \geq 2$. By the definition of M in equation 3.59, we need to ensure that $C(m_1) L_1^{\frac{2}{2m_1-1}} \Delta^{\frac{-2}{2m_1-1}} \geq 2 + 1 = 3$. Further, plugging in equation 3.60, this becomes the following requirement of L_1 :

$$L_1 \geq \frac{3^{m_1+\frac{1}{2}}}{32} C(m_1)^{-m_1+\frac{1}{2}} \tilde{R}^{-1}. \quad (3.62)$$

2. $\Delta/L_2 \leq \frac{K_0^*}{2^{2m_2-2} \sqrt{I_{m_2}}}$. This is the second requirement in Lemma 24 that has not yet been verified to hold. For this condition to hold, the following constraint on L_2 should be met.

$$L_2 \geq C'(m_1, m_2) L_1^{\frac{1}{m_1+1/2}} \tilde{R}^{-\frac{m_1-1/2}{m_1+1/2}}, \quad (3.63)$$

where,

$$C'(m_1, m_2) = 2^{2m_2-2} \left(\frac{C(m_1)}{32} \right)^{\frac{m_1-1/2}{m_1+1/2}} \frac{\sqrt{I_{m_2}}}{K_0^*} \quad (3.64)$$

is a constant (independent of T) that depends on m_1, m_2 . In other words, to make sure that the requirements in Lemma 24 are met, we need in the assumptions the following constraint.

$$L_1 \leq C'(m_1, m_2)^{-(m_1+\frac{1}{2})} L_2^{m_1+\frac{1}{2}} \tilde{R}^{m_1-\frac{1}{2}}. \quad (3.65)$$

We have proved Theorem 22 (Sobolev version).

The constraints on B_1 and B_2 in Theorem 14 are derived from the constraints on L_1, L_2 in Theorem 22 and setting B_1, B_2 as instructed in equation 3.31. Then the proof of Theorem 14 is also completed. \square

3.9.2 Proof of Corollary 16

When $d = 1$, Matérn kernel with regularity parameter ν has Fourier decay rate of $\nu + \frac{1}{2}$ (Definition 15). The algorithm considered in Corollary 16 thus satisfies the regret upper bound on an RKHS induced by a kernel with decay rate $m_2 = \nu_2 + \frac{1}{2}$ which is $\tilde{R} = \tilde{O}(T^{\frac{m_2+\frac{1}{2}}{2m_2}})$. Let m_1 be an integer larger than m_2 . Applying Theorem 14, the lower bound on RKHS of a kernel with Fourier decay rate m_1 is $\Omega(\tilde{R}^{-\frac{m_1-\frac{1}{2}}{m_1+\frac{1}{2}}} T)$. For simplicity, we omit the dependence on B (and constant factors) and focus only on the dependence on T . Plugging in the rate of \tilde{R} , the lower bound then becomes $\Omega(T^{\frac{m_1 m_2 + \frac{3}{2} m_2 - \frac{1}{2} m_1 + \frac{1}{4}}{2m_1 m_2 + m_2}})$. Set $m_1 = \nu_1 + \frac{1}{2}$ as the Fourier decay rate of $k_{\text{Matérn}, \nu_1}$ in Corollary 16. Then, we get the lower bound by substituting $m_2 = \nu_2 + \frac{1}{2}$ and $m_1 = \nu_1 + \frac{1}{2}$, which is $\Omega(T^{\frac{\nu_1 \nu_2 + 2\nu_2 + 1}{(\nu_1+1)(2\nu_2+1)}})$.

3.9.3 Proof of Theorem 17

UCB-Meta [Liu et al., 2021] achieves minimax regret rate in dependence on T (except log factors) in Hölder spaces with Hölder exponent $\alpha > 1$. For $0 < \alpha \leq 1$, it reduces to the minimax optimal continuum-armed bandit algorithm from Auer et al. [2007]. For simplicity, we consider UCB-Meta as the general algorithm for continuum-armed bandits in Hölder spaces. To prove that it is also minimax optimal over RKHS of certain Matérn kernels, we establish the following embedding of RKHS of Matérn kernels to Hölder spaces, via (i) norm equivalency between RKHS of a Matérn- ν kernel and Sobolev space with order m and (ii) Sobolev embedding theorem that specifies the embedding of

Sobolev space with order m to Hölder space with exponent α . Note that Singh [2021] have shown that the minimax bandit algorithm over a Besov or Sobolev space is the same as one that is minimax over the smallest Hölder space that the Besov or Sobolev space embeds onto, although not explicitly for RKHS. For completeness, we still include the following proof. We first state the Sobolev embedding theorem [Adams and Fournier, 2003, Theorem 5.4].

Theorem 25 (Sobolev embedding theorem [Adams and Fournier, 2003]). *Let m be a non-negative integer. Suppose that the dimension $d < p \cdot m$ and $\alpha = m - \frac{d}{p}$. Let Ω be a finite domain with Lipschitz boundary. Then, the Sobolev space $\mathcal{W}^{m,p}(\Omega)$ is embedded onto Hölder space with exponent α :*

$$\mathcal{W}^{m,p}(\Omega) \subset \Sigma^\alpha(\Omega). \quad (3.66)$$

For our problem setting, we set $p = 2$ and $d = 1$. The domain $\mathcal{X} = [0, 1]$ satisfies the Lipschitz boundary condition. Therefore, $\mathcal{W}^m(\mathcal{X}) \subset \Sigma^\alpha(\mathcal{X})$ where $\alpha = m - \frac{1}{2}$. Combining Sobolev embedding theorem with the norm equivalency between Sobolev space and RKHS (Lemma 12), we have the following result.

Corollary 26. *Suppose that $k_s : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a positive-definite translation-invariant kernel, whose Fourier transformation decays polynomially with rate s , $s > d/2$, $s \in \mathbb{N}$. Then, the RKHS $\mathcal{H}_{k_s}(\mathcal{X})$ is embedded onto Hölder space $\Sigma^\alpha(\mathcal{X})$ with exponent $\alpha = s - \frac{d}{2}$:*

$$\mathcal{H}_{k_s}(\mathcal{X}) \subset \Sigma^{s-\frac{d}{2}}(\mathcal{X}). \quad (3.67)$$

The above relationship is also studied in the earlier work of Shekhar and Javidi [2020, Appendix B.1]. Note that Matérn kernels with regularity parameter ν have a Fourier decay rate of $s = \nu + \frac{d}{2}$. Hence, $\mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X}) \subset \Sigma^\alpha(\mathcal{X})$, for $\alpha = \nu$. Therefore, since UCB-Meta achieves on $\Sigma^\alpha(\mathcal{X})$ the regret rate of $\tilde{O}(T^{\frac{\alpha+1}{2\alpha+1}})$ [Liu et al., 2021, Equation (19)], it achieves the same rate $\tilde{O}(T^{\frac{\nu+1}{2\nu+1}})$ on the subset $\mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X})$. Here, we omit the dependence on B , the RKHS norm bound. A function $f \in \mathcal{H}_{k_{\text{Matérn},\nu}}(\mathcal{X}, B)$ also has a finite Hölder norm $\|f\|_{\Sigma^\alpha} = \nu$. The norm $\|f\|_{\Sigma^\nu}$, by definition, poses an upper bound on L (using the notation from Liu et al. [2021, Definition 1], the Hölder-continuity coefficient of the l -th order derivative of f , where l is the largest integer strictly less than α). By Theorem 4 from Liu et al. [2021], we can see that L affects the regret only through a multiplicative term and not through the exponents of T . Therefore, we omit the dependence on B and write the regret rate of UCB-Meta as $\tilde{O}(T^{\frac{\nu+1}{2\nu+1}})$.

3.9.4 Proof of Theorem 18

Recall that Theorem 5.3 in Pacchiano et al. [2020b] provides general regret bounds for CORRAL. The proof of our Theorem 18 is an adaptation to the proof of Theorem 5.3 in Pacchiano et al. [2020b]. We use the same notations as Pacchiano et al. [2020b] unless otherwise specified. M is the number of base algorithms (also aligning with the statement in Theorem 18). δ is the probability of failure. $U : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}^+$ is the cumulative regret function (for a base algorithm), such that $U(t, \delta)$ is the high-probability and anytime regret bound of a base algorithm. ρ is the maximum of reciprocals of the probability that the base algorithm is chosen by the master algorithm over all time steps. η is the learning rate of the master algorithm whose value is determined later in the proof.

In Section 3.5.1.1, we discussed briefly SupKernelUCB [Valko et al., 2013] versus GP-UCB Srinivas et al. [2009]. Despite the convenient implementation and good empirical performance of GP-UCB, SupKernelUCB matches the non-adaptive lower bound in the dependence on T except log factors under the RKHS assumption and thus is minimax optimal while GP-UCB is not. UCB-Meta [Liu et al., 2021] as shown in Theorem 17 is also minimax optimal in the dependence on T except log factors for the Matérn RKHS setting. For this subsection, however, we use SupKernelUCB as base algorithms, since the regret bound of SupKernelUCB has an explicit dependence on B , while for UCB-Meta the dependence on B would rely on an implicit constant (see proof of Theorem 17 in Appendix 3.9.3). We set $d = 1$ as specified in Section 3.5.

Given B and ν of a Matérn- ν kernel, the regret bound of SupKernelUCB is $\tilde{O}(B^{\frac{1}{2}}T^{\frac{\nu+1}{\nu+2}})$ in the RKHS of the Matérn kernel [Valko et al., 2013, Theorem 1]. Note that the original SupKernelUCB (i) is for finite action set and (ii) takes T as input and therefore does not have any time regret guarantees. As mentioned in Section 3.5.1.1, Cai and Scarlett [2021] argue that the aforementioned problem (i) could be extended to the continuum-armed setting by a discretization argument with an extra $O(d(\log(T)))$ term in the regret. The problem (ii) can be theoretically circumvented by the doubling procedure [Auer et al., 1995]. Doubling converts an algorithm with (cumulative) regret bound for fixed T to one with anytime regret bound, suffering only up to constant factors in the regret.¹⁶ Therefore, for theoretical interest, we treat SupKernelUCB as the minimax optimal base algorithm with anytime regret upper bound $\tilde{O}(B^{\frac{1}{2}}T^{\frac{\nu+1}{\nu+2}}), \forall T$.

We acknowledge that this is for theoretical convenience only and it remains an important open problem [Vakili et al., 2021b] to improve the regret bound of the practical GP-UCB algorithm under RKHS assumptions.

We plug in $U(T, \delta) = \tilde{O}(B^{\frac{1}{2}}T^{\frac{\nu+1}{2\nu+1}})$ for the base algorithms for CORRAL. Following the proof of Pacchiano et al. [2020b, Theorem 5.3], we have the following. Note that this upper bound holds with respect to any base algorithm with anytime high-probability regret $U(t, \delta)$. Therefore, we plug in the regret of the best base algorithm, which is $U(t, \delta) = \tilde{O}(B^{*\frac{1}{2}}t^{\frac{\nu^*+1}{2\nu^*+1}})$ because ν^*, B^* belong in the set of candidate values \mathbf{u} .

$$\begin{aligned} R_T &\leq O\left(\frac{M \ln(T)}{\eta} + T\eta\right) - \mathbb{E} \left[\frac{\rho}{40\eta \ln(T)} - \rho U(T/\rho, \delta) \log(T) \right] + \delta T + 8\sqrt{MT \log\left(\frac{4TM}{\delta}\right)} \\ &\leq \tilde{O}\left(\frac{M}{\eta} + T\eta + \delta T + \sqrt{MT}\right) - \mathbb{E} \left[\tilde{O}\left(\frac{\rho}{\eta} - \rho\sqrt{B^*}T^{\frac{\nu^*+1}{2\nu^*+1}}\rho^{-\frac{\nu^*+1}{2\nu^*+1}}\right) \right] \\ &\stackrel{\text{set } \delta=\frac{1}{T}}{=} \tilde{O}\left(\frac{M}{\eta} + T\eta + \sqrt{MT}\right) - \mathbb{E} \left[\tilde{O}\left(\frac{\rho}{\eta} - \sqrt{B^*}T^{\frac{\nu^*+1}{2\nu^*+1}}\rho^{\frac{\nu^*}{2\nu^*+1}}\right) \right] \end{aligned}$$

Maximizing the above equation over ρ results in $\rho \propto \eta^{\frac{2\nu^*+1}{\nu^*+1}} B^{*\frac{\nu^*+\frac{1}{2}}{\nu^*+1}} T$. If we plug this value for ρ in the above equation, then the regret is bounded by:

$$\begin{aligned} R_T &= \tilde{O}\left(\frac{M}{\eta} + T\eta + \sqrt{MT}\right) - \tilde{O}\left(\eta^{\frac{\nu^*}{\nu^*+1}} B^{*\frac{\nu^*+\frac{1}{2}}{\nu^*+1}} T - \eta^{\frac{\nu^*}{\nu^*+1}} B^{*\frac{2\nu^*+1}{2\nu^*+2}} T\right) \\ &\leq \tilde{O}\left(\frac{M}{\eta} + T\eta + \sqrt{MT} + \eta^{\frac{\nu^*}{\nu^*+1}} B^{*\frac{2\nu^*+1}{2\nu^*+2}} T\right) \end{aligned}$$

For the problem of adapting to kernel regularity (represented by ν^* when the kernel is a Matérn kernel), since CORRAL does not have access to ν^* (and B^*), we choose η with respect to the user-specified

¹⁶The doubling procedure is also used in other works that use CORRAL to adapt to unknown parameters of the function space, for example Liu et al. [2021] which studied adaptivity to the Hölder exponent.

parameter $\tilde{\nu}$: $\eta = T^{-\frac{\tilde{\nu}+1}{2\tilde{\nu}+1}}$. Plugging this choice of η back in the above equation, we have:

$$R_T \leq \tilde{O}\left(MT^{\frac{\tilde{\nu}+1}{2\tilde{\nu}+1}} + B^* \frac{2\nu^*+1}{2\nu^*+2} T^{\frac{\tilde{\nu}\nu^*+2\tilde{\nu}+1}{(2\tilde{\nu}+1)(\nu^*+1)}}\right).$$

Absorbing the dependence on M and B in \tilde{O} , we then have the regret rate in equation 3.19.

3.9.5 Proof of Theorem 21

The proof follows from the general form of regret upper bound of RBBE (Theorem 5.1 from Pacchiano et al. [2020a]). The regret bound in Theorem 5.1 in Pacchiano et al. [2020a] is expressed with the “play ratio” $\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_*(t_i)}$, where \mathcal{B} denotes the set of misspecified base algorithms, t_i denotes the last round before base algorithm i is eliminated, and $n_i(t)$ denotes the number of times i is selected until time step $t \leq T$. In the following part, we use Lemma A.3 in Pacchiano et al. [2020a] to calculate the play ratio, then plug it in Theorem 5.1 of Pacchiano et al. [2020a] to get the final regret bound. For reasons why the more straightforward result (Theorem 5.4 in Pacchiano et al. [2020a]) is not used, see the end of this subsection for an explanation.

In the following, each base algorithm i has the following candidate pseudo regret bound (equation (7) in Pacchiano et al. [2020a]):

$$R_i(t) \leq C\theta_i T^{\beta_i}, \quad (3.68)$$

where $C \geq 1$ is some term independent of T or i , and $\theta_i \geq 1$ is some parameter dependent on i . For minimax optimal kernelised bandit algorithms instantiated with ν_i (parameter of the Matérn kernel), $\beta_i = \frac{\nu_i+d}{2\nu_i+d}$. We write down the general regret bound of RBBE here for completeness (Theorem 5.1 [Pacchiano et al., 2020a]). Below, $*$ denotes any well-specified learner, that is, a learner whose actual (pseudo) regret Reg_i is upper bounded by its candidate (which means if well-specified) regret bound $R_i(T)$.

$$\begin{aligned} R_T &\leq \sum_{i=1}^M R_*(n_*(t_i)) + \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_*(t_i)} R_*(n_*(t_i)) + 2M + 2c \sum_{i \in \mathcal{B}} \sqrt{n_i(t_i) \ln\left(\frac{M \ln(T)}{\delta}\right)} \\ &\quad + 2c \sum_{i \in \mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_*(t_i)}} \sqrt{n_i(t_i) \ln\left(\frac{M \ln(T)}{\delta}\right)} \end{aligned}$$

We refer to the five terms in the above summation above as #1...#5.

The terms #1 + #3 can be bounded the same way as in the proof of Theorem 5.4 in Pacchiano et al. [2020a]:

$$\sum_{i=1}^M R_*(n_*(t_i)) + 2M \leq MR_*(T) + 2M \leq \tilde{O}(M\theta_* T^{\beta_*}).$$

The term #4 is bounded also following the proof in Pacchiano et al. [2020a]:

$$\begin{aligned} 2c \sum_{i \in \mathcal{B}} \sqrt{n_i(t_i) \ln\left(\frac{M \ln(T)}{\delta}\right)} &\leq 2c \sqrt{|\mathcal{B}| \ln \frac{M \ln(T)}{\delta}} \sum_{i \in \mathcal{B}} n_i(t_i) \\ &\leq 2c \sqrt{|\mathcal{B}| T \ln \frac{M \ln(T)}{\delta}} \end{aligned}$$

Bounding the term #1 and #5, however, needs changes to the proof of Theorem 5.4 [Pacchiano et al., 2020a], since the play ratio is involved. Lemma A.3 in Pacchiano et al. [2020a] states that for two base learners i, j ,

$$\frac{n_i(t)}{n_j(t)} \leq \max \left\{ \left(2 \frac{\theta_j}{\theta_i} \right)^{\frac{1}{\beta_i}} (n_j(t))^{\frac{\beta_j}{\beta_i} - 1}, 2 \right\}. \quad (3.69)$$

Therefore, the play ratio between a misspecified base learner i and a well-specified learner $*$ can be bounded by:

$$\begin{aligned} \frac{n_i(t)}{n_*(t)} &\leq 2 + \left(2 \frac{\theta_*}{\theta_i} \right)^{\frac{1}{\beta_i}} n_*(t)^{\frac{\beta_*}{\beta_i} - 1} \\ &\leq 2 + 4C_2 B_* n_*(t)^{\frac{\beta_*}{\beta_i} - 1} \\ &\leq 2 + 4C_2 B_* n_*(t)^{2\beta_* - 1}. \end{aligned}$$

The first inequality above is simply plugging $j = *$ (representing a well-specified learner), and using that $\max\{x, y\} \leq x + y$. For the second inequality, recall that the minimax optimal SupKernelUCB algorithm has a regret rate (if the kernel parameter ν and RKHS norm bound B are known) of $\tilde{O}(\sqrt{B\gamma_T T}) = \tilde{O}(\sqrt{BT}^{\frac{\nu+d}{2\nu+d}})$. The \tilde{O} notation hides polynomial terms that are dependent on $\log(T), d$. Therefore, the parameter θ_i in equation 3.68 that depends on the index of the base algorithm i is $\theta_i \propto \sqrt{B_i}$. Given the assumption that $\theta_i \geq 1$, $\frac{\theta_*}{\theta_i} \leq C_1 \sqrt{B_*}$ for some constant C_1 . Since $\beta_i \geq \frac{1}{2}$, $\left(2 \frac{\theta_*}{\theta_i} \right)^{\frac{1}{\beta_i}} \leq 4C_2 B_*$ for some constant C_2 . Also in the last two inequalities, we used $\beta_i \geq \frac{1}{2}$, that is, every base algorithm used in Theorem 21 have at least $\tilde{O}(T^{\frac{1}{2}})$ regret. Therefore, we have the following bound on the sum of play ratio:

$$\sum_{i \in \mathcal{B}} \frac{n_i(t)}{n_*(t)} \leq 2|\mathcal{B}| + 4C_2 B_* |\mathcal{B}| (n_*(t))^{(2\beta_* - 1)} \quad (3.70)$$

$$\leq 2|\mathcal{B}| + 4C_2 B_* |\mathcal{B}| T^{(2\beta_* - 1)} = 2|\mathcal{B}| (1 + 2C_2 B_* T^{(2\beta_* - 1)}) \quad (3.71)$$

We can plug in equation 3.71 to bound #5 as follows.

$$\begin{aligned} 2c \sum_{i \in \mathcal{B}} \sqrt{\frac{n_i(t_i)}{n_*(t_i)}} \sqrt{n_i(t_i) \ln \left(\frac{M \ln(T)}{\delta} \right)} &\leq 2c \sqrt{\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_*(t_i)} \sum_{i \in \mathcal{B}} n_i(t_i) \ln \frac{M \ln(T)}{\delta}} \\ &\leq 2c \sqrt{\sum_{i \in \mathcal{B}} \frac{n_i(t)}{n_*(t)} T \ln \frac{M \ln(T)}{\delta}} \\ &\leq 2c \sqrt{2|\mathcal{B}| (1 + 2C_2 B_* T^{(2\beta_* - 1)}) T \ln \frac{M \ln(T)}{\delta}} \\ &= \tilde{O}(|\mathcal{B}|^{\frac{1}{2}} B_*^{\frac{1}{2}} T^{\beta_*}) \end{aligned}$$

Similarly, the upper bound of term #2 relies on equation 3.71 as well.

$$\begin{aligned}
\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_*(t_i)} R_*(n_*(t_i)) &\leq C \sum_{i \in \mathcal{B}} \frac{n_i(t)}{n_*(t)} \theta_* n_*(t_i) \\
&\leq C \sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{(n_*(t_i))^{1-\beta_*}} \theta_* \\
&\leq C \left(\sum_{i \in \mathcal{B}} \frac{n_i(t_i)}{n_*(t_i)} \right)^{(1-\beta_*)} \theta_* (n_i(t_i))^{\beta_*} \\
&\leq C \theta_* \left(2|\mathcal{B}|(1 + 2C_2 B_* T^{(2\beta_*-1)}) \right)^{(1-\beta_*)} T^{\beta_*} \\
&= \tilde{O}(\theta_* |\mathcal{B}|^{(1-\beta_*)} B_*^{1-\beta_*} T^{(2\beta_*-1)(1-\beta_*)+\beta_*}) \\
&= \tilde{O}(\theta_* |\mathcal{B}|^{(1-\beta_*)} B_*^{1-\beta_*} T^{4\beta_*+2\beta_*^2-1})
\end{aligned}$$

Now that the asymptotic rates of the five terms are derived, we can see that term #2 dominates in the dependence of T and #5 dominates dependence on $|\mathcal{B}|, B_*$, and hence, the regret of RBBE can be bounded as follows.

$$R_T \leq \tilde{O}(\theta_* |\mathcal{B}|^{\frac{1}{2}} B_*^{\frac{1}{2}} T^{4\beta_*+2\beta_*^2-1}) \quad (3.72)$$

$$= \tilde{O}(\theta_* |\mathcal{B}|^{\frac{1}{2}} B_*^{\frac{1}{2}} T^{\frac{2\nu_*^2+4\nu_*+1}{(2\nu_*+1)^2}}) \quad (3.73)$$

$$= \tilde{O}(\theta_* M^{\frac{1}{2}} B_*^{\frac{1}{2}} T^{\frac{2\nu_*^2+4\nu_*+1}{(2\nu_*+1)^2}}) \quad (3.74)$$

Finally, the reason for not using the straightforward results in Theorem 5.4 of [Pacchiano et al. \[2020a\]](#) is as follows. In adaptation to the kernel regularity parameter ν , the candidate regret bounds of base algorithms do *not* have the same exponent of T . The candidate regret bounds having the same rates of T is a requirement for the more straightforward results, hence, those results are not directly applicable to our setting.

Chapter 4

A General “Plug and Play” Framework for M-estimators in Bandits

This chapter is based on an on-going work conducted jointly with Aarti Singh.

4.1 Introduction

The exploration-exploitation dilemma, as introduced in Chapter 1, is at the core of many bandit optimization problems. Through exploration, the algorithm samples different actions to estimate the unknown reward function. Through exploitation, the algorithm selects actions predicted to have high rewards to keep the overall performance good. One prominent way to achieve exploration-exploitation trade-off is through the optimism in the face of uncertainty principle. Algorithms following this principle maintain an optimistic estimate on the unknown environment and act according to this optimistic estimation. The upper confidence bound (UCB) algorithm was first formally stated in [Lai and Robbins \[1985\]](#), although its *theoretical* guarantees were only asymptotic at first. The UCB algorithm maintains an upper confidence bound on the uncertain environment and selects actions with high upper confidence estimate of rewards. This ensures that the algorithm is either choosing actions to learn about uncertain aspects of the environment (exploration), or exploiting actions that are certain to lead to good reward (exploitation). The better the confidence estimates are, the better the performance the learner has. Many works have since followed the guidelines laid out by [Lai and Robbins \[1985\]](#) and designed UCB-type algorithms with optimal performances for a wide range of parametric models, including but not limited to linear [[Dani et al., 2008](#), [Abbasi-Yadkori et al., 2011](#)], kernelised [[Srinivas et al., 2009](#), [Valko et al., 2013](#)], generalized linear [[Filippi et al., 2010](#), [Faury et al., 2020](#)], and even neural networks [[Zhou et al., 2020](#), [Kassraie and Krause, 2022](#)].

The confidence estimates in these works have typically been developed in a case-by-case fashion, with a specific model and estimator in mind. Confidence sets or confidence sequences (which are confidence sets that hold simultaneously over all steps) of the linear parameter in linear bandit problems [[Abbasi-Yadkori et al., 2011](#)] were derived analytically based on the close-form expression of the (regularized) least-square estimator. Similarly for kernelised bandit [Srinivas et al. \[2009\]](#) problems. confidence

interval analysis for logistic (regression) [Faury et al., 2020] bandit was developed specific to the logistic function and might not be trivially adapted to another function. Lee et al. [2024] proposed a general confidence sequences framework that derives confidence sets for the target parameter using likelihood ratio-based PAC-Bayesian approach. However, their scope is within the generalized linear model. In a separate line of work, for heavy-tail noise setting, Li and Sun [2024], Huang et al. [2023] studied the performance of the Huber estimator in the bandit and reinforcement learning setting.

Despite the many advances made in prior works for the specific problems considered, some of which are of independent statistical interest, it remains unclear what analysis and techniques can be effective in a more general setup beyond the scope of the original work. In this paper, we present a unifying framework for M -estimators (empirical risk minimizers) in the bandit setting (sequential and non-i.i.d.). We point out that treatment for the finite-sample time-uniform confidence sequences across different bandit problem settings are essentially two simple steps. The first step is concentration of the M -estimator based on its influence function (score function). The second step is a general “sub- ψ ” framework Howard et al. [2021], Whitehouse et al. [2023b] that can control the deviation of score function. This framework subsumes *all* mentioned prior works and provides a unifying lens on the design of UCB-type algorithms.

In the first step, we draw connection to influence functions. While influence functions are used to derive *asymptotic* variance in classical theory for M -estimators, they can be a useful and versatile tool in characterization of M -estimators behaviors in finite-sample as well. Mathieu et al. [2022] presented a finite-sample analysis for multivariate M -estimators in d -dimension by deriving a connection between the tail probability of an M -estimator and the empirical deviation of the influence function, under general assumptions on the influence (score) functions. Mathieu et al. [2022]’s analysis assumed the data are i.i.d samples from a fixed distribution P , and the estimation goal is the mean of the distribution. Mathieu [2022]’s general main result shows that as long as the deviation of the score function can be controlled (via assumptions on itself and its derivative), one can obtain finite-sample concentration of the M -estimator. One may wonder if the same is true for multivariate M -estimator, when the data are no longer i.i.d., for example in a *sequential decision-making* setup where the action-selection policy changes per round. We answer this question positively by extending the main result of Mathieu et al. [2022] to the non-i.i.d. data and regression setting.

4.2 Problem Setting

Consider the (penalized) M -estimation/empirical risk minimization formulation. Denote a data point as

$$Z = (X, Y),$$

the input $X \in \mathcal{X} \in \mathbb{H}^d$, some Hilbert space. The reward $Y \in \mathcal{Y} \in \mathbb{R}$. In the standard non-contextual bandit set up, at step t , the algorithm chooses action X_t and receives a reward Y_t . Recall that the observation model is

$$Y_t = f_{\theta^*}(X_t) + \eta_t,$$

where $\theta^* \in \Theta \in \mathbb{R}^d$ is the true model parameter unknown to the learner. Denote the sequence of data collected by the learner at time t as $\{Z\}_{i=1\dots t} = \{(X_i, Y_i)\}_{i=1\dots t}$. Define the filtration $\mathcal{F}_t = \{(X_i, \eta_i)\}_{i=1\dots t}$. We make the standard assumption that $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$. Throughout this work,

we consider deterministic algorithms for which X_t is deterministic conditioned on \mathcal{F}_{t-1} . Algorithms based on optimism in face of uncertainty (that is, algorithms using upper confidence bounds) normally satisfy this condition.

The learner constructs an estimator using some loss function

$$\rho : \mathcal{X} \times \mathcal{Y} \times \Theta \rightarrow \mathbb{R}_{\geq 0},$$

after observing a sequence of data $\{Z\}_{i=1..t} = \{(X_i, Y_i)\}_{i=1..t}$, the learner employs an estimator that is the solution to the following (penalized) M-estimation problem. The penalty function is $p(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$.

$$\hat{\theta}_t = \arg \min_{\theta \in \Theta} J_t(\theta), \quad \text{where the empirical risk } J_t(\theta) \triangleq \sum_{i=1}^t \rho(Y_i; f_\theta(X_i)) + p(\theta). \quad (4.1)$$

Note that $\{Z\}_i$ s do not have to be i.i.d in our setting, and can be generated sequentially and in a data-dependent manner. Let:

1. φ denote first derivative of ρ with respect to $f_\theta(x)$.
2. φ' denote the second derivative.

We focus on the M-estimation setting where $\hat{\theta}_t$ can be formulated as the root of:

$$\theta : \quad \nabla_\theta J_t(\theta) = \sum_{i=1}^t \varphi(Y_i; f_\theta(X_i)) \nabla_\theta f_\theta(X_i) + \nabla_\theta p(\theta) = 0 \quad (4.2)$$

The empirical Hessian is as follows.

$$\text{Hess}(J_t)(\theta) = \sum_{i=1}^t \{ \varphi'(Y_i; f_\theta(X_i)) \nabla_\theta f_\theta(X_i) \nabla_\theta f_\theta(X_i)^\top + \varphi(Y_i, f_\theta(X_i)) \nabla_\theta^2 f_\theta(X_i) \} + \nabla_\theta^2 p(\theta). \quad (4.3)$$

4.2.1 Preliminary: Influence Function in Risk Minimization Problems

Let P be a data distribution over (X, Y) . For empirical risk minimization problems, the influence function typically consists of two parts: an inverse Matrix term where the Matrix corresponds to Hessian of the risk; and a first-order term that corresponds to the gradient of the risk and is also called the score function. Below, we illustrate the influence function aligned with our assumptions (equation 4.2).

Consider the *regularized* risk minimization problem stated below. Also assume that the objective is continuous and twice-differentiable (a.e.) in θ . Let $\tilde{\theta}$ be the unique solution to:

$$\tilde{\theta} = \arg \min_{\theta \in \Theta} \mathbb{E}_P[\rho(Z; \theta)] + p_t(\theta). \quad (4.4)$$

Here p_t is the penalty function that could depend on the number of data samples t . If the second-order gradient (the Hessian) of the regularized risk, $\mathbb{E}_P[\nabla_\theta^2 J(\theta)]$, is non-singular, the influence function with bias of a data point $Z = (X, Y) \in (\mathcal{X}, \mathcal{Y})$ on θ^* is given as follows [Avella-Medina, 2017]. To connect with the formulations for the empirical risk, one can treat $P = \hat{P}_t = \frac{1}{t} \sum_{i=1}^t \delta_{Z_i}$. $\tilde{\theta}$ is then the minimizer of the empirical risk: $\hat{\theta}_t = \tilde{\theta}(\hat{P}_t)$, and is biased with respect to the true model parameter θ^* . We put \sim on top of the influence function to highlight its difference from the canonical influence

function in classical statistics without regularization (they differ only with a term induced by the regularization, that is, $\tilde{\mathbb{F}}$ is not zero-mean for θ^*). The bias induced by regularization, as we later show, does not dominate the size of the confidence sequence.

$$\tilde{\mathbb{F}}(Z; P; \tilde{\theta}) = - \left(\mathbb{E}_P[\nabla_{\tilde{\theta}}^2 \rho(Z; \tilde{\theta})] + \nabla_{\tilde{\theta}}^2 p_t(\tilde{\theta}) \right)^{-1} \underbrace{(\nabla_{\tilde{\theta}} \rho(Z; \tilde{\theta}) + \nabla_{\tilde{\theta}} p_t(\tilde{\theta}))}_{\Psi(Z; \tilde{\theta})} \quad (4.5)$$

By setting $P = \hat{P}_t = \frac{1}{t} \sum_{i=1}^t \delta_{Z_i}$, the matrix whose inverse appears in equation 4.5 is exactly the Hessian $\text{Hess}(J_t)(\theta)$ multiplied with $\frac{1}{t}$. The score function is

$$\Psi(Z; \theta) := \varphi(Y; f_{\theta}(X)) \nabla_{\theta} f_{\theta}(X) + \nabla_{\theta} p(\theta).$$

The sum of $\sum_{i=1}^t \Psi(Z_i, \theta)$ is equal to $\nabla_{\theta} J_t(\theta)$, the gradient of the empirical risk.

In the following, if $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a positive-definite matrix. Let $\|\cdot\|_{\mathbf{A}}$ denote the Mahalanobis distance induced by \mathbf{A} : $\forall \nu \in \mathbb{R}^d, \|\nu\|_{\mathbf{A}} = \sqrt{\nu^{\top} \mathbf{A} \nu} \geq 0$.

4.3 Methodology: The Plug-and-Play Framework

4.3.1 Overview

As introduced in Section 4.1, while Mathieu [2022] makes a connection between the concentration of M -estimators and influence functions under the finite-sample setting, the analysis is within mean-estimation problems with i.i.d. data and the result is for fixed-time only. We show that by extending the analysis to regression problems, with sequential non-i.i.d data, and obtaining time-uniform result, this naturally enables a framework to fit M -estimators with bandit algorithms. Specifically, this can be done with only *two steps*. The first step connects the tail behavior of an M -estimator with that of the deviation of the empirical score function. In Section 4.3.3, we present a design principle and results of this first step. Our framework requires the empirical Hessian of the regularized risk to be positive-definite within a certain range. Let θ denote the target parameter. As an overview, we can connect the deviation of θ from $\hat{\theta}_t$ measured in Mahalanobis distance $\|\cdot\|_H$, for a positive-definite matrix H , with the deviation of influence function $\tilde{\mathbb{F}}$, measured in Mahalanobis distance $\|\cdot\|_{H^{-1}}$. That is,

$$\|\theta - \hat{\theta}_t\|_H \simeq \|\nabla_{\theta} J_t(\theta)\|_{H^{-1}} = \left\| \sum_{i=1}^t \Psi(Z_i; \theta) \right\|_{H^{-1}} \quad (\Psi(Z_i; \theta) \text{ defined in equation 4.5}). \quad (4.6)$$

This requires that the matrix H lower bounds the empirical Hessian in Loewner order in a neighborhood measured by \mathbf{d} . Namely, with probability $1 - \delta_H$, the following ‘‘good event’’ is true:

$$\exists c, c', \mathbf{d} > 0, \text{ such that } \forall \vartheta : \|\theta - \vartheta\|_2 \leq \mathbf{d}, \text{Hess}(\vartheta) \succeq cH. \quad (4.7)$$

The details of this step are illustrated in Section 4.3.3. The matrix H is actually a data-dependent quantity, hence we use H_t to represent it in the following subsections.

If, considering a simpler scenario, H is set to be $H_t = t\mathbf{I}_d$, then this recovers Mathieu [2022]’s result for i.i.d. and mean estimation setting $\|\hat{\theta}_t - \theta^*\| \simeq \|\frac{1}{t} \sum_{i=1}^t \Psi(X_i, \theta)\|$. Note that this is parallel in spirit

to the classical asymptotic results $\sqrt{t}(\hat{\theta}_t - \theta^*) \simeq \frac{1}{\sqrt{t}} \sum_{i=1}^t \text{IF}(Z_i; \theta^*)$ but with the first-order term. The inverse Hessian term does not explicitly appear in the result of Mathieu [2022] because their result is for the mean-estimation setting, where $\nabla_{\theta} f_{\theta}(X) \nabla_{\theta} f_{\theta}(X)^{\top} = \mathbf{I}_d$. However, in bandit problems, the optimal shape of confidence sequences is typically data-dependent. This is because the degree of explorations along different directions can vary, hence the uncertainty along these directions should also be different. In this case one should set H_t to be data-dependent.

In the second step (Section 4.3.4), we discuss how to control the deviation of the score function $\|\sum_{i=1}^t \Psi(Z_i; \theta)\|_{H_t^{-1}}$. Prior work Mathieu [2022] considers only the i.i.d. setting where a ℓ_2 norm bound would suffice, and uses classical i.i.d. concentration inequality to control the deviation $\|\sum_{i=1}^t \Psi(X_i; \theta)\|_2$. Classical i.i.d. concentration inequalities would fail in the bandit setting, however. In this work, we discuss which tools can be used to control the deviation in the sequential setting. Specifically, we discuss how H_t and $\sum_{i=1}^t \Psi(Z_i; \theta)$ can be set to satisfy a general “sub- ψ ” condition. This condition is tied to a general, unified framework proposed by Howard et al. [2021] to derive time-uniform martingale concentration inequalities. Whitehouse et al. [2023b] extended the sub- ψ conditions to vector processes. This framework unites many existing results and improves upon some. In the sub- ψ framework, the function ψ characterizes the tail behavior of S_t while H_t acts as its (conditional) variance proxy process.

4.3.2 Model Assumptions

Before we introduce our specific model assumptions, note that for the aforementioned sub- ψ conditions, H_t acts as the variance process of S_t . Plugging in the definition for $S_t = \sum_{i=1}^t \Psi(Z_i, \theta)$, H_t needs to be proportional to $\sum_{i=1}^t \nabla_{\theta} f_{\theta}(X_i) \nabla_{\theta} f_{\theta}(X_i)^{\top} + \nabla^2 p(\theta)$. Namely, the good event in equation 4.7 translates to

$$\sum_{i=1}^t (\varphi'(Y_i; f_{\theta}(X_i)) \nabla_{\theta} f_{\theta}(X_i) \nabla_{\theta} f_{\theta}(X_i)^{\top} + \varphi(Y_i, f_{\theta}(X_i)) \nabla_{\theta}^2 f_{\theta}(X_i)) + \nabla^2 p(\vartheta) \succeq c \left(\sum_{i=1}^t \nabla_{\theta} f_{\theta}(X_i) \nabla_{\theta} f_{\theta}(X_i)^{\top} + \nabla^2 p(\vartheta) \right) \quad (4.8)$$

Note that, typically the sign of φ is indefinite, unlike that of φ' , and the term with the second-order gradient of $f_{\theta}(x)$ is not guaranteed to be positive or PSD. So in order to reason with the above formulation generally, we need to impose upper bounds on the magnitude of $|\varphi|$ as well as $\|\nabla_{\theta}^2 f_{\theta}(x)\|_{op}$.

Throughout the rest of this work, we consider the following setting simplified from 4.2, which still consolidates various models commonly used in bandits including the following:

1. linear model,
2. kernelised model (including neural tangent kernel),
3. generalized linear model (including logistic regression and Poisson regression models).

We start by assuming a generalized linear mode:

$$f_{\theta}(x) = \mu(x^{\top} \theta).$$

$\mu(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is a univariate function. With this setup, the expressions for gradient and Hessian (equation 4.3) of the risk become simpler. Concretely, we consider $\hat{\theta}_t$ as the root of the following problem. We use ℓ_2 regularization: $p(\theta) = \beta\|\theta\|_2^2$.

$$\hat{\theta}_t \in \Theta : \quad \nabla_{\theta} J_t(\theta) = \sum_{i=1}^t \underbrace{\varphi(\mu(X_i^{\top}\theta) - Y_i) X_i}_{\Psi(Z_i; \theta)} + 2\beta\theta = 0. \quad (4.9)$$

The Hessian of the penalized empirical risk then becomes

$$\text{Hess}_t(\theta) = \nabla_{\theta} \left(\sum_{i=1}^t \Psi_i(\theta) + 2\beta\theta \right) = \sum_{i=1}^t \varphi'(\mu(X_i^{\top}\theta) - Y_i) \dot{\mu}(X_i^{\top}\theta) x_i x_i^{\top} + 2\beta \mathbf{I}_d. \quad (4.10)$$

We also make the following assumptions on the input and θ .

B.1 $\forall X \in \mathcal{X}, \|X\|_2 \leq L$.

B.2 $\forall \theta \in \Theta, \|\theta\|_2 \leq S$.

B.3 The empirical Hessian is strictly convex in θ .

With Assumption B.3 and an appropriately large S , $\hat{\theta}_t \in \Theta$ is the unique and existing solution to equation 4.9. In particular, under our sequential decision-making setting, we *do not* assume that the design matrix $\sum_i X_i X_i^{\top}$ is positive definite.

4.3.2.1 Assumptions on risk functions

We make the following assumptions about elements in $\Psi_i(\theta)$ (following from Mathieu et al. [2022]). These assumptions facilitate the design principle (essentially guaranteeing equation 4.8) formally stated in Section 4.3.3.

S.1 $\varphi(x) : \mathbb{R} \rightarrow \mathbb{R}$ and $\mu(x) : \mathbb{R} \rightarrow \mathbb{R}$ are continuous and differentiable in x a.e..

S.2 $\varphi(x)$ is antisymmetric: $\varphi(x) = -\varphi(-x)$, and $\varphi(0) = 0$.

S.3 There exist positive constants γ, r , such that $\varphi'(x) \geq \gamma > 0, \forall |x| \leq r$.

S.4 There exist positive constants c_{μ}, d_{μ} , such that $\mu'(x) \geq c_{\mu} > 0, \forall |x| \leq d_{\mu}$.

S.5 The regularization parameter $\beta > 0$.

Besides the squared loss, our framework also incorporates other losses such as Huber loss and Catoni's loss. Huber loss grows quadratically in a neighborhood around zero, and then linearly outside that neighborhood. This penalizes input with large absolute values and thus mitigates the influence of (rare) extreme input value. For Huber loss:

$$\rho_{H,1}(x) = x^2 \mathbf{1}(|x| \leq 1) + x \mathbf{1}(|x| > 1), \quad (4.11)$$

its derivative $\varphi_{H,1}$ is

$$\varphi_{H,1}(x) = x \mathbf{1}(|x| \leq 1) + \mathbf{1}(|x| > 1). \quad (4.12)$$

One can easily check that $\varphi_{H,1}$ satisfy Assumption S.3 with $\gamma = 1$ and $r = 1$.

Similarly, for Catoni's [Catoni, 2012] wide score function

$$\varphi_{C,1}(x) = \text{sign } x \log\left(1 + |x| + \frac{x^2}{2}\right) \quad (4.13)$$

satisfies Assumption S.3 with $\gamma = 0.8$ and $r = 1$. Catoni's narrow score function satisfies Assumption S.3 with $\gamma = 0.8$ and $r = 0.5$.

The loss functions can be rescaled via a parameter \mathbf{r} , demonstrated as follows. The parameter allows for versatile use of the loss functions. If φ_1 satisfies Assumption S.3 with (γ, r) , then $\varphi_{\mathbf{r}}(x)$ satisfies Assumption S.3 with $(\gamma, \mathbf{r}r)$.

$$\varphi_{\mathbf{r}}(x) = \mathbf{r}\varphi_1\left(\frac{x}{\mathbf{r}}\right). \quad (4.14)$$

In certain settings, for example, with heavy-tail noise (discussed in Section 4.6), one might need to vary \mathbf{r} , so it might take different values \mathbf{r}_t in round t .

4.3.3 Step I: Deviation of Empirical Risk Minimizers from Influence Function

As mentioned at the beginning of Section 4.3, to connect the deviation of the M -estimator with the deviation of the score functions in Mahalanobis distance dependent on matrix H_t , H_t needs to lower bound (in Lowener order) the empirical Hessian matrix around a neighborhood of θ . For the M -estimation problem defined in Section 4.3.2, we present the following design principle.

Design Principle 1 (Convexity of the empirical risk). *There $\exists C_0 > 0$, $\mathbf{d} \geq 2S > 0$, and sequence of positive-definite matrix $(H_t)_t, H_t \succeq \tilde{\beta}\mathbf{I}_d$, such that, the following "good event" is true with probability at least $1 - \delta_H$, $\delta_H \geq 0$:*

$$\forall t \in [T], \forall \vartheta : \|\theta - \vartheta\|_2 \leq \mathbf{d}, \text{Hess}(\vartheta) \succeq C_0 H_t. \quad (4.15)$$

Recall that the simplified model we consider (Section 4.3.2) has Hessian:

$$\text{Hess}_t(\theta) = \sum_{i=1}^t \varphi'_{\mathbf{r}_i}(\mu(X_i^\top \theta) - Y_i) \dot{\mu}(X_i^\top \theta) X_i X_i^\top + \beta \mathbf{I}_d.$$

Note that for the squared loss, $\varphi'(x) = 2$, so Assumption S.3 is satisfied with $r = \infty$. For generalized linear models where μ represents the (inverse) link function, Design principle 1 can be satisfied contingent on $\dot{\mu}(x)$ being lower bounded by a positive constant, details can be found in Section 4.5.

With Design Principle 1, we present the guarantee for this step in Theorem 27.

Theorem 27 (Tail probability of M -estimator and that of influence function). *Consider a M -estimation problem defined in equation 4.9. Suppose Assumptions B.1~B.3 and S.1~S.5 hold for the risk function and the parameters. Let sequence $\{H_t\}_{t \geq 1}$ be a sequence that satisfies Design Principle 1. Then the tail probability of the M -estimator using data collected by a deterministic bandit algorithm (Section 4.2) and the tail probability of the empirical score function satisfy the following.*

Let $\tilde{\beta}$ be a lower bound on $\lambda_{\min}(H_t), t \in [T]$. For a process $(\lambda(t))_{t=1\dots T}, \lambda(t) > 0$:

$$\begin{aligned} \mathbb{P}\left(\exists t, \|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda(t)\right) &\leq \delta_H + \mathbb{P}\left(\exists t : \left\| \sum_{i=1}^t \Psi(Z_i; \theta) \right\|_{H_t^{-1}} \geq C_0 \lambda(t)\right) \\ &\leq \delta_H + \mathbb{P}\left(\exists t : \underbrace{\left\| \sum_{i=1}^t \varphi_{r_i}(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}}}_{\text{deviation of unregularized score function}} + \underbrace{2\sqrt{\tilde{\beta} S}}_{\text{bias term}} \geq C_0 \lambda(t)\right) \end{aligned} \quad (4.16)$$

The proof of Theorem 27 is deferred to Appendix 4.8. In equation 4.16, the term that measures deviation of the score function $\left\| \sum_{i=1}^t \Psi(Z_i; \theta) \right\|_{H_t^{-1}}$ contains two elements. The first (and dominating) element is the deviation of *unregularized* score function $\left\| \sum_{i=1}^t \tilde{\Psi}(Z_i; \theta) \right\|_{H_t^{-1}}$ (where $\tilde{\Psi} = \nabla_{\theta} \rho(Z; \theta)$, without the regularization). The second element is a bias term induced by regularization. The assumption that $\lambda_{\min}(H_t) \geq \tilde{\beta}$ is needed to upper bound the bias term. Hence, Design Principle 1 essentially requires the empirical regularized risk to satisfy a strongly convex curvature condition around θ .

In the second step, we point out that a general framework (the “sub- ψ ” framework) for time-uniform martingale concentration can be used to provide high-probability bound on the deviation of unregularized score functions (at least for standard light-tailed noise settings). Together, they offer a unifying solution for time-uniform confidence sequence for M -estimators.

4.3.4 Step II: Bounding Deviation of Influence Function with Sub- ψ Condition

When the target parameter θ is θ^* the underlying model parameter, $Y_t - \mu(X_t^\top \theta^*)$ becomes the observation noise η_t . η_t is conditionally zero-mean. When H_t satisfy certain conditions, the process $\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}}$ in equation 4.54 can be formulated as a self-normalized martingale vector process. More specifically, if $\sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i$ satisfy a “sub- ψ ” tail condition [Howard et al., 2021, Whitehouse et al., 2023b] with respect to the matrix process H_t , the deviation can be controlled via existing versatile tools, as explained in Section 4.3.4.1. For this step, we assume that the observation noise $\eta_t | \mathcal{F}_{t-1}$ follow a light-tailed distribution, which is standard for bandit problems.

4.3.4.1 Concentration of martingale process and the sub- ψ condition

Martingale concentration results are essential tools for theoretical analysis under the sequential decision-making setting where classical i.i.d. concentration inequalities will no longer suffice. Time-uniform concentration refers to results that are valid simultaneously for all possible values of sample sizes. In cases where one needs to maintain a valid confidence set over all t , such as for algorithms that follow the optimism in face of uncertainty principle (put simply, the UCB-type algorithms) in the bandit setting, time-uniform martingale concentration is a particularly important theoretical tool used by many previous works [Abbasi-Yadkori et al., 2011, Fauray et al., 2020, Janz et al., 2024] to obtain confidence sequences of the unknown model parameter. In short, the martingale dependence frees one from the

i.i.d. assumption, while the time-uniform aspect frees one from the need to invoke union bound over all t causing a looser bound (for example in Dani et al. [2008]).

Martingale (vector) concentration tail bounds have largely been developed individually in a case-specific fashion, the recent work of Howard et al. [2021] presented a unified framework that characterizes a martingale process by the tail behavior of its increment, and obtain the tail bound as the probability that the martingale crosses a certain threshold. Using this framework, Howard et al. [2021] provided a single general algorithm that captures and in some case improves existing martingale inequalities. The characterization is called the sub- ψ condition. On a high level, it captures how a martingale process $(S_t)_{t \geq 0}$ grows with respect to some variance proxy process $(V_t)_{t \geq 0}$. The function $\psi(\lambda)$ itself behaves like the cumulant generating function (CGF) $\log \mathbb{E}[\exp(\lambda \Delta S_t)]$. In the general master theorem of Howard et al. [2021], the deviation of S_t is controlled by both the $\psi(\lambda)$ function and the variance proxy of V_t . The condition for a vector process $(S_t)_t \in \mathbb{R}^d$ to be sub- ψ with a CGF-like function $\psi : [0, \lambda_{\max}] \rightarrow \mathbb{R}_{\geq 0}$, with respect to some variance proxy process $(V_t)_t \in \mathbb{R}^{d \times d}$ is as follows [Whitehouse et al., 2023b]. (Assuming (S_t) and (V_t) are adapted to some filtration $(\mathcal{F}_t)_t$.)

$$\forall v \in \mathbb{S}^{d-1} : \exp(\lambda \langle v, S_t \rangle - \psi(\lambda) \langle v, V_t v \rangle) \leq L_t^{\lambda, v}, \forall t. \quad (4.17)$$

Where $(L_t^{\lambda, v})_t$ is a non-negative super-martingale process adapted to $(\mathcal{F}_t)_t$. This general definition consolidates many tail behaviors, ranging from sub-Poisson to sub-Gamma (equivalent to sub-Exponential). Some examples of ψ functions include:

1. Sub-Gaussian: $\psi_N(\lambda) = \frac{\lambda^2}{2}$.
2. Sub-Poisson: $\psi_{P,c} = \frac{\exp(c\lambda) - c\lambda - 1}{c^2}$.
3. Sub-Gamma: $\psi_{G,c}(\lambda) = \frac{\lambda^2}{2(1-c\lambda)}$.

What rises most often in the bandit setting is not exactly the quantity $\|S_t\|$, but the “self-normalized” process $\|S_t\|_{V_t^{-1}}$. While S_t itself is a martingale, the self-normalized process $\|S_t\|_{V_t^{-1}}$ is often not. This poses additional technical difficulties in controlling the latter quantity. Recently, Whitehouse et al. [2023b] extended the general framework of Howard et al. [2021] to self-normalized process in vector spaces in \mathbb{R}^d . By the general sub- ψ definition for vector processes introduced by Whitehouse et al. [2023b], $\sum_{i=1}^t \eta_i X_t$ is sub-Gaussian with variance proxy $\sum_{i=1}^t X_t X_t^\top + \beta \mathbf{I}_d$ (as appeared in linear bandits with ridge estimator Abbasi-Yadkori et al. [2011]).

Although previously in the sequential decision-making community, Abbasi-Yadkori et al. [2011], Fauray et al. [2020] have developed self-normalizing concentration for $S_t = \sum_{i=1}^t \eta_i X_i$ where η_i are conditionally sub-Gaussian, and Janz et al. [2024] has studied where η_i are conditionally sub-Exponential, the sub- ψ framework is by far the most general and versatile as it directly captures the growth of S_t with respect to the variance process V_t .

4.3.4.2 Bounding the deviation under light-tailed setting

Tying it back to our framework, when the tail behavior of $\varphi(\mu(X^\top \theta) - Y)$ is light-tail, one can directly use results from the sub- ψ framework to control the “self-normalized” process $\|\Psi(Z_i; \theta)\|_{H_t^{-1}}$. In order to do so, H_t needs to grow at least proportionally to the variance proxy of the sum of the score functions. To formalize this method, we present design principle for the second step below.

Design Principle 2 (Sub- ψ relation between score function and the Hessian). *There exists $(H_t)_t$ satisfying Design Principle 1, such that $(S_t, V_t)_t$ is sub- ψ with a CGF-like ψ function (examples include ψ_N , $\psi_{P,c}$ and $\psi_{G,c}$), and :*

$$S_t = \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i$$

$$H_t = \Omega(V_t)$$

As long as such a sub- ψ relationship exists, one can control the deviation of

$$\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}},$$

and thus the deviation of the M -estimator.

4.3.4.3 Treatment for heavy tail

In situations where the noise η follows a conditionally heavy-tail distribution, Design Principle 2 is not satisfied directly, and one cannot directly apply the sub- ψ vector process concentration results to control the deviation of the self-normalized process. On a high level, this is because the parameters τ_t that satisfy Design Principle 1 with appropriate H_t have possibly large (growing sub-linearly with T) magnitude of $\varphi_{\tau}(\eta)$ (defined in Section 4.3.2.1), even if $\varphi_1(\eta)$ is bounded or grows slowly. In Section 4.6, we discuss the heavy-tail noise case in more detail and treatments for this case designed by prior works original for Huber estimator in bandit setup.

4.3.5 Summary

Under the simple two-step framework, if there is a time-uniform upper bound $\zeta_t(\delta)$ on the “self-normalized” process such that with probability at least $1 - \delta$, for all t , the deviation of the score function is bounded as in $\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}} \leq \zeta_t(\delta)$, we can obtain a time-uniform confidence sequence of the M -estimator. Combining with the first step, Theorem 27, then with probability at least $1 - \delta_H - \delta$, the deviation of M -estimator $\|\theta - \hat{\theta}_t\|_{H_t}$ is upper bounded by $\frac{1}{c_0} \left(\zeta_t(\delta) + 2\sqrt{\tilde{\beta}S} \right)$.

In the subsequent sections, we demonstrate that our framework subsumes various settings where UCB-type algorithms use M -estimators to model the reward. In each setting, we show that the design principles are satisfied with appropriate H_t .

4.4 Case Study: Linear and Kernelised Model

We consider the following linear reward model, parametrized by linear parameter $\theta \in \mathbb{R}^d$.

$$Y_t = X_t^\top \theta + \eta_t.$$

We also make the standard assumption that the noises are conditionally zero-mean and sub-Gaussian with constant parameter σ^2 : $\eta_t | \mathcal{F}_{t-1} \in \text{SG}(\sigma^2)$, $\sigma > 0$ [Dani et al., 2008, Abbasi-Yadkori et al., 2011].

The linear estimator is recovered by setting $\mu(x) = x$ in 4.9. For a (penalized) M -estimator with loss function ρ , the empirical risk, its gradient and Hessian are as follows.

$$J_t(\theta) = \sum_{i=1}^t \rho(Y_i - X_i^\top \theta) + \beta \|\theta\|_2^2, \quad (4.18)$$

$$\nabla_\theta J_t(\theta) = \sum_{i=1}^t \varphi(X_i^\top \theta - Y_i) X_i + 2\beta\theta, \quad (\text{using Assumption S.2}) \quad (4.19)$$

$$\text{Hess}_t(\theta) = \sum_{i=1}^t \varphi'(X_i^\top \theta - Y_i) X_i X_i^\top + 2\beta \mathbf{I}_d. \quad (4.20)$$

4.4.1 Ridge Linear Least-Square

Recall that the linear UCB algorithm [Dani et al., 2008, Abbasi-Yadkori et al., 2011] selects the action that has the largest projected linear reward among all θ in a valid confidence ellipsoid

$$C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{H_t} \leq \zeta_t(\delta)\},$$

for some confidence width $\zeta_t(\delta)$. Because of its convenient ellipsoid shape, the confidence set induces a clean form for the upper confidence bound of reward given an action x in the linear case:

$$\text{UCB}_t(x) = x^\top \hat{\theta}_t + \sqrt{\zeta_t} \|x\|_{A_t^{-1}} \quad \text{where } A_t = \sum_{i=1}^t X_i X_i^\top + \beta \mathbf{I}_d. \quad (4.21)$$

$\hat{\theta}_t$ is the ridge linear least-square estimator. This penalized M -estimator is obtained when we set the loss function $\rho(x) = x^2$ with ℓ_2 penalty. For the least-square estimator,

$$\nabla_\theta J_t(\theta) = -2 \sum_{i=1}^t (Y_i - X_i^\top \theta) X_i + 2\beta\theta, \quad (4.22)$$

$$\text{Hess}_t(\theta) = 2 \sum_{i=1}^t X_i X_i^\top + 2\beta \mathbf{I}_d \succeq 2\beta \mathbf{I}_d. \quad (4.23)$$

Since $\text{Hess}_t(\theta)$ does not depend on θ , so

$$H_t = \sum_{i=1}^t X_i X_i^\top + \beta \mathbf{I}_d,$$

satisfy that $\text{Hess}_t(\theta) \succeq 2H_t$ deterministically, $\forall \theta$. H_t satisfies Design principle 1 with $\delta_H = 0$ and $C_0 = 2$. Theorem 27 then gives

$$\mathbb{P} \left(\exists t : \|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda(t) \right) \leq \mathbb{P} \left(\exists t : 2 \underbrace{\left\| \sum_{i=1}^t (X_i^\top \theta - Y_i) X_i \right\|_{H_t^{-1}}}_{\text{self-normalized}} + 2\sqrt{\beta} S \geq 2\lambda(t) \right). \quad (4.24)$$

The righthand side can be controlled whenever the tail behavior of the residual $Y_i - X_i^\top \theta$ can be controlled. If $\theta = \theta^*$ the groundtruth parameter for which $Y_i = X_i^\top \theta^* + \eta_i$ where η_i is the zero-mean noise variable, the above self-normalized term is simply $\left\| \sum_{i=1}^t \eta_i X_i \right\|_{H_t^{-1}}$.

Given that the noise is conditionally zero-mean, and that the action X_t of UCB algorithm is deterministic when conditioned on \mathcal{F}_{t-1} , this quantity $S_t = \sum_{i=1}^t \eta_i X_i$ is a martingale process. When the noise are conditionally sub-Gaussian, $S_t = \sum_{i=1}^t \eta_i X_i$ is sub- ψ_N with $V_t = \sigma^2 \sum_{i=1}^t X_i X_i^\top$ (and consequently with $V_t + \beta \mathbf{I}_d$) [Whitehouse et al., 2023a]. Hence, the result of Whitehouse et al. [2023a] gives that

$$\mathbb{P} \left(\exists t, \left\| \left(\sum_{i=1}^t \eta_i X_i \right) \right\|_{H_t^{-1}} \geq \mathcal{O}(\sqrt{\log \log(1 + tL^2/\beta) + d \log(1 + tL^2/\beta)}) \right) \leq \delta. \quad (4.25)$$

Note that Theorem 2 in the earlier work on UCB algorithm for linear bandit, Abbasi-Yadkori et al. [2011], also provide a result under the specific sub-Gaussian case. A comparison of the two results can be found in Section 4 of Whitehouse et al. [2023a]. We provide a simple summary in Section 4.9.1 for completeness.

To conclude, our framework recovers the high-probability confidence sequence width: $\|\theta^* - \hat{\theta}_t\|_{H_t} \leq \mathcal{O}(\sqrt{d \log(t)} + 2\sqrt{\beta}S)$. This confidence width is known to yield the regret upper bound $\tilde{O}(d\sqrt{T})$, which is optimal in its dependence on t and d for the standard continuum-armed linear setting. As translating the confidence set to the regret upper bound follows standard arguments in literature [Srinivas et al., 2009] and is not our focus, we omit the details here.

4.4.2 Kernelised Model

The kernelised model can be seen as a transformation of the linear model from the original input space to a Kernel Reproducing Hilbert Space (RKHS), also commonly known as the kernel trick. The transformation is defined by a bivariate kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that is positive semidefinite (which implies symmetry). The following result states that there is a unique RKHS induced by the kernel function.

Theorem 28 (Theorem 12.11 [Wainwright, 2019]). *Given any such kernel function k , there is a unique Hilbert space \mathcal{H}_k in which the kernel satisfies the reproducing property. This space is known as the reproducing kernel Hilbert space associated with K .*

The kernel function also defines a feature mapping for the input. Take some function $\Phi : \mathcal{X} \rightarrow \ell^2(\mathbb{N})$ which is a function between the original input domain to a square-summable space possibly of infinite dimension. The feature map defined by the kernel is $x \mapsto \Phi(x)$, where

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle.$$

For all $f \in \mathcal{H}$, the reproducing property states $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$. Hence, the linear relationship now exists between f and the feature map of x . Assumption B.2 in the RKHS setting assumes that the reward functions inside a RKHS ball with finite (constant) radius.

$$f \in \mathcal{H}_{k,B} \triangleq \{f : f \in \mathcal{H}_k, \|f\|_k \leq B\}. \quad (4.26)$$

We assume that the reward is a noisy observation of $f^*(x)$, for input x and an underlying function unknown to the learner:

$$Y_t = f^*(X_t) + \eta_t, \quad f^* \in \mathcal{H}_{k,B}.$$

The Mahalanobis distance deviation of the M -estimator parallel to the linear case, except in the RKHS, is of form $\|f - \hat{f}_t\|_{H_t} = \sqrt{\langle f - \hat{f}_t, H_t(f - \hat{f}_t) \rangle_{\mathcal{H}}}$, where the inner product is now the inner product in the RKHS: $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The kernel ridge estimator can be seen as an extension of the linear ridge least-square estimator (Section 4.4.1), except with the feature map $\Phi(x)$ instead of x , and $f \in \mathcal{H}$ instead of $\theta \in \mathbb{R}^d$ (Section 4.4.2.2).

4.4.2.1 The GP-UCB Algorithm Uses Confidence Ellipsoid Sequence in RKHS

We will see that the GP-UCB [Srinivas et al., 2009] algorithm is exactly performing the linear UCB algorithm (4.21) using the confidence ellipsoid on θ^* , but in the RKHS where $f^* \in \mathcal{H}_{k,B}$.

In RKHS, the ellipsoid-shaped confidence set for f is

$$C_t(f) = \{f : \|f - \hat{f}\|_{H_t, \mathcal{H}} \leq \zeta_t(\delta)\} \quad (4.27)$$

for some width $\zeta_t(\delta)$, then resulting upper confidence bound will be:

$$\text{UCB}_{\text{kernel},t}(x) = \langle \hat{f}_t, \Phi(x) \rangle_{\mathcal{H}} + \sqrt{\zeta_t} \|\Phi(x)\|_{A_t^{-1}, \mathcal{H}} \quad \text{where } A_t = \sum_{i=1}^t \Phi(X_i)\Phi(X_i)^\top + \beta \text{id}_{\mathcal{H}}.$$

$\text{id}_{\mathcal{H}} : \mathcal{H} \rightarrow \mathcal{H}$ is identity mapping in the RKHS. The optimal GP-UCB algorithm in Srinivas et al. [2009] is essentially using this confidence ellipsoid in RKHS to calculate upper confidence bounds on the rewards. Whitehouse et al. [2023a] had already made such a point that the GP-UCB algorithm is leveraging the confidence ellipsoid in the RKHS, but for clarity and completeness we prove this in Section 4.9.2. This motivates us to focus on the confidence ellipsoid in form of $\|f - \hat{f}_t\|_{H_t}$, as recovering the correct confidence width should suffice to recover the regret performance for GP-UCB [Srinivas et al., 2009].

4.4.2.2 Our Framework for M -estimator in Kernel Reproducing Hilbert Space

The kernel ridge estimator is an M -estimator in the RKHS. Namely, \hat{f}_t is the solution to $\nabla_f J_t(f) = 0$ where:

$$\nabla_f J_t(f) = 2 \sum_{i=1}^t (\langle f, \Phi(X_i) \rangle_{\mathcal{H}} - Y_i) \Phi(X_i) + 2\beta f, \quad (4.28)$$

$$\text{Hess}_t(f) = 2 \sum_{i=1}^t \Phi(X_i)\Phi(X_i)^\top + 2\beta \text{id}_{\mathcal{H}}, \quad (4.29)$$

In particular, observe that the Hessian, now an operator in the RKHS, is still positive-definite (Assumption B.3 satisfied). Specifically,

$$\forall g \in \mathcal{H}, \langle g, \text{Hess}_t(f)g \rangle_{\mathcal{H}} = \left\langle g, 2 \left(\sum_{i=1}^t \Phi(X_i)g(X_i) + \beta g \right) \right\rangle_{\mathcal{H}} = 2 \sum_{i=1}^t g(X_i)^2 + \beta \|g\|_{\mathcal{H}} > 0.$$

Hence, kernel ridge estimator under the kernelised setting is in essence the same as the linear ridge estimator. For completeness, we state the following in entirety. Specifically, H_t can be set to the RKHS operator

$$H_t = \sum_{i=1}^t \Phi(X_i)\Phi(X_i)^\top + \beta \text{id}_{\mathcal{H}},$$

so for any f , $\text{Hess}_t(f) \succeq 2H_t$ deterministically, and Design Principle 1 is satisfied with $\delta_H = 0, C_0 = 2$. Then, we have from Theorem 27

$$\mathbb{P}\left(\exists t : \|f - \hat{f}_t\|_{H_t, \mathcal{H}} \geq \lambda(t)\right) \leq \mathbb{P}\left(\exists t : 2\|H_t^{-1/2}\left\{\sum_{i=1}^t (f(X_i) - Y_i)\Phi(X_i)\right\}\| + 2\sqrt{\beta}B \geq 2\lambda(t)\right).$$

Where H_t^{-1} is the inverse operator of H_t . When $f = f^*$, the residual term is simply

$$\left\|\left(\sum_{i=1}^t \Phi(X_i)\Phi(X_i)^\top + \beta \text{id}_{\mathcal{H}}\right)^{-1/2} \left\{\sum_{i=1}^t \eta_i \Phi(X_i)\right\}\right\|_{\mathcal{H}}.$$

The self-normalized sub- ψ_N martingale concentration results in RKHS that can be directly plugged into our framework was provided by Whitehouse et al. [2023a]. Specifically, Their Corollary 1 gives the following bound. Let $\gamma_t(\beta)$ denote the maximum information gain [Srinivas et al., 2009, Valko et al., 2013]. With probability at least $1 - \delta$, for all $t \geq 0$,

$$\left\|\left(\sum_{i=1}^t \Phi(X_i)\Phi(X_i)^\top + \beta \text{id}_{\mathcal{H}}\right)^{-1/2} \left\{\sum_{i=1}^t \eta_i \Phi(X_i)\right\}\right\|_{\mathcal{H}} \leq \sigma \sqrt{2 \log\left(\frac{1}{\delta} \det\left(\mathbf{I}_t + \frac{1}{\beta} \mathbf{K}_t\right)\right)} \leq \mathcal{O}(\sqrt{\gamma_T}). \quad (4.30)$$

The last inequality was by the definition of γ_t .

Putting it together, our framework recovers the width for the confidence sequence $\|f^* - \hat{f}_t\|_{H_t, \mathcal{H}} \leq \tilde{\mathcal{O}}(\sqrt{\gamma_t} + 2\sqrt{\beta}B)$. Treating B as a constant, this is known (by standard argument analogous to the linear case [Srinivas et al., 2009]) to yield a $\tilde{\mathcal{O}}(\gamma_T \sqrt{T})$ regret bound for the RKHS setting.

4.4.2.3 A connection to Neural Networks via Neural Tangent Kernel

As discussed in Section 3.1, Neural Tangent Kernel [Arora et al., 2019, Lee et al., 2019, Chen and Xu, 2020, Vakili et al., 2021a] literature attempts to analyze the behavior of neural networks $f_{\text{NN}}(\theta)$. For example, a sufficiently-wide, overparametrized neural networks trained by gradient descent with infinitesimal step size can be approximated by a kernel regression predictor with a deterministic kernel, the Neural Tangent Kernel (NTK). In particular, recall that the NTK as a kernel has the following definition by feature map:

$$k_{\text{NTK}}(x, x') = \frac{1}{m} \langle \nabla_{\theta} f_{\text{NN}}(x; \theta_0), \nabla_{\theta} f_{\text{NN}}(x'; \theta_0) \rangle \quad \text{for sufficiently large } m. \quad (4.31)$$

Where the corresponding feature map of the NTK is $\Phi_{\text{NTK}}(x) = \frac{1}{\sqrt{m}} \nabla_{\theta} f_{\text{NN}}(x; \theta_0)$.

The expression of NTK is defined in a recursive way dependent on the network structure, but there were efforts to capture their complexity or smoothness (hence relating to γ_T for the NTK) in a more traditional statistical sense. For example, Vakili et al. [2021a] argue that the NTK of certain wide, fully-connected networks with ReLU-type activation function is *isomorphic* with a Matérn- ν kernel [Matern et al., 1960] with parameter $\nu = 1/2$, which is the roughest inside the Matérn kernel family. Although it is commonly acknowledged that the NTK representation does not fully capture the network behavior, as in practice the network training dynamics often are able to escape the linear regime, the NTK is a step toward approximating and understanding the behavior of large neural networks.

Our framework work for kernelised model when the kernel is an NTK as well. In the NTK realm, the recovered confidence width is $\sqrt{\gamma_{\text{NTK},T}}$ where $\gamma_{\text{NTK},T}$ is the maximum information gain for the neural tangent kernel. [Kassraie and Krause \[2022\]](#) obtains $\gamma_{\text{NTK},T} = \tilde{\mathcal{O}}(T^{\frac{d-1}{d}})$ for fully-connected ReLU neural networks. d is the dimension of the (original) input space $\mathcal{X} \in \mathbb{R}^d$. We refer interested readers to [Whitehouse et al. \[2023a\]](#), [Lattimore \[2023\]](#) for further insights on the regret bounds for NTK bandit problems.

4.5 Case Study: Generalized Linear Models

Generalized linear models (GLMs) also received lots of attention in the bandit setting [[Filippi et al., 2010](#), [Abeille et al., 2021](#), [Li et al., 2017](#), [Lee et al., 2024](#)]. Through a link function that can take various non-linear forms, GLMs can represent more complex functions than linear. In GLM regression, $\mathbb{E}[Y|X]$ belongs 1-dimensional exponential family with parameter $X^\top \theta^*$. Specifically, Y has the following likelihood:

$$\mathbb{P}(Y | X, \theta^*) = \exp\left(\frac{Y X^\top \theta^* - m(X^\top \theta^*)}{v(\eta)} + c(Y, \eta)\right), \quad (4.32)$$

for some functions v, c and a scalar parameter η . Here, $m(\cdot)$ is a twice-differentiable function. Under different definitions of $m(\cdot)$, $P(Y | X^\top \theta^*)$ corresponds to different distributions under the exponential family. $\mu(\cdot) = \dot{m}(\cdot)$ is called the (inverse) link function, it also satisfies that

$$\mathbb{E}[Y | X] = \dot{m}(X^\top \theta^*) = \mu(X^\top \theta^*), \quad \text{Var}(Y | X) = \ddot{m}(X^\top \theta^*) = \dot{\mu}(X^\top \theta^*).$$

Below are two examples of GLMs that are commonly used and studied in bandit optimization. We will focus on these two models in the subsequent parts.

Logistic model. $m(x) = \log(1 + \exp(x))$: $Y | X$ follows a Bernoulli distribution with parameter $\mu(X^\top \theta^*)$, where $\mu(x) = \exp(-x) + 1^{-1}$. This corresponds to the logistic regression model. Y takes value in $\{0, 1\}$ and the model predicts $Y = 1$ whenever $P(Y = 1 | X) = \mu(X^\top \theta^*) \geq 0.5$.

Poisson model. $m(x) = \exp(x)$: $Y | X$ follows a Poisson distribution with parameter $\mathbf{p} = \exp(X^\top \theta^*)$.¹ $\mathbb{E}[Y | X] = \mathbf{p} = \exp(X^\top \theta^*)$. This corresponds to the Poisson regression model which is often used to model count data. In this case, $\mathbb{P}(Y | X) \propto \exp(Y X^\top \theta^* - \exp(X^\top \theta^*))$ and $Y \in \{0, 1, \dots\}$.

The log likelihood following equation 4.32 is:

$$\log \ell(\theta) = \sum_{i=1}^n \left[\frac{Y_i X_i^\top \theta - m(X_i^\top \theta)}{v(\eta)} + c(Y_i, \eta) \right] \quad (4.33)$$

$$= \frac{1}{v(\eta)} \sum_{i=1}^n [Y_i X_i^\top \theta - m(X_i^\top \theta)] + \text{constant} \quad (4.34)$$

$$\propto \sum_{i=1}^n [Y_i X_i^\top \theta - m(X_i^\top \theta)]. \quad (4.35)$$

¹Note that here \mathbf{p} denotes the Poisson parameter and not the deviation variable λ in our main proof.

For GLMs, the MLE estimator that minimizes the (regularized) negative log-likelihood is typically used. In other words, the empirical risk can be written as

$$J_t(\theta) = \sum_{i=1}^t (m(X_i^\top \theta) - Y_i(X_i^\top \theta)) + \beta \|\theta\|_2. \quad (4.36)$$

Consequently, the gradient of the empirical risk and the Hessian are as follows. We see that they conform to the formulations in equation 4.9, where $\varphi(x) = x$ simply and μ is the (inverse) link function in the GLM.

$$\nabla J_t(\theta) = \sum_{i=1}^n X_i (\mu(X_i^\top \theta) - Y_i) + 2\beta\theta, \quad (4.37)$$

$$\text{Hess}(J_t)(\theta) = \sum_{i=1}^n \dot{\mu}(X_i^\top \theta) X_i X_i^\top + 2\beta \mathbf{I}_d \quad (4.38)$$

Note that $\varphi(x) = x$ naturally satisfies Assumption S.3 with $r = \infty$. Since the magnitudes of both the action and the parameter are bounded, that is, $\|X\| \leq L$ (Assumption B.1) and $\|\theta\| \leq S$ (Assumption B.2). If μ satisfy Assumption S.4 with $c_\mu > 0, d_\mu \geq LS$, then $\dot{\mu}(x^\top \theta)$ is lower bounded by a positive constant $c_\mu, \forall x \in \mathcal{X}, \theta \in \Theta$. This then becomes the same common assumption made for the derivative of the link function in prior works on GLM bandits [Filippi et al., 2010, Faury et al., 2020] is

$$c_\mu = \inf_{\theta \in \Theta, x \in \mathcal{X}} \dot{\mu}(x^\top \theta) > 0. \quad (4.39)$$

The logistic and Poisson models discussed above satisfy this assumption with $c_\mu > 0$.

4.5.1 Connection with Deviation of Score Function

For GLMs (equation 4.36), let us choose the matrix H_t to be

$$H_{t,\min} = c_\mu \sum_{i=1}^n X_i X_i^\top + \beta \mathbf{I}_d.$$

Then, Design Principle 1 is satisfied with $\delta_H = 0$ and $C_0 = 1$ and $\mathbf{d} = 2S$. Also, note that the observation model in the GLM also satisfy the following.

$$Y = \mu(X^\top \theta^*) + \eta, \quad (4.40)$$

Where η is the (conditionally) zero-mean residual. Theorem 27 then gives (here, we abbreviate $H_{t,\min}$ to H_{\min} for conciseness):

$$\mathbb{P} \left(\exists t, \|\theta - \hat{\theta}_t\|_{H_{\min}} \geq \lambda(t) \right) \leq \mathbb{P} \left(\exists t : \left\| \sum_{i=1}^t (\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_{\min}^{-1}} + 2\sqrt{\beta}S \geq C_0 \lambda(t) \right) \quad (4.41)$$

$$= \mathbb{P} \left(\exists t : \left\| \sum_{i=1}^t \eta_i X_i \right\|_{H_{\min}^{-1}} + 2\sqrt{\beta}S \geq C_0 \lambda(t) \right) \quad (4.42)$$

Therefore, next we analyze the tail behavior of $\eta_t | \mathcal{F}_{t-1}$ in order to make use of the sub- ψ martingale vector concentration result (Section 4.3.4.1).

4.5.2 Bounding the Deviation of the Score Function

Sub- ψ behavior of residual in Poisson model

Recall that in the Poisson model, $Y | X$ follows a Poisson distribution with parameter $\mathbf{p} = \exp(X^\top \theta^*)$. And $\mathbb{E}[Y | X] = \mathbf{p} = \exp(X^\top \theta^*)$.

Lemma 29. *Consider the noise $\eta_t = Y_t - \mu(X_t^\top \theta^*)$ in Poisson bandit model. Let*

$$S_t = \sum_{i=1}^t \Delta S_i, \quad \Delta S_i = \eta_i X_i.$$

Then S_t is sub-Poisson [Whitehouse et al., 2023b] with $\psi_{P,c}$ with cumulative variance process

$$V_t = \sum_{i=1}^t \mathbb{E}_{i-1} \Delta S_i \Delta S_i^\top = \sum_{i=1}^t X_i X_i^\top \mathbb{E}_{i-1} [\eta_i^2] = \sum_{i=1}^t \dot{\mu}(X_i^\top \theta^*) X_i X_i^\top.$$

The sub-Poisson parameter c is $c = L$.

The proof of Lemma 29 is deferred to Section 4.9.4.

Sub- ψ behavior of residual in Logistic model

Recall that in the logistic model, Y takes value in $\{0, 1\}$. And $Y = 1$ whenever $P(Y = 1 | X) = \mu(X^\top \theta^*) \geq 0.5$. In this case, the residual η takes value in $\{-p, 1 - p\}$, where $p = \mu_{\text{logistic}}(X^\top \theta^*)$. Therefore, conditioned on \mathcal{F}_{t-1} , the norm of ΔS_t is bounded as $\|\eta_t X_t\|_2 \leq L$. Bounded variables are easier to analyze: by also noting that $\mathbb{E}_{t-1}[\eta_t X_t] = 0_d$, we can directly call on the second example for sub- ψ vector processes in Whitehouse et al. [2023b] and show that $S_t = \sum_{i=1}^t \eta_i X_i$ is sub- ψ with respect to variance proxy process $V_t = \sum_{i=1}^t \mathbb{E}_{i-1}[\eta_i^2] X_i X_i^\top = \sum_{i=1}^t \dot{\mu}(X_i^\top \theta^*) X_i X_i^\top$, with ψ being the sub-Poisson ψ function: $\psi_{P,c}$. For this model, the sub-Poisson parameter is $c = L$.

Now we have established the sub- ψ relationship between S_t and V_t . Note that V_t equals the unregularized part of $\text{Hess}_t(\theta^*)$. And we also have

$$V_t = \text{Hess}_t(\theta^*) \preceq \frac{\bar{C}}{c_\mu} H_{\min}.$$

This means the GLMs we consider also satisfy Design Principle 2. We re-state the result for time-uniform self-normalized concentration of sub-Gamma vector martingale below. Note that, sub- $\psi_{P,c}$ naturally implies sub-Gamma (sub- $\psi_{G,c}$) relation. Using the specific sub-Poisson condition gives one a tighter bound (in constants), but there lack a close-form expression for the upper bound corresponding to $\psi_{P,c}$. For simplicity, we invoke the result corresponding to sub- $\psi_{G,c}$ here, and refer readers who are interested in the comparison between $\psi_{P,c}$ and $\psi_{G,c}$ rates to the Appendix D of Whitehouse et al. [2023b].

Theorem 30. *From Whitehouse et al. [2023b] verbatim. For $\psi_{G,c}$ process S_t with variance proxy V_t ,*

$$\|S_t\|_{V_t^{-1}} = \mathcal{O}(\sqrt{\log \log(\lambda_{\max}(V_t)) + d \log \kappa(V_t)} + \frac{c}{\sqrt{\lambda_{\min}(V_t)}} [\log \log(\lambda_{\max}(V_t)) + d \log \kappa(V_t)])$$

To apply this result to our setup, we first need to analyse the eigenvalues of the variance proxy V_t . Let $\bar{C} = \max_{x,\theta} \dot{\mu}(x^\top \theta)$ denote the upper bound of $\dot{\mu}$, The maximum eigenvalue is bounded by

$$\begin{aligned} \lambda_{\max}\left(\sum_{i=1}^t \dot{\mu}(X_i^\top \theta^*) X_i X_i^\top\right) &\leq \sum \lambda_{\max}(\dot{\mu}(X_i^\top \theta^*) X_i X_i^\top) \\ &\leq t\bar{C}L^2. \end{aligned}$$

Due to the sequential and non-iid nature of action sequence in the bandit setting, for the minimum eigenvalue we have $\lambda_{\min}(V_t) \geq 0$. Then, Theorem 30 gives the following result.

Theorem 31. *Recall that $\text{Hess}_t(\theta^*) = \sum_{s=1}^t \dot{\mu}(X_s^\top \theta^*) X_s X_s^\top + 2\beta \mathbf{I}_d$. For Logistic and Poisson models, with high probability $1 - \delta$, for all $t \geq 0$:*

$$\left\| \sum_{s=1}^t \eta_s X_s \right\|_{\text{Hess}_t(\theta^*)^{-1}} = \mathcal{O}(d \log(t)). \quad (4.43)$$

4.5.3 Conclusion

Note that, since $\text{Hess}_t(\theta^*) \succeq \frac{\bar{C}}{c_\mu} H_{t,\min}$, hence

$$\left\| \sum_{s=1}^t \eta_s X_s \right\|_{H_{\min}^{-1}} \leq \sqrt{\frac{\bar{C}}{c_\mu}} \left\| \sum_{s=1}^t \eta_s X_s \right\|_{\text{Hess}_t(\theta^*)^{-1}}.$$

Combining the above with our previous analysis in equation 4.42 then yields that with high probability $1 - \delta$, for all $t \geq 0$:

$$\|\theta - \hat{\theta}_t\|_{H_{\min}} = \mathcal{O}\left(\sqrt{\frac{\bar{C}}{c_\mu}} d \log(t) + \sqrt{\beta} S\right).$$

It then follows directly from $H_{\min} \geq c_\mu V_t$, $V_t = \sum X_i X_i^\top + 2\beta \mathbf{I}_d$ that:

$$\|\theta - \hat{\theta}_t\|_{V_t} \leq \frac{1}{c_\mu} \|\theta - \hat{\theta}_t\|_{H_{\min}} = \mathcal{O}\left(\frac{\sqrt{\bar{C}}}{c_\mu} d \log(t) + \sqrt{\beta} S\right) \quad (4.44)$$

Note that in the confidence sequence width above we retain the term c_μ . In literature of GLM bandits, $\frac{1}{c_\mu}$ is also commonly referred to as κ . Arguments were made that κ , although a T -independent constant, could be large for some models such as the logistic model. Prior work such as Faury et al. [2020] make efforts to improve the dependence on κ in the confidence width and consequently the cumulative regret. In our current work, however, we focus on presenting the unifying framework related to influence function, and defer the improvement related to κ as future work. We anticipate, based on Ostrovskii and Bach [2021], that self-concordance assumption should be able to improve the proof of Theorem 27, if one swaps out the step that uses the mean-value theorem.

4.6 Case Study: Heavy-tail Noise

When the noise have heavy-tail (symmetric) distribution, for example, when $\eta_t | \mathcal{F}_{t-1}$ satisfies $\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$ and $\text{Var}(\eta_t | \mathcal{F}_{t-1}) = \sigma^2$, the two design principles are not directly satisfied, bounding the deviation

of the M -estimator becomes more complicated. Although not the main focus for our current scope, we look at solutions to the heavy-tail case from existing literature for linear models and how they relate to our framework. This subsection provides a high-level, heuristic discussion.

In face of heavy-tail noises, one can use robust losses such as Catoni's [Catoni, 2012] or Huber loss [Huber, 1992], where the parameter $r < \infty$ can be chosen by the learner (equation 4.14). For the heavy-tail noise, there are (technical) distinctions between non-adaptive (and i.i.d.) setting and adaptive settings such as bandit optimization:

1. In the setting where actions are drawn from the same, non-degenerate distribution Σ , $\lambda_{\min}(\Sigma) > 0$, the design matrix $\sum_{i=1}^t X_i X_i^\top$ grows linearly with t (its minimal eigenvalue $\sim t\lambda_{\min}(\Sigma)$). In this case, it would suffice to keep the loss parameter \mathbf{r}_i same across different values of i . Consider a H_t for the linear model that is proportional to $\sum_{i=1}^t X_i X_i^\top + \tilde{\beta} \mathbf{I}_d$. As long as $r_i = O(\sigma)$, one can directly use matrix Hoeffding-style concentration inequality [Tropp et al., 2015] to guarantee Design Principle 1 is satisfied with fail probability δ_H proportional to $\exp(-t)$. Hence, after an order of $\log(T)$ rounds, δ_H can be upper bounded by a small constant.² This reduces to the conclusion of Mathieu [2022] under the pure i.i.d. and mean-estimation setting (where $f_\theta(X) = \theta$) and when noises have bounded variance.
2. In fully adaptive bandit setting where actions are selected sequentially and in a data-dependent fashion, the distribution of $\sum_{i=1}^t X_i X_i^\top$ can be quite arbitrary, growing at different levels in different directions (effect of exploration and exploitation trade-off). At the worst-case, the minimal eigenvalue of $\sum_{i=1}^t X_i X_i^\top$ grows at a constant rate rather than linearly. Here, matrix Hoeffding-style concentration cannot be directly applied to reach a desired bound on δ_H . Instead, one must carefully design \mathbf{r}_t , such that the good event still happens with a large probability but without the magnitude of r having an undesirable dependence on T . A dynamic scheme for setting \mathbf{r}_t proposed in Li and Sun [2024], Huang et al. [2023] can satisfy our design principle with constant δ_H for this setting. Below, we examine how the dynamic parameters can satisfy Design Principle 1.

4.6.1 Data-dependent Loss Function Parameter Satisfies Design Principle 1

In this subsection, we examine how setting a dynamic loss parameter [Li and Sun, 2024, Huang et al., 2023] can be used to satisfy Design Principle 1, given that the loss function also satisfies our Assumption S.3. If the loss function $\rho(\cdot)$ satisfies Assumption S.3, the Hessian of the unregularized empirical risk satisfies:

$$\sum_{i=1}^t \varphi'_{\mathbf{r}_i}(X_i^\top \theta - Y_i) X_i X_i^\top \succeq \gamma \sum_{i=1}^t \mathbf{1}(|X_i^\top \theta - Y_i| \leq \mathbf{r}_i r) X_i X_i^\top.$$

In order for the Huber estimator to have a finite-sample optimal theoretical guarantee in the i.i.d. regression setting, Sun et al. [2017] set the Huber estimator parameter to be adaptive to the sample size. Li and Sun [2023], Huang et al. [2023] extend the method of Sun et al. [2017] to the bandit setting and design a dynamic, data-dependent scheme to set \mathbf{r}_t for the Huber estimator.

²Mathieu et al. [2022] made a similar conclusion for a simpler, finite-armed bandit setting, where a forced exploration phase of $\Omega(\log(T))$ is needed.

On a high level, the dynamic radius parameter r_t is set proportional to the reciprocal of $\|X_t\|_{H_t^{-1}}$. A dynamic scaling is placed such that the unregularized Hessian matrix now becomes $\sum_{i=1}^t \varphi'_{\tau_i} \left(\frac{X_i^\top \theta - Y_i}{\varsigma_i} \right) \frac{X_i X_i^\top}{\varsigma_i}$. ς_t is a data-dependent scaling factor of round t (for details, refer to line 5 of Algorithm 1 in Li and Sun [2024]). Intuitively, this regularizes the weight of $\tilde{X}_i = \frac{X_i}{\varsigma_t}$ among the new design matrix $\sum_{i=1}^t \tilde{X}_i$. Absent this scaling, the analysis would require controlling the probability that absolute values of the residuals stay below τ_i for all rounds, which would inflate the confidence width by an undesirable $\frac{1}{\delta_H}$. The dynamic loss parameter then essentially sets

$$\tau_t = \tau_0 \sqrt{\frac{1 + w_t^2}{w_t^2}}, \quad \text{where } w_t = \|\tilde{X}_t\|_{(\sum_{i=1}^{t-1} \tilde{X}_i \tilde{X}_i^\top + \beta \mathbf{I}_d)^{-1}}. \quad (4.45)$$

With the above alterations to the estimator, Design Principle 1 is satisfied for $\mathbf{d} = 2S$ (Section 4.3.2), $C_0 = \frac{1}{3}$ and $H_t = \sum_{i=1}^t \tilde{X}_i \tilde{X}_i$ for the Huber loss. τ_0 is set proportional to \sqrt{d} as shown in Lemma C.1 of Huang et al. [2023]. The results in Huang et al. [2023] are for the Huber estimator only, and hence the above mentioned constants (such as $C_0 = \frac{1}{3}$) are hard-coded. However, the analysis essentially uses the property of $\varphi_{H,1}$, that it satisfies Assumption S.3 with $\gamma = 1$ and $r = 1$. Thus, it is reasonable to expect that for future efforts, it can be extended to other losses in a straightforward fashion. For example, $\varphi_{\text{Catoni's narrow},1}$ satisfies Assumption S.3 with $\gamma = 0.8$ and $r = 0.5$. For different values of γ and r , τ_0 should simply be multiplied with $\frac{1}{r}$ and the value of C_0 should change with γ .

If the Design Principle 1 is satisfied for $H_t = \sum_{i=1}^t \tilde{X}_i \tilde{X}_i$, Theorem 27 then gives

$$\mathbb{P} \left(\exists t : \|\theta^* - \hat{\theta}_t\|_{H_t} \geq \lambda(t) \right) \leq \mathbb{P} \left(\exists t : \underbrace{\left\| \sum_{i=1}^t \varphi_{\tau_i} \left(\frac{X_i^\top \theta^* - Y_i}{\varsigma_i} \right) \tilde{X}_i \right\|_{H_t^{-1}}}_{\text{self-normalized}} + 2\sqrt{\beta}S \geq C_0(\gamma)\lambda(t) \right). \quad (4.46)$$

Remark: the prior work Li and Sun [2024] offer an analysis similar to our Theorem 27 in the proof of their Theorem 2.1. While the proof techniques are similar, their focus is on the specific setting of Huber regression for heavy-tailed rewards in bandit and RL. Our result, developed independently from Mathieu [2022], focuses on the unifying analysis and offers a more comprehensive study that ties together confidence analysis across various models and estimators under a common set of conditions. The connection we establish with the general sub- ψ conditions provided by our framework (Section 4.3.4) also eliminates the need to union bound over T while controlling the deviation. This should improve upon the result of Li and Sun [2024], see discussion in Section 4.6.2.

4.6.2 Bounding Martingale Vector Process with Heavy-Tail Noise

The self-normalized quantity in equation 4.46 can be written as follows, where $\tilde{\eta}_i = \frac{\eta_i}{\varsigma_i}$,

$$\left\| \sum_{i=1}^t \varphi_{\tau_i} \left(\frac{X_i^\top \theta - Y_i}{\varsigma_i} \right) \tilde{X}_i \right\|_{H_t^{-1}} = \left\| \sum_{i=1}^t \varphi_{\tau_i}(\tilde{\eta}_i) X_i \right\|_{H_t^{-1}} = \left\| \sum_{i=1}^t \tau_i \varphi_1 \left(\frac{\tilde{\eta}_i}{\tau_i} \right) X_i \right\|_{H_t^{-1}}.$$

Consider a concrete example of heavy-tailed noise, where η_i has bounded second moment only. The sub- ψ conditions are not directly satisfied – the magnitude of $\tau_t = \tau_0 \sqrt{1 + \frac{1}{w_t^2}}$ could be as large

as $\mathcal{O}(\sqrt{t})$. However, with the data-dependent loss parameters, there exists a solution (albeit more complex) for the heavy-tail case. Below, we summarize on a high level the core techniques from [Li and Sun, 2024, Huang et al., 2023] for bounding the self-normalized process for heavy-tail noise. We also point out that connecting to the sub- ψ framework should be able to remove a $\mathcal{O} \log(T)$ factor caused by the union bound.

In particular, they show that $\left\| \sum_{i=1}^t \mathbf{r}_i \varphi_1\left(\frac{\tilde{\eta}_i}{\mathbf{r}_i}\right) X_i \right\|_{H_t^{-1}}$ can be decomposed into two sums. The two terms can then be bounded individually (they satisfy the sub- ψ condition). This is because although the magnitude of r_t could be as large as $\mathcal{O}(\sqrt{t})$ in the worst case, it also interacts with X_t and H_t . Specifically, let $S_t = \sum_{i=1}^t \mathbf{r}_i \varphi_1\left(\frac{\tilde{\eta}_i}{\mathbf{r}_i}\right) X_i$. Then,

$$\|S_t\|_{H_t^{-1}}^2 = \|S_{t-1}\|_{H_{t-1}^{-1}}^2 + \underbrace{2\mathbf{r}_0 \varphi_1\left(\frac{\tilde{\eta}_t}{\mathbf{r}_t}\right) \|S_{t-1}\|_{H_{t-1}^{-1}}}_{\#1} + \underbrace{\mathbf{r}_0^2 \varphi_1^2\left(\frac{\tilde{\eta}_t}{\mathbf{r}_t}\right)}_{\#2} \quad (4.47)$$

An informal derivation for this part is deferred to Section 4.9.3. Bounding the two sums rely on similar technical tools, so we focus the upcoming discussion on the sum of the first term as an example. Specifically, consider bounding the sum of $\sum_{i=1}^t \mathbf{r}_0 \varphi_1\left(\frac{\tilde{\eta}_i}{\mathbf{r}_i}\right)$. Assume that φ_1 is bounded. The random variable $\varphi_1\left(\frac{\tilde{\eta}_i}{\mathbf{r}_i}\right)$ has (conditioned on \mathcal{F}_{i-1}) mean 0, and an upper bound of its variance is proportional to $\frac{1}{\mathbf{r}_i^2}$. Prior works use fixed-time Freedman's inequality to bound the summation. However fixed-time Freedman's inequality does not yield time-uniform results and needs to be followed by a union bound over $t \in [T]$ for the bandit setting. This union bound results in a factor of $\log\left(\frac{t^2}{\delta}\right)$. We expect that this can be avoided by using the sub- ψ conditions, and provide a sketch below.

For a process $(\tilde{S}_t, \tilde{V}_t)$ where $\tilde{S}_t = \sum_{i=1}^t \varphi_1\left(\frac{\tilde{\eta}_i}{\mathbf{r}_i}\right)$ and $\tilde{V}_t = \Theta\left(\sum_{i=1}^t \sigma_i^2\right)$, it is trivial to see that they satisfy a sub- ψ relationship, where $\psi = \psi_{P, c=1}$ is the sub-Poisson function. (A variant of Design Principle 1 is satisfied, in other words.) Hence, Theorem 1(b) of Howard et al. [2021] gives a time-uniform freeman-style martingale concentration of \tilde{S}_t that, with high probability $1 - \delta$,

$$\forall t : \tilde{S}_t \leq \sqrt{2\tilde{V}_t \log\left(\frac{1}{\delta}\right)} + \frac{1}{3} \log\left(\frac{1}{\delta}\right). \quad (4.48)$$

For linear bandits, the cumulative variance $\tilde{V}_t = \mathcal{O}\left(\frac{d \log(1 + \frac{T}{d})}{\mathbf{r}_0^2}\right)$, by the elliptical potential lemma. Compared to the results from fixed-time Freedman-style inequality, equation 4.48 is better by a factor of $\mathcal{O}(\log(t))$. Sum of the second term can be bounded using similar techniques. Although the mean of the second term $\mathbb{E}[\varphi_1^2\left(\frac{\tilde{\eta}_t}{\mathbf{r}_t}\right)]$ is greater than 0, making the sum a submartingale, the submartingale can be decomposed into a martingale and a predictable increasing component (the sum of the mean).

4.7 Discussion

In this work, we draw inspiration from the use of influence functions on analyzing the behavior of M -estimators. By extending the analysis to the bandit setting and employing an existing general solution for martingale process tail bounds, we present a framework that shows a general and natural way to analyze UCB-type bandit algorithms with M estimators. Our findings offer a unifying lens that ties together various settings (linear, kernelised, and generalized linear models, heavy-tailed noise) from the literature that were previously studied case by case.

An interesting future direction on top of our work is to bring theoretical insights to empirical experiments. The empirical study of Koh and Liang [2017] approximate influence functions for neural networks and use it for several downstream tasks such as understanding model behaviors and adversarial training even if the theoretical assumptions about model differentiability and convexity were violated. Inspired by their work, it will be of practical interest to see whether influence functions can be computed and used to guide bandit algorithms under various problem structures.

4.8 Proof of Theorem 27

Proof. The proof of Theorem 27 is mostly inspired by the proof of Mathieu [2022]. In time step t , given the H_t in Design Principle 1, define $u_t(\theta)$ to be the unit vector under the Mahalanobis distance induced by H_t . Recall that $\hat{\theta}_t$ is the unique solution to the risk minimization problem defined in equation 4.9.

$$u_t(\theta) := \frac{\theta - \hat{\theta}_t}{\|\theta - \hat{\theta}_t\|_{H_t}}. \quad (4.49)$$

One can verify with simple algebra that $\|u_t(\theta)\|_{H_t} = \sqrt{u_t(\theta)^\top H_t u_t(\theta)} = 1$.

Next, define the following quantity which is the directional derivative of the empirical risk evaluated at $\theta - \lambda u_t(\theta)$ along the direction $u_t(\theta)$:

$$f_t(\lambda, \theta) := \langle \nabla_{\theta} J_t(\theta - \lambda u_t(\theta)), u_t(\theta) \rangle. \quad (4.50)$$

For a scalar $\lambda \in \mathbb{R}$, note that under the event where $\lambda \leq \|\theta - \hat{\theta}_t\|_{H_t}$, the point $\theta - \lambda u_t(\theta)$ will be between θ and the global minimizer $\hat{\theta}_t$. Given the *convexity* of the risk-minimization problem, the directional derivative measured at that point, along a direction moving *away* from the global minimizer $\hat{\theta}_t$, will be non-negative. In other words,

$$\left\{ \|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda \right\} \Rightarrow \{f_t(\lambda, \theta) \geq 0\}, \quad \text{hence } \mathbb{P}\left(\|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda\right) \leq \mathbb{P}(f_t(\lambda, \theta) \geq 0). \quad (4.51)$$

Relationship between $f_t(0, \theta)$ and the deviation of influence function.

Note that, when $\lambda = 0$, $f_t(0, \theta)$ is the directional derivative evaluated at target parameter θ and is directly related to the deviation of influence function.

$$\begin{aligned} f_t(0; \theta) &= \langle \nabla J_t(\theta), u_t(\theta) \rangle \\ &= \nabla J_t(\theta)^\top I_d u_t(\theta) \\ &= \nabla J_t(\theta)^\top H_t^{-1/2} H_t^{1/2} u_t(\theta) \quad (H_t \text{ is PD and symmetric}) \\ &\leq \|H_t^{-1/2} \nabla J_t(\theta)\| \cdot \|H_t^{1/2} u_t(\theta)\| \\ &= \|\nabla J_t(\theta)\|_{H_t^{-1}} \cdot \|u_t(\theta)\|_{H_t} \\ &= \|\nabla J_t(\theta)\|_{H_t^{-1}} \quad (\text{because } \|u_t(\theta)\|_{H_t} = 1 \text{ by definition}), \end{aligned} \quad (4.52)$$

$$= \left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i + 2\tilde{\beta}\theta \right\|_{H_t^{-1}} \quad \text{deviation of “influence function”} \quad (4.53)$$

$$\leq \left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}} + 2\tilde{\beta} \|\theta\|_{H_t^{-1}}. \quad (4.54)$$

Here, equation 4.53 is exactly the norm (in Mahalanobis distance) of the influence function without the inverse matrix term. And equation 4.54 shows that it can be further decomposed into a “self-normalized” vector process (which we later connect to sub- ψ martingale vector process in Section 4.3.4.1), as well as a simpler bias term. Specifically, for the bias term we have:

$$\begin{aligned}
2\tilde{\beta}\|\theta\|_{H_t^{-1}} &= 2\tilde{\beta}\|H_t^{-1/2}\theta\|_2 \\
&\leq 2\tilde{\beta}\|\theta\|_2\|H_t^{-1/2}\|_{op} \\
&= 2\tilde{\beta}\|\theta\|_2\frac{1}{\sqrt{\lambda_{\min}(H_t)}} \\
&\leq 2\sqrt{\tilde{\beta}}\|\theta\|_2 \leq 2\sqrt{\tilde{\beta}}S.
\end{aligned} \tag{4.55}$$

The last inequality was because $\lambda_{\min}(H_t) \geq \tilde{\beta}$, from Design Principle 1.

Putting equation 4.55 and equation 4.54 together, the deviation of the influence function is bounded as such:

$$\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top\theta) - Y_i)X_i + 2\tilde{\beta}\theta \right\|_{H_t^{-1}} \leq \left\| \sum_{i=1}^t \varphi(\mu(X_i^\top\theta) - Y_i)X_i \right\|_{H_t^{-1}} + 2\sqrt{\tilde{\beta}}S. \tag{4.56}$$

Connection between $f_t(\lambda, \theta)$ and $f_t(0, \theta)$

The connection between $f_t(\lambda, \theta)$ and $f_t(0, \theta)$ can essentially be quantized by the minimum eigenvalue of $\text{Hess}(\theta)$. By Assumption S.1, $f_t(\lambda, \theta)$ is differentiable in λ . Taking the derivative of $f_t(\lambda, \theta)$ with respect to λ , we get:

$$\begin{aligned}
f'_t(\lambda; \theta) &= \frac{\partial}{\partial \lambda} f_t(\lambda; \theta) \\
&= -u_t(\theta)^\top \text{Hess}(\theta - \lambda u_t(\theta)) u_t(\theta) \\
&= -\|u_t(\theta)\|_{\text{Hess}(\theta - \lambda u_t(\theta))}^2 \\
&< 0.
\end{aligned} \tag{4.57}$$

The last inequality was because $\text{Hess}(\theta - \lambda u_t(\theta))$ is a positive-definite matrix (trivially from Assumption S.5). This means $f_t(\lambda, \theta)$ is *decreasing* with λ . By the mean value theorem, there exists $c \in (0, \lambda)$, such that

$$f_t(\lambda, \theta) = f_t(0, \theta) + \lambda \cdot f'_t(c, \theta) \tag{4.58}$$

$$\leq f_t(0, \theta) - \lambda \inf_{c \in [0, \lambda]} |f'_t(c; \theta)| \quad (f'_t(c, \theta) < 0). \tag{4.59}$$

If Design Principle 1 is satisfied, then with probability $1 - \delta_H$, $|f'_n(c; \theta)|$ is lower bounded with a positive constant:

$$\begin{aligned}
|f'_t(c; \theta)| &= \|u_t(\theta)\|_{\text{Hess}(\theta - cu_t(\theta))}^2, \quad c \in (0, \lambda) \\
&= u_t(\theta)^\top \text{Hess}(\theta - cu_t(\theta)) u_t(\theta) \\
&= \frac{(\theta - \hat{\theta}_t)^\top \text{Hess}(\theta - cu_t(\theta)) (\theta - \hat{\theta}_t)}{\|\theta - \hat{\theta}_t\|_{H_t}^2} \quad \text{by definition of } u_t(\theta) \text{ in equation 4.49} \\
&= \frac{\|\theta - \hat{\theta}_t\|_{\text{Hess}(\theta - cu_t(\theta))}^2}{\|\theta - \hat{\theta}_t\|_{H_t}^2} \\
&\geq C_0.
\end{aligned} \tag{4.60}$$

The last inequality was due to the conditions specified in Design Principle 1. Let $\vartheta = \theta - cu_t(\theta)$, then

$$\begin{aligned} \|\vartheta - \theta\| &= \|cu_t(\theta)\| = \left\| c \frac{\theta - \hat{\theta}_t}{\|\theta - \hat{\theta}_t\|_{H_t}} \right\| \\ &= \frac{c}{\|\theta - \hat{\theta}_t\|_{H_t}} \|\theta - \hat{\theta}_t\| \\ &\leq \|\theta - \hat{\theta}_t\| \quad (c < \lambda \text{ and } \lambda < \|\theta - \hat{\theta}_t\|_{H_t}) \\ &\leq 2S \quad (\text{triangular inequality}). \end{aligned}$$

Therefore, the events satisfy

$$\{f_t(\lambda, \theta) \geq 0\} \Rightarrow \left\{ f_t(0, \theta) \geq \lambda \inf_{c \in [0, \lambda]} |f'_t(c; \theta)| \right\} \Rightarrow \{f_t(0, \theta) \geq \lambda C_0\} \quad (4.61)$$

Anytime-valid confidence sequence based on deviation of the influence function

Recall that equation 4.51 combined with equation 4.61 gives

$$\{\|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda\} \Rightarrow \{f_t(0, \theta) \geq \lambda C_0\}. \quad (4.62)$$

Now, for a sequence of time steps $t = 1 \dots T$ and the accompanying sequence of matrices in Principle 1 $(H_t)_{t=1 \dots T}$, we can deduce that, for a process $(\lambda(t))_{t=1 \dots T}$:

$$\begin{aligned} \mathbb{P}(\exists t, \|\theta - \hat{\theta}_t\|_{H_t} \geq \lambda(t)) &\leq \mathbb{P}(\exists t : f_t(0, \theta) \geq \lambda \inf_{c \in (0, \lambda)} |f'_t(c; \theta)|) \\ &= 1 - \mathbb{P}(\forall t : f_t(0, \theta) \leq \lambda \inf_{c \in (0, \lambda)} |f'_t(c; \theta)|) \\ &\leq 1 - \mathbb{P}(\{\text{Good event in Lemma 1}\} \cap \{\forall t : f_t(0, \theta) \leq C_0 \lambda\}) \\ &\leq \delta_H + \mathbb{P}(\exists t : f_t(0, \theta) \geq C_0 \lambda(t)) \\ &\leq \delta_H + \mathbb{P}(\exists t : \underbrace{\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i + 2\tilde{\beta}\theta \right\|_{H_t^{-1}}}_{\text{deviation of IF}} \geq C_0 \lambda(t)) \\ &\leq \delta_H + \mathbb{P} \left(\exists t : \left(\underbrace{\left\| \sum_{i=1}^t \varphi(\mu(X_i^\top \theta) - Y_i) X_i \right\|_{H_t^{-1}}}_{\text{deviation of unregularized score function}} + \underbrace{2\sqrt{\tilde{\beta}}S}_{\text{bias term}} \geq C_0 \lambda(t) \right) \right) \end{aligned} \quad (4.63)$$

□

4.9 Auxiliary Derivations

4.9.1 Comparison of Sub- ψ_N Concentration Results

1. [Abbasi-Yadkori et al. \[2011\]](#) derived self-normalized concentration for vector process with sub-Gaussian increments specifically. According to Lemma 9 of [Abbasi-Yadkori et al. \[2011\]](#), for

$\delta > 0$,

$$\mathbb{P} \left(\exists t, \left\| \left(\sum_{i=1}^t \eta_i X_i \right) \right\|_{H_t^{-1}} \geq \sigma \sqrt{2d \log \left(\frac{1 + tL^2/\beta}{\delta} \right)} \right) \leq \delta. \quad (4.64)$$

2. As mentioned in Section 4.3.4.1, Whitehouse et al. [2023b] provides a more general sub- ψ characterization that unifies existing martingale vector concentration results. Since the noise is conditionally sub-Gaussian, the process $S_t = \sum_{i=1}^t \eta_i X_i$ is sub- ψ_N with respect to variance proxy V_t , where ψ_N represents the sub-Gaussian condition. Their result is in terms of the condition number of the variance proxy $\kappa(V_t) = \frac{\lambda_{\max}(V_t)}{\lambda_{\min}(V_t)}$, while the result from Abbasi-Yadkori et al. [2011] is in terms of its determinant.³

$$\mathbb{P} \left(\exists t, \left\| \left(\sum_{i=1}^t \eta_i X_i \right) \right\|_{H_t^{-1}} \geq \mathcal{O}(\sqrt{\log \log \lambda_{\max}(H_t/\beta) + d \log \kappa(H_t)}) \right) \leq \delta, \quad (4.65)$$

$$\Rightarrow \mathbb{P} \left(\exists t, \left\| \left(\sum_{i=1}^t \eta_i X_i \right) \right\|_{H_t^{-1}} \geq \mathcal{O}(\sqrt{\log \log(1 + tL^2/\beta) + d \log(1 + tL^2/\beta)}) \right) \leq \delta \quad (4.66)$$

Note that $\lambda_{\max}((\sum X_i X_i^\top + \beta \mathbf{I}_d)/\beta) \leq \frac{tL^2}{\beta} + 1$. And $\lambda_{\min}(H_t) \geq \beta$. $\kappa(H_t) \leq \frac{tL^2}{\beta} + 1$.

4.9.2 GP-UCB Uses Confidence Ellipsoid in RKHS

It is easy to see that, using the close-form representation of the kernel ridge estimator

$$\hat{f}_t = \sum_{i=1}^t (\mathbf{K}(\mathbf{X}_t, \mathbf{X}_t) + 2\beta \mathbf{I}_t)^{-1} \Phi(X_i) Y_i,$$

the predicted mean (the first term) is $\langle \hat{f}_t, \Phi(X_t) \rangle_{\mathcal{H}} = k(x, \mathbf{X}_t) (\mathbf{K}_t + \beta \mathbf{I}_t)^{-1} \mathbf{Y}_t$. This exactly recovers the first term ($\mu_t(x)$) in the UCB expression from Srinivas et al. [2009] (by setting $\beta = \sigma^2$ their prior variance). For the second term,

$$\begin{aligned} \|\Phi(x)\|_{A_t^{-1}, \mathcal{H}} &= \sqrt{\langle \Phi(x), A_t^{-1} \Phi(x) \rangle_{\mathcal{H}}} \\ &= \sqrt{\langle \Phi(x), (\Phi_t^\top \Phi_t + \beta \text{id}_{\mathcal{H}})^{-1} \Phi(x) \rangle_{\mathcal{H}}} \\ &\stackrel{(1)}{=} \sqrt{\langle \Phi(x), \left(\frac{1}{\beta} \text{id}_{\mathcal{H}} - \frac{1}{\beta} \Phi_t^\top (\beta \mathbf{I}_t + \Phi_t \Phi_t^\top)^{-1} \Phi_t \right)^{-1} \Phi(x) \rangle_{\mathcal{H}}} \\ &\stackrel{(2)}{=} \sqrt{\langle \Phi(x), \left(\frac{1}{\beta} \text{id}_{\mathcal{H}} - \frac{1}{\beta} \Phi_t^\top (\beta \mathbf{I}_t + k(\mathbf{X}_t, \mathbf{X}_t))^{-1} \Phi_t \right)^{-1} \Phi(x) \rangle_{\mathcal{H}}} \\ &= \sqrt{\frac{1}{\beta} k(x, x) - \frac{1}{\beta} k(x, \mathbf{X}_t) (\beta \mathbf{I}_t + \mathbf{K}(\mathbf{X}_t, \mathbf{X}_t))^{-1} \mathbf{K}(\mathbf{X}_t, x)}. \end{aligned}$$

In the above, (1) is from using Woodbury matrix identity and (2) is from using the kernel trick: $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. The final expression exactly recovers the second (variance) term in the UCB expression: $\sigma_t(x) = \sqrt{k(x, x') - \mathbf{k}_t(x)^\top (\mathbf{K}_t + \sigma^2 \mathbf{I})^{-1} \mathbf{k}_t(x')}$, up to a factor of $\sqrt{1/\beta}$.

³Whitehouse et al. [2023b] remark that these two bounds (for the sub-Gaussian case) are in general non-comparable, which is tighter would be decided on a case-by-case basis depending on the structure of V_t .

4.9.3 Auxiliary Derivations for Section 4.6.2

$$\begin{aligned}
\|S_t\|_{H_t^{-1}}^2 &= \|S_{t-1}\|_{H_{t-1}^{-1}}^2 + 2r_t\varphi_1\left(\frac{\eta_t}{r_t}\right)S_{t-1}^\top H_t^{-1}X_t + r_t^2\varphi_1^2\left(\frac{\eta_t}{r_t}\right)X_t^\top H_t^{-1}X_t \\
&= \|S_{t-1}\|_{H_{t-1}^{-1}}^2 + 2r_t\varphi_1\left(\frac{\eta_t}{r_t}\right)\frac{1}{1+w_t^2}S_{t-1}^\top H_{t-1}^{-1}X_t + r_t^2\varphi_1^2\left(\frac{\eta_t}{r_t}\right)\frac{w_t^2}{1+w_t^2} \text{ by Woodbury matrix identity} \\
&= \|S_{t-1}\|_{H_{t-1}^{-1}}^2 + 2\frac{r_0}{\sqrt{1+w_t^2}}\varphi_1\left(\frac{\eta_t}{r_t}\right)\|S_{t-1}\|_{H_{t-1}^{-1}} + r_t^2\varphi_1^2\left(\frac{\eta_t}{r_t}\right)\frac{w_t^2}{1+w_t^2} \\
&= \|S_{t-1}\|_{H_{t-1}^{-1}}^2 + \underbrace{2r_0\varphi_1\left(\frac{\eta_t}{r_t}\right)\|S_{t-1}\|_{H_{t-1}^{-1}}}_{\#1} + \underbrace{r_0^2\varphi_1^2\left(\frac{\eta_t}{r_t}\right)}_{\#2}
\end{aligned}$$

4.9.4 Proof of Lemma 29

Proof. Recall that the condition for a vector process $S_t \in \mathbb{R}^d$ to be sub- ψ with some cumulant generating function (CGF)-like function ψ , with respect to some variance proxy process $V_t \in \mathbb{R}^{d \times d}$ is as follows:

$$\exp(\lambda\langle v, S_t \rangle - \psi(\lambda)\langle v, V_t v \rangle) \leq L_t^{\lambda, v}, \forall v \in \mathbb{S}^{d-1}. \quad (4.67)$$

Where $L_t^{\lambda, v}$ is a non-negative supermartingale. By induction, let $\exp(\lambda\langle v, S_{t-1} \rangle - \psi(\lambda)\langle v, V_{t-1} v \rangle) = L_{t-1}^{\lambda, v}$, if we could prove that

$$\mathbb{E}_{t-1}[\exp(\lambda\langle v, S_t \rangle - \psi(\lambda)\langle v, V_t v \rangle)] = L_{t-1}^{\lambda, v} \times \mathbb{E}_{t-1}[\exp(\lambda\langle v, \Delta S_t \rangle - \psi(\lambda)\langle v, \Delta V_t v \rangle)] \leq L_{t-1}^{\lambda, v}, \quad (4.68)$$

that would conclude the proof. Which means we need to show that

$$\mathbb{E}_{t-1}[\exp(\lambda\langle v, \Delta S_t \rangle - \psi(\lambda)\langle v, \Delta V_t v \rangle)] \leq 1,$$

for $\Delta S_t = \eta_t X_t$ and $\Delta V_t = \mathbb{E}_{i-1} \Delta S_t \Delta S_t^\top = \mathbb{E}_{i-1}[\eta_t^2] X_t X_t^\top$.

Recall that $\eta_i \mid \mathcal{F}_{i-1}$ follows a centered Poisson distribution. According to the MGF of Poisson distribution,

$$\begin{aligned}
\mathbb{E}_{t-1}[\exp(\lambda\langle v, \Delta S_t \rangle)] &= \mathbb{E}_{t-1}[\exp(\lambda v^\top X_t \eta_t)] \\
&= \exp(\mathbf{p}(e^{\lambda v^\top X_t} - \lambda v^\top X_t - 1)), \quad \mathbf{p} = \mu(X_t^\top \theta^*) \\
&\leq \exp \mathbf{p} \left(\sum_{k=2}^{\infty} \frac{\lambda^k |v^\top X_t|^k}{k!} \right) \quad (\text{Taylor expansion}).
\end{aligned}$$

Since $\Delta V_t = \mathbb{E}_{t-1}[\eta_t^2] X_t X_t^\top$, we have

$$\begin{aligned}
\psi_{P,c}(\lambda)\langle v, \Delta V_t v \rangle &= \frac{e^{c\lambda} - c\lambda - 1}{c^2} \text{Var}_{t-1}(\eta_t)(v^\top X_t)^2 \\
&= \mathbf{p}(e^{c\lambda} - c\lambda - 1)(v^\top X_t)^2 c^{-2} \quad \text{variance of Poisson distribution} \\
&= \mathbf{p} \left(\sum_{k=2}^{\infty} \frac{c^{k-2} \lambda^k (v^\top X_t)^2}{k!} \right)
\end{aligned}$$

Therefore, if we choose c larger or equal to $\max_{x \in \mathcal{X}} |v^\top x|$, then

$$\begin{aligned} \mathbb{E}_{t-1}[\exp(\lambda \langle v, \Delta S_t \rangle)] &\leq \exp(\psi_{P,c}(\lambda) \langle v, \Delta V_t v \rangle) \Rightarrow \\ \mathbb{E}_{t-1}\left[\frac{\exp(\lambda \langle v, \Delta S_t \rangle)}{\exp(\psi_{P,c}(\lambda) \langle v, \Delta V_t v \rangle)}\right] &\leq 1 \Rightarrow \\ \mathbb{E}_{t-1}[\exp(\lambda \langle v, \Delta S_t \rangle - \psi_{P,c}(\lambda) \langle v, \Delta V_t v \rangle)] &\leq 1. \end{aligned}$$

Since $v \in \mathbb{S}^{d-1}$, we know that $|v^\top X_t| \leq L$, so a simple choice for c is $c = L$. This concludes the proof. \square

Chapter 5

Conclusion and Future Directions

This thesis provides generalizations for the bandit optimization framework in algorithm design and theoretical guarantees. The first part of this thesis (Chapter 2 and Chapter 3) generalizes the reward structures and problem assumptions. Chapter 2 presents a family of minimax optimal algorithms for Hölder function spaces, which capture all degrees of smoothness and encompass both linear and Lipschitz spaces studied by prior works as extreme cases, closing the gap for bandit optimization of smooth differentiable functions. In Chapter 3 we remove the potentially-stringent assumption which assumes that the algorithm knows the exact smoothness of the reward function space. We answer the theoretical question of how well an algorithm can do without the a priori knowledge of the smoothness. Our results span across Reproducing Kernel Hilbert Space, Sobolev space and Hölder space, and provide a conclusive characterization of an algorithm’s best possible performance without the aforementioned key knowledge of the reward function space. The second part of this thesis (Chapter 4) has a somewhat different flavor, as it provides a novel generalized way to think of theoretical analysis for a family of existing bandit algorithms that were developed to operate under different problem structures and were seemingly distinct. The analytical framework we propose makes novel connections to existing statistical methods, highlights the fundamental factors that underline different settings, and provides as a unifying interpretation of such bandit algorithms.

This thesis establishes groundwork for future explorations in several interesting and non-trivial directions. Future extensions can be found in the discussion section of each chapter. Here, we summarize several important directions, from more specific ones to higher-level contemplations. The first direction is the improvement of algorithm and theoretical analysis from “worst-case” to benign cases, where additional structures can be leveraged. In Chapter 2, we focus on the “worst-case” bounds. Such bounds are satisfied for all functions in the target function space, but might not be optimal for some easier-to-optimize instances within. For example, an additional “growth” condition of functions around the optima was proposed in prior works [Kleinberg et al., 2008, Bubeck et al., 2010] to further represent how easy it is to optimize a continuous function. To adapt to these conditions, prior works develop methods for Lipschitz reward functions that achieve better cumulative regret bounds than the worst-case bounds. These methods (such as the adaptive discretization method in Bubeck et al. [2010]) can potentially be seen as orthogonal to our algorithms in Chapter 2. Thus, it is interesting to see whether our algorithms can leverage these methods and further take advantage of benign cases in the broader Hölder function class.

The second future direction, which is related to our framework in Chapter 4, is optimality for individual cases. The current framework, despite its advantage of unifying across different scenarios, does not yield optimal results in certain cases. More specifically, the dependence on certain parameters is not as tight as the state-of-the-art. For an example, see the remarks in Section 4.5.3. It is worth extending and improving it so that it is able to recover optimal results for each of the specific settings.

More broadly, with the rise of increasingly powerful machine learning methods and the growing number of domains that have embraced them, many of which involve expensive queries and scarce data, there is a need to advance both the theory and methods of sequential decision-making. To adapt these methods to the evolving landscape, it is essential to develop uncertainty quantification techniques that are both general, so they can accurately capture uncertainty under complex structures, and robust, so that their guarantees remain valid even when the problem structures are misspecified.

Furthermore, it is worth understanding when theoretical insights about these uncertainty quantification methods remain practical enough to guide active algorithms in experiments, even when the assumptions required to obtain theoretical guarantees are not strictly satisfied. For example, the connection to influence functions explored in Chapter 4 grew from our contemplation of these issues, and the use of influence functions in both classical statistics and empirical tasks with neural networks (Section 4.7). Although our study remains in a purely theoretical regime, future efforts should study whether the insights of using influence functions to guide sequential decision-making (bandit) algorithms can also be valuable empirically. Chapter 4 represents our step towards these broader goals for sequential decision-making.

To conclude, we believe that the contributions in this thesis and the future efforts for which we pave the way enhance the bandit framework in essential and meaningful directions.

Bibliography

- Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.
- M. Abeille, L. Faury, and C. Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.
- R. A. Adams and J. J. Fournier. *Sobolev spaces*. Elsevier, 2003.
- A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corraling a band of bandit algorithms. *arXiv preprint arXiv:1612.06246*, 2016.
- A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corraling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.
- A. Akhavan, M. Pontil, and A. B. Tsybakov. Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *arXiv preprint arXiv:2006.07862*, 2020.
- R. Arora, T. V. Marinov, and M. Mohri. Corraling stochastic bandit algorithms. *arXiv preprint arXiv:2006.09255*, 2020.
- R. Arora, T. V. Marinov, and M. Mohri. Corraling stochastic bandit algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 2116–2124. PMLR, 2021.
- S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.
- P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331. IEEE, 1995.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. In *International Conference on Computational Learning Theory*, pages 454–468. Springer, 2007.
- M. Avella-Medina. Influence functions for penalized m-estimators. *Bernoulli*, 23(4B):3178–3196, 2017.

- F. Berkenkamp, A. P. Schoellig, and A. Krause. No-regret bayesian optimization with unknown hyperparameters. *arXiv preprint arXiv:1901.03357*, 2019.
- A. Bietti and F. Bach. Deep equals shallow for relu networks in kernel regimes. *arXiv preprint arXiv:2009.14397*, 2020.
- I. Bogunovic and A. Krause. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34:3004–3015, 2021.
- S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. X-armed bandits. *arXiv preprint arXiv:1001.4475*, 2010.
- S. Bubeck, G. Stoltz, and J. Y. Yu. Lipschitz bandits without the lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011.
- A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.
- X. Cai and J. Scarlett. On lower bounds for standard and robust gaussian process bandit optimization. In *International Conference on Machine Learning*, pages 1216–1226. PMLR, 2021.
- D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557. PMLR, 2019.
- O. Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l’IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.
- L. Chen and S. Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.
- X. Chen, D. Simchi-Levi, and Y. Wang. Utility fairness in contextual dynamic pricing with demand learning. *arXiv preprint arXiv:2311.16528*, 2023.
- L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.
- S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.
- W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.
- V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, pages 355–366, 2008.
- S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.
- L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

- S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.
- D. J. Foster, Z. Li, T. Lykouris, K. Sridharan, and E. Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, pages 4734–4742, 2016.
- D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 14741–14752, 2019.
- J. A. Grant and D. S. Leslie. On thompson sampling for smoother-than-lipschitz bandits. *arXiv preprint arXiv:2001.02323*, 2020.
- Y. Gur, A. Momeni, and S. Wager. Smoothness-adaptive stochastic bandits. *arXiv preprint arXiv:1910.09714*, 2019.
- H. Hadiji. Polynomial cost of adaptation for x-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.
- M. Hoffmann, R. Nickl, et al. On adaptive inference and confidence bands. *The Annals of Statistics*, 39(5):2383–2409, 2011.
- S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.
- Y. Hu, N. Kallus, and X. Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010, 2020.
- J. Huang, H. Zhong, L. Wang, and L. Yang. Tackling heavy-tailed rewards in reinforcement learning with function approximation: Minimax optimal and instance-dependent regret bounds. *Advances in Neural Information Processing Systems*, 36:56576–56588, 2023.
- P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.
- A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- D. Janz, D. Burt, and J. González. Bandit optimisation of functions in the matern kernel rkhs. In *International Conference on Artificial Intelligence and Statistics*, pages 2486–2495. PMLR, 2020.
- D. Janz, S. Liu, A. Ayoub, and C. Szepesvári. Exploration via linearly perturbed loss minimisation. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2024.
- K. Kandasamy, W. Neiswanger, R. Zhang, A. Krishnamurthy, J. Schneider, and B. Póczos. Myopic posterior sampling for adaptive goal oriented design of experiments. In *International Conference on Machine Learning*, pages 3222–3232. PMLR, 2019.
- P. Kassraie and A. Krause. Neural contextual bandits without regret. *arXiv preprint arXiv:2107.03144*, 2021.
- P. Kassraie and A. Krause. Neural contextual bandits without regret. In *International Conference on Artificial Intelligence and Statistics*, pages 240–278. PMLR, 2022.

- P. Kassraie, J. Rothfuss, and A. Krause. Meta-learning hypothesis spaces for sequential decision-making. *arXiv preprint arXiv:2202.00602*, 2022.
- R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.
- R. D. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *Advances in Neural Information Processing Systems*, pages 697–704, 2005.
- P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.
- A. Krishnamurthy, J. Langford, A. Slivkins, and C. Zhang. Contextual bandits with continuous actions: Smoothing, zooming, and adapting. *arXiv preprint arXiv:1902.01520*, 2019.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- T. Lattimore. A lower bound for linear and kernel regression with adaptive covariates. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2095–2113. PMLR, 2023.
- T. Lattimore and C. Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.
- T. Lattimore, C. Szepesvari, and G. Weisz. Learning with good feature representations in bandits and in rl with a generative model. In *International Conference on Machine Learning*, pages 5662–5670. PMLR, 2020.
- J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.
- J. Lee, S.-Y. Yun, and K.-S. Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *arXiv preprint arXiv:2407.13977*, 2024.
- L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.
- L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2017.
- L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185): 1–52, 2018.
- X. Li and Q. Sun. Variance-aware robust reinforcement learning with linear function approximation with heavy-tailed rewards. *arXiv preprint arXiv:2303.05606*, 2023.
- X. Li and Q. Sun. Variance-aware decision making with linear function approximation under heavy-tailed rewards. *Transactions on Machine Learning Research*, 2024.

- Y. Liu and A. Singh. Adaptation to misspecified kernel regularity in kernelised bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4963–4985. PMLR, 2023.
- Y. Liu, Y. Wang, and A. Singh. Smooth bandit optimization: Generalization to holder space. In *International Conference on Artificial Intelligence and Statistics*, pages 2206–2214. PMLR, 2021.
- A. Locatelli and A. Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492. PMLR, 2018.
- M. G. Low et al. On nonparametric confidence intervals. *The Annals of Statistics*, 25(6):2547–2554, 1997.
- B. Matern et al. Spatial variation. stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden fran Statens Skogsforskningsinstitut*, 49(5), 1960.
- T. Mathieu. Concentration study of m-estimators using the influence function. *Electronic Journal of Statistics*, 16(1):3695–3750, 2022.
- T. Mathieu, D. Basu, and O.-A. Maillard. Bandits corrupted by nature: Lower bounds on regret and robust optimistic algorithms. *Transactions on Machine Learning Research*, 2022.
- W. Neiswanger and A. Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic Learning Theory*, pages 963–982. PMLR, 2021.
- D. M. Ostrovskii and F. Bach. Finite-sample analysis of m-estimators using self-concordance. *Electronic Journal of Statistics*, 15:326–391, 2021.
- A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020a.
- A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33:10328–10337, 2020b.
- P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742. PMLR, 2017.
- S. Shekhar and T. Javidi. Multi-scale zero-order optimization of smooth functions in an rkhs. *arXiv preprint arXiv:2005.04832*, 2020.
- S. Singh. Continuum-armed bandits: A function space perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 2620–2628. PMLR, 2021.
- N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.
- Q. Sun, W. Zhou, and J. Fan. Adaptive huber regression: Nonasymptotic optimality and phase transition. *arXiv preprint arXiv:1706.06991*, 2017.

- J. A. Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
- A. B. Tsybakov. Introduction to nonparametric estimation, 2009. URL <https://doi.org/10.1007/b13794>. Revised and extended from the, 9(10), 2004.
- A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.
- S. Vakili, M. Bromberg, J. Garcia, D.-s. Shiu, and A. Bernacchia. Uniform generalization bounds for overparameterized neural networks. *arXiv preprint arXiv:2109.06099*, 2021a.
- S. Vakili, J. Scarlett, and T. Javidi. Open problem: Tight online confidence intervals for rkhs elements. In *Conference on Learning Theory*, pages 4647–4652. PMLR, 2021b.
- M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.
- M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Y. Wang, S. Balakrishnan, and A. Singh. Optimization of smooth functions with noisy observations: Local minimax rates. *Advances in Neural Information Processing Systems*, 31, 2018.
- H. Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.
- J. Whitehouse, A. Ramdas, and S. Z. Wu. On the sublinear regret of gp-ucb. *Advances in Neural Information Processing Systems*, 36:35266–35276, 2023a.
- J. Whitehouse, Z. S. Wu, and A. Ramdas. Time-uniform self-normalized concentration for vector-valued processes. *arXiv preprint arXiv:2310.09100*, 2023b.
- C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.
- D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.
- Y. Zhu and R. Nowak. Pareto optimal model selection in linear bandits. *arXiv preprint arXiv:2102.06593*, 2021.