Carnegie Mellon University

Department of Statistics & Data Science
Machine Learning Department

Trustworthy Scientific Inference with Machine Learning

Luca Masserano

Thesis Committee:

Ann B. Lee (Chair)
Barnabás Póczos (MLD Mentor)
Mikael Kuusela
Jing Lei
Cosma Shalizi
Rafael Izbicki (Federal University of São Carlos)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics and Machine Learning.

To my family.

Abstract

The application of AI and machine learning to complex scientific problems is becoming increasingly widespread across various fields. A key challenge of scientific inference is to derive parameter constraints that are both valid — meaning they include the true parameter regardless of its (unknown) value at a specified confidence level, even in finite samples — and precise — meaning they are as small as possible given the data-generating process. However, standard machine learning approaches often fail to ensure that these properties hold, thereby limiting the reliability of downstream scientific conclusions. In this dissertation, we introduce several novel techniques to leverage regression, classification, and generative models to construct confidence sets with strong statistical guarantees. The methods we develop allow one to derive confidence sets that are simultaneously (1) valid across the entire parameter space and in finite samples, (2) robust to prior probability shifts, (3) as precise as possible when prior knowledge aligns with the target distribution, and (4) computationally efficient. By bridging modern machine learning with classical statistical tools, we provide a principled path towards integrating AI into scientific inference and discovery pipelines, enabling advancements in fields such as astronomy, high-energy physics, biology, and beyond.

Acknowledgments

I remember when, more than five years ago, I started writing the statement of purpose to apply for PhD programs. I was in my room at Collegio di Milano, wondering how best to impress the admissions committees of prestigious schools, with Ksenija lying on the bed behind me, probably thinking, "What have I gotten myself into?". I revisited that document, which at some point said:

> *A doctoral degree would represent an extraordinary accomplishment, concluding a journey that brought me from grass pitches to the highest achievable degree and opening the doors to a fulfilling research career.*

I definitely did not know much of what was about to unroll in front of me for the following five years. One of the many things I did not know was how many wonderful people I would have met along the way, and how much each of them would have shaped the person I have become. Now that this journey is over, this is my attempt to thank all of them.

First and foremost, I would like to express my deepest and most sincere gratitude to my advisor, Ann Lee, for her guidance and support throughout my PhD, and for showing me the ropes of how to do research. I am particularly grateful for her patience and for deeply caring about me, especially in moments where I struggled professionally or personally. I am honored to have worked besides her during these five years, and to have witnessed what true, unlimited and tireless passion for research looks like.

I would like to thank Barnabás Póczos, Rafael Izbicki, Mikael Kuusela and Tommaso Dorigo for being invaluable mentors and close collaborators over the years, and for sharing many insights and intuitions that made me a better researcher. I also want to extend my gratitude to Jing Lei and Cosma Shalizi for being part of my thesis committee and for providing their precious feedback. In addition, I would like to thank all the members of the statistical methods for the physical sciences (STAMPS) research group. Being part of it has been a tremendous learning experience, and I cannot wait to see its success in the future now that it has become an official CMU research center.

Finally, my gratitude goes also to Alessandro Rinaldo for supporting me through tough times during the first two years of the PhD; to Michele Doro, Aishik Ghosh and Joshua Speagle for their feedback and collaboration on applications in astroparticle physics, high-energy

physics and astronomy; and to Igor Prünster, Antonio Lijoi and Daniele Durante for pushing me to explore the idea of a PhD and for immensely helping me during the application process.

To my longtime friends, Giovanni and Federico. We live thousands of kilometers apart and see each other one or two times per year, but every time we talk it seems like we have always been together. I feel extremely fortunate to have both of you in my life and to be able to witness how your lives are unfolding. Your friendship and advice are inestimable to me, and I hope one day to go back living closer to each other to share our lives as we used to do. Thank you for everything you have meant and will always mean to me.

My gratitude goes also to Frank, Edoardo and Gerard, who are still my dear friends despite my notorious sloppiness in replying to messages. Thank you for your patience — I look forward to sharing many more precious moments with you in the years ahead.

I have been fortunate to share this journey with many incredible friends and colleagues. This acknowledgement section would not be complete without sincerely thanking them.

To Lorenzo, Maya and Rebecca. We became very close only during the last two years of the PhD, but you have quickly turned into my family in Pittsburgh and are the dearest friends I have here. Thank you for the countless days spent together, the fantastic meals we shared, the tires we changed, the long hours we worked together, the continuous support while struggling with the PhD and with life, the sleepovers, the parties, the drama, the honest advice and much, much more. I wish we had had the entire five years together, but I am sure we will find ways to cultivate our friendship in the future. I cannot wait to see what amazing things you will achieve. Thank you for making me feel part of something during times when I felt lost.

A special thanks goes also to Alec, Diego, Ian, James, Lucas, Mateo, Mike, Odalys, Sasha, Sid and YJ for all the memories, support, and work together along this journey. I would also like to thank Nic for his invaluable mentorship especially during the first year of the PhD and for becoming a friend and a person I look up to. In addition to the ones already mentioned above, I would like to acknowledge the colleagues I have had the pleasure to collaborate with during my time in the department: Alex, Antonio, David and Joshua. My gratitude goes also to the other members of my PhD cohort, who helped me keep my sanity during the fully remote first year and a global pandemic: Akshay, Anni, Catherine, Galen, Holly, Julia, Kayla, Konrad, Kyle, Raghav, Tiger, Victoria and Vinny. I would like to thank many other people, including, but not limited to, Federica, Kenta, Matteo B., Matteo G. and Neil. Finally, I cannot forget Syama and Konstantinos, whom I met during internships at Amazon and have become close friends and precious mentors.

Questa tesi è dedicata a mio padre, a mia madre, a mia sorella, a mio fratello, ai miei nonni e a Ksenija, senza i quali non avrei mai potuto farcela.

Ai miei genitori, che con i sacrifici di una vita e l'amore incondizionato mi avete dato prima la forza e gli strumenti per affrontare i momenti bui, poi la determinazione senza fine per andarsi a prendere un obiettivo, a prescindere da quanto difficile questo appaia. Il vostro esempio — anche nella capacità di affrontare le difficoltà ed i silenzi di avere un figlio così lontano — è per me fonte di ispirazione e di orgoglio quotidiano.

A mia sorella e mio fratello, con cui continuo a crescere e condividere le nostre vite ormai

da adulti. Nononstante la distanza che ci separa da tredici anni, so di poter sempre contare su di voi. Sono così felice e fortunato di avervi accanto in questo viaggio.

Ai miei nonni, che con la loro saggezza mi hanno insegnato a “non mollare mai”.

Ed infine, alla mia K. *Samo ti i ja stvarno znamo kroz šta smo prošli tokom ovih pet godina. Zapravo ne znam o čemu si razmišljala onog dana dok sam ja pisao „statement of purpose“ u Koledu, ali dobro se sećam kako nijedno od nas nije imalo nikakvu sumnju o tome šta ćemo raditi kada smo shvatili da ćemo biti razdvojeni. Suočili smo se sa razdaljinom i pobedili je, nastavljajući da gradimo našu vezu kroz ono ‘držim te na FaceTime-u dok radim’, jutarnje i večernje poruke, putovanja svakog raspusta i mnogo toga što je stvorilo našu svakodnevicu uprkos udaljenosti. Ali pre svega, uvek smo se držali za ruke, koračali usklađeno i slepo verovali u našu međusobnu ljubav. Hvala ti što si bila moje svakodnevno utočište, moja najbolja prijateljica i najbolji partner kojeg sam ikada mogao poželeti. Ne mogu ti opisati koliko se osećam srećnim što si pored mene, i jedva čekam da započnemo naš zajednički život.*

Contents

Abstract i

Acknowledgments ii

Contents v

List of Tables ix

List of Figures x

1 Introduction 1
- 1.1 Summary of Contributions . . . 4

2 Background: Likelihood-Free Frequentist Inference 6
- 2.1 Introduction . . . 6
- 2.2 Statistical Inference in a Traditional Setting . . . 9
- 2.3 Likelihood-Free Frequentist Inference via Odds Estimation . . . 10
  - 2.3.1 Estimating an Odds Function across the Parameter Space . . . 10
  - 2.3.2 Test Statistics based on Odds . . . 11
  - 2.3.3 Fast Construction of Neyman Confidence Sets . . . 12
  - 2.3.4 Diagnostics: Checking Coverage across the Parameter Space . . . 14
- 2.4 Theoretical Guarantees . . . 15
  - 2.4.1 Critical Value Estimation . . . 15
  - 2.4.2 P-Value Estimation . . . 16
  - 2.4.3 Power of `BFF` . . . 17
- 2.5 Handling Nuisance Parameters . . . 19
- 2.6 Experiments . . . 20
  - 2.6.1 Gaussian Mixture Model: Unknown Null Distribution . . . 20
  - 2.6.2 Poisson Counting Experiment: Nuisance Parameters and Diagnostics 22
  - 2.6.3 Muon Energy Estimation: Intractable and High-Dimensional Likelihood 24
- 2.7 Conclusions and Discussion . . . 26
  - 2.7.1 Related Work . . . 28

3 Confidence Sets from Prediction Algorithms and Posterior Estimators 30
3.1 Introduction . . . 30
3.2 Related Work . . . 33
3.3 Methodology . . . 34
3.3.1 Foundational Tools from Classical Statistics . . . 34
3.3.2 Confidence Sets from Predictions and Posteriors . . . 34
3.3.3 Statistical Properties: Coverage and Power . . . 37
3.3.4 Computational Properties . . . 38
3.4 Experiments . . . 40
3.4.1 Confidence Sets from Neural Posteriors . . . 40
3.4.2 Confidence Sets for Muon Energies using CNN Predictions . . . 42
3.5 Conclusions and Discussion . . . 43

4 Optimal Confidence Sets from Generative Models 45
4.1 Introduction and Problem Setting . . . 45
4.2 Results . . . 49
4.2.1 Case Study I: Reconstructing Gamma-Ray-Induced Air Showers with Ground-Based Detector Arrays . . . 49
4.2.2 Case Study II: Inferring Properties of Milky Way Stars in Simulation-Based Inference . . . 51
4.2.3 Case Study III: Inferring Stellar Parameters from Cross-Matched Astronomical Catalogs ("LFI Beyond SBI") . . . 53
4.3 Methods . . . 55
4.3.1 Experimental Set-Up . . . 55
4.3.2 A Protocol for Valid Scientific Inference . . . 56
4.4 Conclusions . . . 57

5 Inference under Nuisance Parameters and Generalized Label Shift 61
5.1 Introduction . . . 61
5.2 Related Work . . . 64
5.3 Methodology . . . 65
5.3.1 Classification as Hypothesis Testing . . . 65
5.3.2 The Rejection Probability Across the Entire Parameter Space . . . 65
5.3.3 Selecting the Optimal Cutoff under GLS . . . 66
5.3.4 Constructing Robust Set-Valued Classifiers . . . 67
5.4 Theoretical Results . . . 68
5.4.1 Validity and Robustness to GLS . . . 68
5.5 Experiments . . . 69
5.5.1 Synthetic Example . . . 69
5.5.2 Single-Cell RNA Sequencing . . . 70
5.5.3 Atmospheric Cosmic-Ray Showers . . . 72
5.6 Conclusion and Discussion . . . 74

6 The `lf2i` package 77
6.1 Description of the Main Components . . . 77

6.2 Related Software 79

7 Extensions and Future Work **81**
7.1 Anytime-Valid Sequential Likelihood-Free Inference 81
7.2 LF2I for Data Assimilation 83
7.3 Improving on the LRT by Leveraging Prior Distributions in Particle Physics 84

Bibliography **86**

A Additional Results for Chapter 2 **107**
A.1 Estimating Odds 107
A.2 Estimating p-values 107
A.3 Constructing Confidence Sets 108
A.4 Theoretical Guarantees of Power for `ACORE` with Calibrated Critical Values 108
A.5 Analysis of Critical Values for Experiments 2.6.1 and 2.6.2 111
A.6 Additional Proofs 113
A.7 Loss Functions 119

B Additional Results for Chapter 3 **120**
B.1 Additional Experiments 120
B.1.1 Property III: Estimating the Conditional Variance Matters 120
B.1.2 Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture 121
B.1.3 Confidence Sets for Muon Energies using CNN Predictions 122
B.2 Details on Models, Training, and Computational Resources 122
B.2.1 Synthetic Examples for Statistical Properties 122
B.2.2 Synthetic Example for Computational Properties 123
B.2.3 Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture 124
B.2.4 Confidence Sets for Muon Energies using CNN Predictions 124

C Additional Results for Chapter 4 **125**
C.1 Relation to Other Methodology 125
C.2 Constructing Confidence Procedures with Frequentist Coverage 127
C.2.1 Fast Construction of Confidence Procedures from Posterior Estimates 128
C.2.2 Validity of Frequentist Bayes Procedure 132
C.3 Power of Frequentist Bayes Procedure 134
C.4 Details on Synthetic Examples of Section 4.1 135
C.4.1 Synthetic Example of Figure 4.2 135
C.4.2 Synthetic Example of Figure 4.6 136
C.5 Supplement for Case Study I 137
C.5.1 Experimental Setup 137
C.5.2 Data 138
C.5.3 Details on Training 139
C.6 Supplement for Case Study II 139
C.6.1 Experimental Set-Up 139

C.6.2 Data . . . 140
C.6.3 Details on Training . . . 141
C.6.4 Additional Results . . . 141
C.7 Supplement for Case Study III . . . 141
C.7.1 Data . . . 141

D Additional Results for Chapter 5 **144**
D.1 The Bayes Factor as a Frequentist Test Statistic . . . 144
D.2 Proofs . . . 144
D.3 Estimating the Rejection Probability Function . . . 145
D.4 Diagnostics of Estimated ROC Curves . . . 146
D.5 The Standard Bayes Classifier . . . 146
D.6 Additional Results and Details on Cosmic Ray Experiment . . . 147
D.6.1 Experimental Set-Up with Ground-Based Detector Arrays . . . 147
D.6.2 Details on the algorithms used in Section 5.5.3 . . . 148
D.6.3 Additional Results . . . 149
D.7 Additional Results and Details on the RNA Sequencing experiment . . . 149
D.7.1 Data Simulation Procedure . . . 149
D.7.2 Details on the algorithms used in Section 5.5.2 . . . 151
D.7.3 Additional Results . . . 151
D.8 Computational Analysis: Training and Inference Times . . . 151
D.9 Synthetic Example: Deep Dive . . . 154
D.9.1 Impact of the Nuisance Parameter . . . 154
D.9.2 Additional Results . . . 155
D.9.3 ν-Conditional Coverage and validity under GLS . . . 155
D.9.4 When does $\gamma > 0$ for NAPS increase power? . . . 156
D.9.5 Performance of NAPS under SLS . . . 161

List of Tables

4.1 **Scientific inference challenges addressed in this chapter.** Each case study in Sections 4.2.1, 4.2.2 and 4.2.3 (with the set-up listed in the right column) illustrates a unique scientific challenge, which we resolve with our proposed approach. *Right Column,* **I:** Ground-based detector array for measuring atmospheric cosmic-ray showers (proposed SWGO experiment; Abreu et al. 2019). **II:** Two differing models of the galaxy, simulated using BRUTUS (Speagle et al., 2025). **III:** Galactic map displaying the stars included in a cross-match between Gaia Data Release 3 (Gaia Collaboration et al., 2023) and APOGEE Data Release 17 (Majewski et al., 2017). 60

C.1 Galactic model parameters 140
C.2 True stellar parameters for the displayed star in Section 4.2.2 140
C.3 True stellar parameters for the additional example star in Section C.6.4 142

D.1 Training and inference times for NAPS for the experiments of Sections 5.5.2 and 5.5.3. 154

List of Figures

1.1 **AI is increasingly being used across several fields of science to improve our understanding of natural phenomena.** Some of the most notable examples come from high-energy physics (*left panel*: event recorded at the CMS detector at the LHC in Geneva), biology (*central panel*: example of a protein structure prediction from `AlphaFold`), and astronomy and astrophysics (*right panel*: illustration of gravitational waves.) 2

1.2 **Likelihood-Free Inference setup:** given a collection of data pairs $\{(\theta_i, X_i)\}_{i=1}^{B} \sim p(X \mid \theta)\pi(\theta)$ from a mechanistic model that implicitly encodes the intractable likelihood $\mathcal{L}(\theta; X)$, LFI aims to infer the true $\theta^\star$ that generated a new $x^{\text{obs}} \sim p(X \mid \theta)p^{\text{obs}}(\theta)$. . . 3

2.1 **The three-branch fully modular framework for likelihood-free frequentist inference (LF2I). Center branch:** Draw a sample $\mathcal{T}$ of size B from the simulator to estimate an arbitrary test statistic $\lambda(\mathcal{D}; \theta)$. Here we show how to do so by estimating the likelihood via the odds function $\mathbb{O}(X; \theta)$. **Left branch:** Draw a second sample $\mathcal{T}'$ of size B' to estimate the critical values C_θ or p-values $p(\mathcal{D}; \theta)$ for all $\theta \in \Theta$. **Left + Center:** Once data D are observed, we can construct confidence sets $\widehat{\mathcal{R}}(D)$ with finite-n validity according to Equation (2.12). **Right branch:** The LF2I diagnostics branch independently checks whether the coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ of the confidence set is indeed correct across the entire parameter space. 9

2.2 **Neyman construction of confidence sets by inverting hypothesis tests. Left:** For each $\theta_0 \in \Theta$, we find the critical value C_{θ_0} that rejects the null hypothesis H_{0,θ_0} at level α; that is, C_{θ_0} is the α-quantile of the distribution of the test statistic under the null (a likelihood ratio $\text{LR}(\mathcal{D}; \theta_0)$ in this case). **Right:** The horizontal solid lines represent acceptance regions for each $\theta_0 \in \Theta$. Suppose we observe data D. The confidence set for θ (red vertical solid line) consists of all θ_0-values for which the observed test statistic $\text{LR}(D; \theta_0)$ (black curve) falls in the acceptance region. 11

2.3 **GMM with unknown null distribution**. Each panel shows the estimated coverage across the parameter space of 90% confidence sets for θ. Rows represent experiments with different observed sample sizes: $n = 10, 100, 1000$ (top, center, bottom). Columns represent three different approaches. **Left:** "LR with Monte Carlo samples" achieves nominal coverage everywhere but is computationally expensive, especially in higher dimensions. **Center:** "Chi-square LRT" clearly under-covers, i.e. confidence sets are not valid even for large n, other than at $\theta = 0$ where the mixture collapses to one Gaussian. **Right:** "LR with C_{θ_0} via quantile regression" returns finite-sample confidence sets with the nominal coverage of 90% for all values of θ, but using a total of 1000 simulations, instead of a MC sample of 1000 simulations at each grid point. 21

2.4 **Poisson counting experiment with nuisance parameters.** The diagnostics branch provides guidance as to which LFI approach to use for the problem at hand by pinpointing regions of the parameter space Θ where inference is unreliable. The panels show empirical coverage as a function of both μ, the parameter of interest, and ν, the nuisance parameter. Nominal coverage is 90%. **Left:** `h-ACORE`, which uses profiled likelihoods, is overly conservative in terms of actual coverage ($\approx 96\%$) across Θ. **Center:** `h-BFF`, which marginalizes over ν, under-covers in several regions (red crosses). **Right:** `ACORE` χ^2_1, which uses cutoffs from the chi-square distribution, has almost no constraining power, yielding empirical coverage close to 100% everywhere. 23

2.5 **Constraining power.** Relative size of the confidence sets constructed in Section 2.6.2. `ACORE` χ^2_1 and `h-ACORE` yield the widest intervals (they are indeed overly conservative according to Figure 2.4). `h-BFF` provides tighter confidence sets, but their size cannot be trusted when the method under-covers. LF2I diagnostics can identify the parameter regions where the approach is not valid (red crosses in Figure 2.4). The dark-orange histogram reports `h-BFF` results after removing those points. 24

2.6 **Muon energy estimation.** LF2I guarantees nominal coverage and yields smaller confidence intervals relative to SMC-ABC. **Left:** Data point example of a muon with incoming energy $\theta \approx 3.2$ TeV entering a calorimeter with $32 \times 32 \times 50$ cells. **Center:** LF2I (blue, orange, red in the right two panels) achieves coverage at the nominal level (68.3%), whereas SMC-ABC (green and purple) is consistently over-covering across the parameter space. **Right:** Median lengths of constructed intervals. While being extremely computationally intensive, SMC-ABC has also the least constraining power regardless of the data set used. SMC-ABC on the full calorimeter data is not reported as it was computationally infeasible to run. 25

3.1 **Schematic diagram of `Waldo`. Left (blue)**: For a training set $\mathcal{T}$, we estimate the conditional mean $\mathbb{E}[\theta \mid \mathcal{D}]$ and variance $\mathbb{V}[\theta \mid \mathcal{D}]$ using a prediction algorithm (e.g., DNN) or posterior estimator (e.g., normalizing flows). This gives us the `Waldo` test statistic $\hat{\tau}_{\texttt{Waldo}}$ in Equation (3.4). **Center (green):** For a calibration set $\mathcal{T}'$, we estimate critical values $\widehat{C}_{\theta_0,\alpha}$ for all tests $H_0 : \theta = \theta_0$ across the parameter space Θ via a quantile regression of $\hat{\tau}_{\texttt{Waldo}}$ on θ. **Bottom:** Given an observation D, Neyman inversion converts the tests (which compare test statistics with critical values) into a confidence region for θ. **Right (red):** For a validation set $\mathcal{T}''$, we provide an independent assessment of the conditional validity of constructed confidence regions by computing coverage diagnostics across the entire parameter space. See Section 3.3.2 and Algorithm 3.1 for details. 32

3.2 **Property I: `Waldo` guarantees conditional coverage across Θ, regardless of the specified prior.** Prior: $\theta \sim \mathcal{N}(0, 2)$. Likelihood: $\mathcal{D} \mid \theta \sim \mathcal{N}(\theta, 1)$. **Left:** median of upper/lower bounds of constructed parameter regions. **Right:** empirical coverage computed numerically using 100,000 samples for each θ over a fine grid in Θ (i.e., not using coverage diagnostics). 38

3.3 **Property II: `Waldo` exploits prior information and achieves higher power.** Power curves computed by recording the number of times a wrong value of θ is correctly outside the confidence set over 1,000 repetitions. Likelihood: $\mathcal{D} \sim \mathcal{N}(40, 1)$. **Left:** Wald and `Waldo` are equivalent when $\theta \sim \mathcal{U}(35, 45)$. **Right:** `Waldo` has higher power when $\theta \sim \mathcal{N}(40, 1)$. 39

3.4 **Quantile regression (QR) is orders of magnitude more efficient than Monte Carlo (MC) in terms of the number of simulations B' required to achieve correct coverage.** Each panel shows the fraction of samples (out of 1,000 total) for which the selected method to estimate critical values achieves approximately correct coverage ($\mathbb{P}(\theta \in \mathcal{R}(\mathcal{D}) \mid \theta) \in [0.95 \pm 0.03]$). Prior: $\theta \sim \mathcal{N}(0, 0.1 \cdot I)$. Likelihood: $\mathcal{D} \mid \theta \sim \mathcal{N}(\theta, 0.1 \cdot I)$. In both cases, we used normalizing flows to estimate the posterior. 40

3.5 **`Waldo` converts posterior distributions into confidence regions with correct conditional coverage and high power**. **Left Panel - Top:** Examples of 95% credible regions (blue) from posteriors estimated with normalizing flows and a Gaussian $\mathcal{N}(0, 2 \cdot I)$ prior (gray) for different values of the true unknown parameter θ^* (red star). **Right Panel - Top:** Credible regions have conditional coverage close to the nominal level only in a neighborhood of the prior, and severely undercover everywhere else. **Left Panel - Bottom:** Corresponding 95% `Waldo` confidence sets (green), derived from the same posterior estimates used for the top row. **Right Panel - Bottom:** Conditional coverage for `Waldo` confidence sets achieves the nominal 1-α level everywhere, where $\alpha = 0.05$. 41

3.6 **`Waldo` guarantees the nominal coverage level, and yields smaller confidence intervals (more precise estimates of muon energy) with the higher-granularity ("full") calorimeter data. Left:** Energy deposited by a $\theta \approx 3.2$ TeV muon entering a calorimeter with $32 \times 32 \times 50$ cells. **Center:** `Waldo` (blue, orange, red in the right two panels) guarantees nominal coverage (68.3%), while 1σ prediction intervals (green) under- or over-cover in different regions of Θ. **Right:** Median lengths of constructed intervals: shorter intervals imply higher precision in the estimates. Prediction sets are on average wider than the corresponding confidence sets, using the same data. 42

4.1 **The likelihood-free inference setting. Panel A**: With a forward model, we can make predictions on data X given parameters θ. The inverse problem is to infer the parameters θ of a model given observed data X. **Panel B**: In likelihood-free inference (LFI), the likelihood $p(X \mid \theta)$ is intractable. We consider two LFI scenarios, where the likelihood is implicitly encoded either by (i) a simulator (the inverse problem is then known as simulator-based inference or SBI; brown), or by (ii) labeled data from observational studies (we refer to the latter inverse problem as "LFI beyond SBI", green). 46

4.2 **Our proposed approach to valid scientific inference. Panel A:** *(Left)* The typical workflow for inferring parameters with neural density estimators is to first learn the posterior, $\widehat{\pi}(\theta|X)$, from train data. Then, for new observed data X_{obs}, one slices $\widehat{\pi}(\theta|X_{obs})$ to compute a highest-posterior density (HPD) set. The purple and pink intervals at the bottom depict 95% and 68% HPD sets, respectively, for an observation whose true parameter (indicated by a red star) lies in the tail of the prior $\pi(\theta)$. *(Right)* The actual chance (coverage probability, y-axis) that the two HPD sets contain the true parameter value can be far less than what the nominal coverage of 95% and 68%, respectively, suggest, for a wide range of different θ-values (x-axis). **Panel B**: *(Left)* Recalibration — our approach effectively transforms the posterior to a p-value function, which we then slice to obtain valid ("Frequentist-Bayes"; FreB) confidence sets. *(Right)* The actual chance (coverage probability, y-axis) that FreB sets contain the true parameter value is indeed close to the desired coverage probability for every instance of θ (x-axis). . . . 47

4.3 **Posterior-based methods lack local coverage guarantees and thus fail to reliably reconstruct gamma-ray showers from unfamiliar sources. Panel A:** *(Top)* Distribution of three gamma-ray sources in energy and zenith angle. An example gamma-ray event/shower at high energies is indicated by a red marker. *(Bottom)* Detector data for example event, showing arrival times at different locations. **Panel B:** *(Top)* Estimated local coverage of 90% HPD sets of individual events (averaged over azimuth) reveals undercoverage, especially at higher energies. *(Bottom)* Distribution of coverage across events from each gamma-ray source; coverage drops when training and target sources are different. **Panel C:** *(Top)* Local coverage of 90% FreB sets instead shows uniform validity across the parameter space. *(Bottom)* Coverage distribution per gamma-ray source confirms consistent validity regardless of source. **Panel D:** Comparison for a high-energy event from the Crab Nebula (for the same example event as in Panel A, Top): the 90% HPD set (purple) is overconfident and biased (actual coverage is 78%), while the 90% FreB set (green) provides valid and informative uncertainty. 50

4.4 **FreB resolves tension between differing galactic models. Panel A:** *(Left)* The age-metallicity relationships implied by two Galactic models. The red curves indicate conditional means of metallicity given age. **Panel A:** *(Right)* Surface-level priors induced by the galactic models along line of sight $(70^\circ, 30^\circ)$. Log gravitational constant ($\log g$), effective temperature (T_{eff}), and surface metallicity ($[Fe/H]_{\text{surf}}$) are shown. The true label for a typical object is marked in red, unseen at inference time. **Panel B:** Tension between Models H and D's posteriors at $X \sim p(X|\theta)$. Solid contours for each show 90% credible regions of high posterior density, marginalized. The HPD regions feature 0% conditional coverage. **Panel C:** 90% FreB sets for θ for Models H and D. Each subplot shows cross-sections of the FreB sets at the true label. Conditional coverage for each FreB set is close to the nominal 90% level. 52

4.5 **FreB is robust to label bias in observational studies. Panel A:** Kiel diagrams displaying the training distribution of stellar gravities $\log g$ against the corresponding effective temperatures T_{eff}) for two data settings, where the labeled data are biased towards the asymptotic giant branch stars (*left*, "AGB Label Bias"), and where the labeled and unlabeled target data have the same distribution (*center*, "No Label Bias"). *(Right)* An example spectrum for a Sun-like star, for which the true label marked in red is unknown. **Panel B:** *(Left)* 90% HPD sets under the two selection settings, with the HPD set under the AGB selection bias not including the true label (red). *(Right)* Local coverage plot of 90% HPD sets in the held-out main sequence (MS) parameter space, showing under-coverage for all labels. **Panel C:** *(Left)* 90% FreB sets under the two selection settings, with the FreB set under both settings covering the true (red) label, but with higher constraining power with well-aligned training data. *(Right)* Local coverage plot of 90% FreB sets in the held-out main sequence (MS) parameter space, showing nearly nominal coverage for all labels. 54

4.6 **FreB sets are simultaneously robust against misaligned priors and small in size for well-aligned priors.** Synthetic two-dimensional example where the task is to infer the location θ of a mixture of two Gaussians with different covariances, $X \sim \frac{1}{2}\mathcal{N}(\theta, \sigma_1^2 I) + \frac{1}{2}\mathcal{N}(\theta, \sigma_2^2 I)$, using a posterior learned with a Flow Matching generative model trained with a localized prior, $\pi(\theta) = \mathcal{N}(0, 2)$. **Panel A:** 95% and 68% HPD sets for two scenarios where the prior is misaligned (*left*) versus well-aligned (*center*) with the true θ. *(Right)* Local coverage plot of 95% HPD sets shows that the actual coverage of these sets can be very far from the nominal 95% level, when the truth is further away from the center where the prior is concentrated. **Panel B:** Corresponding FreB sets obtained from the same posterior estimated via the same generative model as in Panel A. For all instances of θ and for all levels of α, domain scientists are guaranteed to achieve the desired coverage level, here illustrated for the 95% case in the *right* plot. That is, FreB sets are robust against misaligned priors. Moreover, the size of FreB sets is smaller for well-aligned priors (compare *center* plot with the *left* plot). 58

5.1 **Synthetic Example. Left (no GLS):** Standard prediction sets $\mathcal{R}_\alpha(x)$ (red) guarantee marginal coverage at the nominal level. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) are also marginally valid, but the "universality" of conditional validity across the entire nuisance parameter space comes at the price of more conservative prediction sets and lower power. **Right (with GLS):** Standard prediction sets are no longer valid and undercover for all α levels (red curve is below the black bisector), while NAPS are still valid. Furthermore, we can increase power while maintaining validity (NAPS $\gamma > 0$; green) by constructing $(1-\gamma)$ confidence sets of the nuisance parameter ν and deriving less conservative cutoffs given an observation. Here $\gamma = \alpha \times 0.01$. 62

5.2 **Coverage under different batch protocols ν for the RNA-Seq example.** Each marker represents the proportion of samples in the test set whose true label was included in the constructed prediction sets. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) are valid regardless of the protocol, which is unknown at inference time. All other methods for prediction sets with marginal coverage (red), class-conditional coverage (pink), and conformal adaptive prediction sets (gold) undercover for at least two batch protocols. 71

5.3 **Dependence of the ROC on the energy of the cosmic-ray shower. Left:** Receiver operating characteristic evaluated according to our method at different energy values (shades of blue). By estimating the entire ROC, we can control FPR or TPR at specified confidence levels for all $\nu \in \mathcal{N}$, which is not possible with the "marginal" ROC curve (red). **Right:** Diagnostic P-P plot evaluated at four bins over energy for nuisance-aware ROC (shades of blue) and ROC that ignores nuisances (shades of red). To check if $\mathbb{P}_{\text{target}}(\lambda(X) \leqslant C \mid y, \nu)$ is well estimated, we plot PIT values against a Uniform$(0, 1)$ distribution (dashed bisector; see Appendix D.4 for details). This is clearly not the case if one ignores nuisance parameters. 73

5.4 **Constraining the cosmic ray shower parameters. Top left**: Illustration of the Southern Wide-field Gamma-ray Observatory (SWGO; Abreu et al. (2019); image credit: Richard White) array of detectors with an incoming gamma ray (red). **Bottom Left**: Test statistic under $y_0 = 0$ (hadron) as a function of energy. At high energies, the class-conditional test statistics are well separated, implying that it is easier to distinguish gamma showers (red) from hadron showers (gold). **Right**: Confidence set for ν at different $(1 - \gamma)$ confidence levels obtained via the framework of Masserano et al. (2023). The true value of ν is the black star. 74

5.5 **Classification metrics within true and within predicted Gamma rays $(y = 1)$.** Results are binned according to whether the shower energy is below (left) or above (right) the median value. **Top panel:** Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) achieve high precision and low false discovery rates (FDR), especially at high confidence levels. In addition, by constraining the nuisance parameters $\nu = (E, A, Z)$, we can increase performance (NAPS $\gamma > 0$; green) with uniformly better results relative to the standard Bayes classifier (black dashed line). **Bottom panel:** Our set-valued classifier makes explicit its level of uncertainty on the label y by returning ambiguous prediction sets (bottom row) for hard-to-classify x_{target}. Even so, NAPS with $\gamma > 0$ is able to achieve a higher number of true positives and lower number of false negatives relative to the Bayes classifier. Here $\gamma = \alpha \times 0.3$. 75

6.1 **Main components of the `lf2i` package.** 78

7.1 **Confidence sequences and credible regions for the mean of a $Beta(10, 30)$ distribution.** HPD sets correctly concentrate around the true parameter as the posterior estimates improve from additional training data. Confidence sequences achieve validity, but remain very conservative. 83

7.2 **Median 68.3% confidence interval length for LRT (red) and for BF (solid blue curve) with a truncated normal prior distribution (dashed blue curve).** Observations are sampled from different values of μ to show the gain of power around the prior and the loss of power far from it. 85

A.1 **Comparison of critical values** obtained via Monte Carlo, the Chi-Square asymptotic assumption of Wilks' Theorem, and LF2I Quantile Regression, for the GMM example of Section 2.6.1. 112

A.2 **Critical values of `h-ACORE` estimated via quantile regression** as a function of the parameter of interest μ and the nuisance parameter ν, for the example of Section 2.6.2. The figures show the same 2D surface from two different angles. 112

B.1 **Property III: Estimating the conditional variance matters. Left:** Power curves at 95% confidence level when the true Pareto shape $\theta^* = 5$, implying a very skewed data distribution. **Right:** Test statistics and critical values as a function of θ. In this example, we set $n = 10$. 120

B.2 **a) When the prior is uninformative, `Waldo` can still correct for possible approximation errors in the estimated posterior. b)-c) When the prior is consistent with the data, `Waldo` tightens the confidence sets, improving the precision with respect to the case using a Uniform prior. a) and b)** Posterior credible regions and `Waldo` confidence sets using different priors. **c)** Average area of credible regions and `Waldo` confidence sets across 100 independent samples, reported as the percentage of points retained among those in the evaluation grid. 121

B.3 **Coverage diagnostics for Gaussian mixture model example with uniform prior**. We achieve correct conditional coverage for `Waldo` (**left**) but not for credible regions (**right**) even though the prior is is uniform, due to estimation and approximation errors, which `Waldo` can correct via recalibration. 122

B.4 **Confidence and prediction sets for the muon energy reconstruction experiment**. Boxplots of the upper and lower bounds of prediction sets (green) versus `Waldo` confidence sets (red) for full the calorimeter data, all divided in 19 bins over true energy. We clearly see the bias occurring in the prediction sets (especially at high energies) and the correction applied by `Waldo`. 123

C.1 **The three-branch modular framework for valid scientific inference with neural density estimators (NDE). Left branch:** Leverage a NDE to learn the posterior distribution $\pi(\theta \mid X)$ from a labeled training set $\mathcal{T}$. **Center branch:** From a universal labeled set $\mathcal{T}'$, learn amortized p-values to allow amortization for all miscoverage levels. Alternatively, learn critical values at a fixed level α. **Left + Center:** Given a new datapoint x, construct Frequentist-Bayes sets by taking level sets of the amortized p-value function, or by retaining all the values of θ for which $\widehat{\pi}(\theta \mid X)$ is larger than the corresponding critical value. **Right branch:** The coverage diagnostics branch independently checks whether the instance-wise coverage $\mathbb{P}_{X|\theta}(\theta \in B_\alpha(X))$ of the confidence set is indeed correct across the entire parameter space. 129

C.2 **Example features collected for a single gamma-ray event.** For each detector (represented by the pixels in each figure), we plot three measurements of the induced atmospheric shower. **(Left)** Average arrival time of secondary shower particles. **(Center)** Number of detections of "main" shower particles (photons, electrons, and positrons). **(Right)** Number of detections of "secondary" shower particles (muons, all other possible shower particles). 139
C.3 See caption of Figure 4.4. 142

D.1 **Left:** Artistic representation of the SWGO array. The inlay shows the individual detector unit. **Right:** Although we have access to all secondary particles in our simulated cosmic ray showers, we only include the particles that hit our simulated detector setup (blue rectangles) in the analysis. This layout pictured here is an illustrative example. 148
D.2 **Classification metrics within predicted Hadrons ($y_{\text{pred}} = 0$).** Results are binned according to whether the shower energy is below (left) or above (right) the median value. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) achieve high precision and low false discovery rates (FDR), especially at high confidence levels. In addition, by constraining the nuisance parameters $\nu = (E, A, Z)$, we see performance (NAPS $\gamma > 0$; green) increase in the lower energy bin but with a corresponding tradeoff in the higher energy bins. Both approaches yield better results relative to the oracle Bayes classifier (black dashed line). 149
D.3 **Classification metrics within true Hadrons ($y = 0$).** Results are binned according to whether the shower energy is below (left) or above (right) the median value. Our set-valued classifier makes explicit its level of uncertainty on the label y by returning ambiguous prediction sets (bottom row) for hard-to-classify x_{target}. Even so, NAPS with $\gamma > 0$ is able to achieve a comparable number true negatives in the higher energy bins and lower number of false positives in both energy bins relative to the Bayes classifier. Here $\gamma = \alpha \times 0.3$ 150
D.4 **Classification metrics within predicted positive class**: Precision (top) and FDR (bottom) for observations predicted to be $\texttt{CD4}^+$ T-cells (i.e. prediction set output is $\{1\}$), additionally separated by protocol (columns). Metrics are shown for nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value. 152
D.5 **Classification metrics within true positive class**: TPR (top), FNR (middle) and proportion of ambiguous sets (bottom) for true $\texttt{CD4}^+$ T-cells, additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value. 152

D.6 **Classification metrics within predicted negative class**: NPV (top) and False Omission Rate (bottom) for observations predicted to be `Cytotoxic` T-cells (i.e. prediction set output is $\{0\}$), additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value. 153

D.7 **Classification metrics within true negative class**: TNR (top), FPR (middle) and proportion of ambiguous sets (bottom) for true `Cytotoxic` T-cells, additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value. 153

D.8 **Impacts of Nuisance Parameters on the Inference Task Top Left**: Conditional densities $p(x \mid Y, \nu)$ for various values of Y and ν according to the problem setup. The marginal density $p(x \mid Y = 0)$ shown in red is induced by a Uniform$(1, 10)$ prior on ν. **Top Right**: Posterior probability $P(Y = 1 \mid X, \nu)$ as a function of X for different values of the nuisance parameter ν. The marginal posterior $P(Y = 1 \mid X)$ is shown in red for a Uniform$(1, 10)$ prior on ν. **Bottom Left**: ROC curves for the Bayes Classifier holding ν fixed (blue, orange, and green curves) and for a Uniform$(1, 10)$ prior on ν (red). $Y = 1$ is taken to be the positive class. **Bottom Right**: Under the classification rule that $\widehat{y}_i = 1$ if $x_i > x^*$, this figure shows how the FPR of that classifier will vary with ν. Each curve represents a different cut x^* for the classification rule. 156

D.9 **Actual vs Nominal Coverage for Several Prediction Set Methods**: We compare the actual coverage of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). We show marginal coverage (top), and conditional coverage for $Y = 0$ events (middle) and $Y = 1$ events (bottom) 157

D.10 **Power vs Nominal Coverage for Several Prediction Set Methods**: We compare the power of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). Power for $Y = 0$ events (bottom) is defined as $\mathbb{P}(1 \notin \text{Prediction Set} \mid Y = 0)$ and vice versa for $Y = 1$ (middle). Marginal power (top) is the sum of these two power metrics weighted by $\mathbb{P}(Y = 1)$. 158

D.11 **Precision vs Nominal Coverage for Several Prediction Set Methods**: We compare the precision of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). We define precision for prediction set $= \{0\}$ as $\mathbb{P}(Y = 0 \mid \text{prediction set} = \{0\})$ and vice versa for prediction set $= \{1\}$ outputs. Events where prediction set $= \{0, 1\}$ or prediction set $= \varnothing$ are not considered here. 159

D.12 **Marginal, Class-conditional, and ν-conditional Coverage of Several Prediction Set Methods**: We examine marginal coverage under the training prior on ν (top), $Y = 0$ conditional coverage (middle) and $Y = 1$ conditional coverage (bottom) for standard prediction sets (red, top right only), class-conditional prediction sets (pink, middle left and bottom left), class-conditional prediction sets with estimated ν (dark red, middle column), and NAPS (blue, right column). In each figure, we also show coverage when additionally conditioning on certain values of ν (dotted and dashed lines) 160

D.13 **Effect of γ on NAPS Power. Left**: We show how the optimization of $x_0(\nu; \alpha, \gamma)$ depends on γ and $S_0(x; \gamma)$. The two curves show the relationship between $x_0(\nu; \alpha, \gamma)$ and ν under two values of γ. When $\gamma = 0$, we must optimize over the entire space of ν to derive $x_0^*(\alpha)$ (or equivalently, $S_0(x; \gamma = 0) = [1, 10]$ for all x. This leads to a $x_0^*(\alpha)$ value indicated by the blue star. When $\gamma = 0.0025$, we consider two hypothetical confidence sets $S_0(x_1; \gamma)$ and $S_0(x_2; \gamma)$ for ν, indicated by the two pairs of green dotted lines. In each case, we only optimize $x_0(\nu; \alpha, \gamma)$ over the values of ν in the confidence set; however, to maintain coverage at $1 - \alpha$, optimization is done over the green curve instead of the blue curve. Optimization over $S_0(x_1; \gamma)$ yields $x_0^*(\alpha)$ indicated by the red star, while optimization over $S_0(x_2; \gamma)$ yields $x_0^*(\alpha)$ indicated by the green star. **Right**: When $S_0(x; \gamma)$ is taken to be the $(\gamma/2, 1-\gamma/2)$ quantiles of the truncated $\mathcal{N}(4, 0.1)$ distribution for all x, we can derive a relationship between $x_0^*(\alpha)$ and γ. In this case, the calibrated cutoff is minimized at $\gamma \approx 0.001$. 161

D.14 **Comparison of NAPS and Class-Conditional Prediction Sets under Standard Label Shift**: We plot the test set marginal coverage (top row) and marginal power (bottom row, defined as $\mathbb{P}_{\text{target}}(1 - Y \notin \text{Prediction set})$). We compare NAPS (blue) to Class-Conditional PS (pink). This comparison is done for several levels of SLS (columns), where we shift the distribution Y in the evaluation set from $\mathbb{P}_{\text{train}}(Y = 1) = 0.5$. The distribution of the nuisance parameter ν is *the same* for training versus target data; that is, we have an SLS setting. 162

1

Introduction

Recent advancements in artificial intelligence have opened unprecedented opportunities across scientific disciplines, empowering researchers to analyze natural phenomena in greater depth by leveraging complex, large-scale datasets spanning multiple modalities. In many science applications, the key challenge is often to test currently accepted theoretical models by designing experiments that can help in proving, disproving, or enriching our understanding of the physical laws governing our universe. This has been the case in recent years, for example, with the discovery of the Higgs Boson (Aad et al., 2012b; Chatrchyan et al., 2012), the AI-drive detection of gravitational waves (Huerta et al., 2021) and the development of the first deep learning model able to accurately predict the three-dimensional structure of proteins Jumper et al. (2021).

As an integral part of these efforts, scientific inference often focuses on using data to infer key parameters that govern complex data-generating processes. This data usually comes in the form of a labeled set $\{(\theta_i, X_i)\}_{i=1}^B$ collected either *i)* from a mechanistic model (i.e., a simulator) that implicitly encodes the likelihood function $\mathcal{L}(\theta; X)$ (e.g., Agostinelli et al. (2003); Song et al. (2023)), or more generally from a statistical model that cannot be evaluated (see, e.g., Davison et al. (2012)); or *ii)* from observational studies where labels can be measured with high precision at least for a subset of the data (see, e.g., Laroche and Speagle (2024)). In both cases, $F_\theta : \theta \mapsto X$ is an implicit map representing the intractable likelihood function, which defines the "causal" relationship between parameters and observable data. Scientists often have a deep knowledge of F_θ, reason why obtaining high-fidelity simulations as in *i)* can be relatively easy, although $\mathcal{L}(\theta; X)$ cannot be evaluated or is not available in closed form. Likelihood-Free Inference (LFI) deals with the hard associated inverse problem: given a *new* set of observations $D = \{x_1^{\text{obs}}, \dots, x_n^{\text{obs}}\}$ from the same distribution, the goal is to infer the parameter of interest $\theta^\star$ that generated D. See Figure 1.2 for a depiction of the typical LFI setup. While at first sight this setting might not look so dissimilar from a standard statistical inference problem, in reality it carries at least two major differences. First, as we mentioned, the likelihood is not available analytically, hence standard statistical techniques such as maximum likelihood estimation or Bayesian inference cannot be used out of the box. Second, standard statistical inference is usually concerned with inferring a single global parameter $\theta^\star$ from a sample

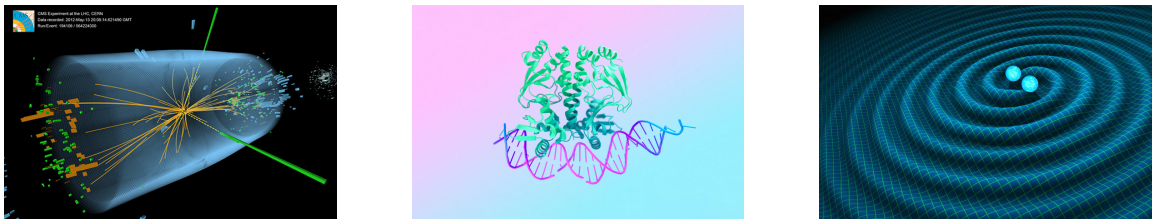


Figure 1.1: **AI is increasingly being used across several fields of science to improve our understanding of natural phenomena.** Some of the most notable examples come from high-energy physics (*left panel*: event recorded at the CMS detector at the LHC in Geneva), biology (*central panel*: example of a protein structure prediction from `AlphaFold`), and astronomy and astrophysics (*right panel*: illustration of gravitational waves.)

of size n. Throughout this thesis, instead, we will deal with the more general problem of inferring several different parameter values $\theta_j^\star, j = 1, \dots, J$, each of which generated a separate set of observations $D_j, j = 1, \dots, J$, each of size n. As such, the (mis-)alignment between the target marginal distribution $p_{\text{obs}}(\theta)$ and the source marginal distribution $\pi(\theta)$ will also play a key role in addition to the likelihood $\mathcal{L}(\theta; X)$, as we will see in all the chapters.

The most well-known approach to LFI has traditionally been Approximate Bayesian Computation (ABC; Beaumont et al. (2002); Rubin (1984); Sunnåker et al. (2013)). Loosely speaking, ABC estimates the posterior distribution $p(\theta \mid x^{\text{obs}})$ by retaining parameters associated with simulations that are close enough to the observation x^{obs}, where the distance is defined relative to a 1-dimensional summary statistic. More recently, the vast majority of research in LFI has been focusing on how to leverage machine learning (ML) algorithms to directly estimate key inferential quantities, such as

1. parameters θ in a prediction setting, as in Ho et al. (2019); Kieseler et al. (2022); Gerber and Nychka (2021);

2. posterior distributions $p(\theta \mid x^{\text{obs}})$ via neural density estimators and generative models, as in Papamakarios and Murray (2016); Lueckmann et al. (2017); Greenberg et al. (2019); Wildberger et al. (2024); Corso et al. (2023);

3. likelihoods $\mathcal{L}(\theta; X)$ and likelihood ratios $\mathcal{L}(\theta_1; X)/\mathcal{L}(\theta_2; X)$, as in Izbicki et al. (2014); Cranmer et al. (2015); Durkan et al. (2020b); Thomas et al. (2021); Walchessen et al. (2023).

These approaches can handle complex, unstructured and high-dimensional data thanks to the expressive power of neural network architectures, and can approximate complicated distributions without resorting to explicit dimensionality reduction and pre-determined summary statistics as in ABC. In addition, some of them are also amortized, meaning that the training phase happens only once and the models can then be evaluated on an arbitrary number of different observations. This last property is especially important in modern large-scale data settings, such as those arising from recent telescope surveys (Pontoppidan et al., 2022).

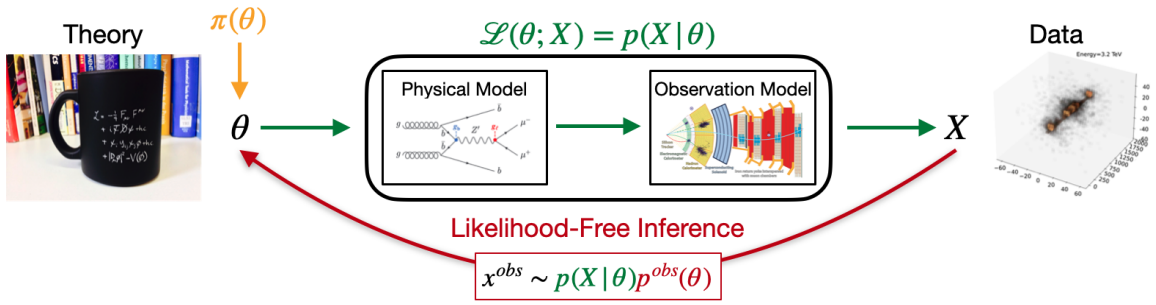


Figure 1.2: **Likelihood-Free Inference setup:** given a collection of data pairs $\{(\theta_i, X_i)\}_{i=1}^{B} \sim p(X \mid \theta)\pi(\theta)$ from a mechanistic model that implicitly encodes the intractable likelihood $\mathcal{L}(\theta; X)$, LFI aims to infer the true $\theta^\star$ that generated a new $x^{\text{obs}} \sim p(X \mid \theta)p^{\text{obs}}(\theta)$.

Nonetheless, all of these LFI methods fail to address a key challenge of scientific inference: providing constraints for parameters of interest that are both valid — meaning they include the true parameter regardless of its (unknown) value at a specified confidence level, even in finite samples — and precise — meaning they are as small as possible given the data-generating process. For example, as it was clearly shown by Hermans et al. (2021) through an extensive empirical analysis, all modern neural density estimators and variants of ABC can yield overconfident and biased posteriors, thereby making them unfit to draw reliable scientific conclusions. On one hand, this problem is caused by the reliance of ML algorithms on training data: the set $\{(\theta_i, X_i)\}_{i=1}^{B}$ is collected by sampling in regions of the parameter space dictated by a (working) prior distribution $\theta \sim \pi(\theta)$. If $\pi(\theta)$ is not consistent with $p_{\text{obs}}(\theta)$, meaning that it places most of its mass far from $\theta^\star$, then it will introduce a possibly harmful bias. On the other hand, the advances in these methods are mainly driven from a machine learning perspective, which causes a discrepancy between the ML evaluation criteria — targeting the exactness of an approximation — and the scientific evaluation criteria — which should instead target trustworthy uncertainty quantification.

In this thesis, we propose several advances to fill these gaps by developing LFI procedures to construct confidence sets that are simultaneously

1. Valid across the entire parameter space and in finite samples (in fact, even if $n = 1$);
2. Robust to prior probability shifts — i.e., validity is guaranteed under mis-specification of the prior with respect to the target distribution over θ;
3. As precise as possible when prior knowledge aligns with the target distribution;
4. Computationally efficient and scalable to high-dimensional data and parameter spaces, without compromising amortization.

More specifically, we will show how to leverage arbitrary machine learning models, such as regression, classification and generative models, to obtain confidence sets $\mathcal{R}(x^{\text{obs}})$ such that,

for $\alpha \in (0, 1)$,

$$\mathbb{P}_{X|\theta}\left(\theta \in \mathcal{R}(x^{\text{obs}})\right) = 1 - \alpha, \quad \forall \theta \in \Theta \tag{1.1}$$

and the expected size of this set $\mathbb{E}[|\mathcal{R}(x^{\text{obs}})|]$ is small in some suitable sense. In what follows, we will assume that the (intractable) likelihood model $\mathcal{L}(\theta; X)$, which defines the data-generating process, is well-specified and does not change between the training and inference stages[1]. On the other hand, we explicitly allow for the prior to be mis-specified.

We provide algorithms, modular frameworks and theoretical guarantees that aim at equipping recent advancements in the AI literature with sound statistical properties. By bridging modern machine learning with classical statistical inference tools, we effectively provide a principled path towards integrating AI into scientific discovery, enabling advancements in fields such as astronomy, high-energy physics, biology, and beyond.

1.1 Summary of Contributions

Chapter 2. Background: Likelihood-Free Frequentist Inference We begin by introducing the LF2I framework, which proposes an amortized procedure to implement the Neyman construction of confidence sets via likelihood-based test statistics and critical values based on quantile regression. In addition, we also discuss a independent diagnostics procedure which allows to check the empirical coverage of any parameter region across the entire parameter space.

Chapter 2 is based on Dalmasso, Masserano*, Zhao, Izbicki, and Lee (2024), which appeared on the Electronic Journal of Statistics, Vol. 18, No. 2.*

Chapter 3. Confidence Sets from Prediction Algorithms and Posterior Estimators Starting from the framework introduced in Chapter 2, we enhance it to leverage state-of-the-art prediction algorithms and posterior estimators via a surrogate of the Wald test statistic. By doing so, we are able to tackle complex scientific questions such as inferring the energy of a subatomic particle using convolutional neural networks on 3D data.

Chapter 3 is based on Masserano, Dorigo, Izbicki, Kuusela, and Lee (2023), which appeared at the 26th International Conference on Artificial Intelligence and Statistics (AISTATS).

Chapter 4. Optimal Confidence Sets from Generative Models In Chapter 3 we showed empirically that our confidence sets from posterior estimators exhibit validity across the parameter space without being conservative when the prior distribution is aligned with the target distribution over θ. Here, we first introduce an alternative method to do Neyman inversion by estimating p-values that are amortized across parameters, data and confidence levels altogether. In this way, practitioners can construct confidence sets that are simultaneously valid at all levels α without having to retrain a calibration model for each. Second, we prove that our confidence sets — if constructed by using the estimated posterior

[1]We will relax this assumption in Chapter 5.

distribution as a test statistic for Neyman inversion — are in fact optimal, i.e. they achieve the smallest possible size on average with respect to the marginal distribution on the data induced by the prior distribution. We demonstrate the potential of these Frequentist-Bayes sets on three challenging case studies of practical relevance.

Chapter 4 is based on Masserano, Carzon*, Shen*, Herling Ribeiro*, Dorigo, Doro, Speagle, Izbicki, and Lee (2025), which is currently in submission to a major scientific journal.*

Chapter 5. Inference under Nuisance Parameters and Generalized Label Shift We then relax the assumption of a well-specified likelihood model by considering a more general setup that reflects a richer mechanistic model: $\theta = (Y, \nu) \to X$, where $\nu \in \mathcal{N}$ are continuous or discrete nuisance parameters that are not of direct interest but critically influence the data-generating process. These nuisance parameters are available at the training stage, but are *not* observed at the inference stage when estimating Y from x_{obs}. We refer to a shift that simultaneously affects Y and ν as *generalized label shift* (GLS), and assume that $p_{\text{train}}(X \mid Y, \nu) = p_{\text{obs}}(X \mid Y, \nu)$. Within this setting, we propose a new method for robust uncertainty quantification that casts classification as a hypothesis testing problem under nuisance parameters. The key idea is to estimate the classifier's receiver operating characteristic (ROC) across the entire nuisance parameter space, which allows us to devise cutoffs that are invariant under GLS. Our method endows a pretrained classifier with domain adaptation capabilities and returns valid prediction sets while retaining high power.

Chapter 5 is based on Masserano, Shen*, Doro, Dorigo, Izbicki, and Lee (2024), which appeared at the 41st International Conference on Machine Learning (ICML).*

Chapter 6. The `lf2i` Package A central goal of this thesis is to provide methods that are not only methodologically or theoretically appealing, but that are also easy to use in practice, so that domain scientists can benefit from them during their investigations. As such, we devoted a crucial effort into developing and maintaining a friendly Python package that provides scalable implementations of all the methods presented in this thesis. In this Chapter, we briefly review the main structure and contributions of this package.

Chapter 6 is based on Masserano (2023), which is available as an open source Python package on PyPI and GitHub at https://github.com/lee-group-cmu/lf2i.

Chapter 7. Extensions and Future Work We conclude by discussing a few methodological extensions and novel applications that we have been working on and that will set the ground for future explorations.

Chapter 7 is partially based on Carzon, Masserano, Ghosh, Whiteson, Izbicki, and Lee (2025), which is currently in submission to a major physics journal.

2

Background: Likelihood-Free Frequentist Inference

2.1 Introduction

Hypothesis testing and uncertainty quantification are the hallmarks of scientific inference. Methods that achieve good statistical performance (e.g., high power) often rely on being able to explicitly evaluate a likelihood function, which relates parameters of the data-generating process to observed data. However, in many areas of science and engineering, complex phenomena are modeled by forward simulators that *implicitly* define a likelihood function. For example,[1] given input parameters θ from some parameter space Θ, a stochastic model F_θ may encode the interaction of atoms or elementary particles, or the transport of radiation through the atmosphere or through matter in the Universe by combining deterministic dynamics with random fluctuations and measurement errors, to produce synthetic data X.

Simulation-based inference with an intractable likelihood is commonly referred to as *likelihood-free inference* (LFI). The most well-known approach to LFI is Approximate Bayesian Computation (ABC; see Beaumont (2010); Marin et al. (2012); Sisson et al. (2018); Sunnåker et al. (2013) for reviews). These methods use simulations sufficiently close to the observed data $D = \left\{x_1^{\text{obs}}, \ldots, x_n^{\text{obs}}\right\}$ to infer the underlying parameters, or more precisely, the posterior distribution $p(\theta \mid D)$. Recently, the arsenal of LFI methods has been expanded with new machine learning algorithms that instead use the output from simulators as training data. The objective here is to learn a "surrogate model" or *approximation* of the likelihood $p(D \mid \theta)$ or posterior $p(\theta \mid D)$. The surrogate model, rather than the simulations themselves, is then used for inference. Machine-learning (ML) based methods have revolutionized LFI in terms of the complexity and dimensionality of the problems that can be tackled (see Cranmer et al. (2020) for a recent review). Nevertheless, neither ABC nor ML-based LFI approaches guarantee confidence sets with frequentist coverage, which are crucial to ensure reliability of downstream scientific conclusions. Suppose that we have a high-fidelity simulator F_θ, which

[1] **Notation.** Let F_θ represent the stochastic forward model for a sample point $X \in \mathcal{X} \subseteq \mathbb{R}^p$ at parameter $\theta \in \Theta \subseteq \mathbb{R}^d$. We refer to F_θ as a "simulator", as the assumption is that we can sample data from the model. We denote i.i.d."observable" data from F_θ by $\mathcal{D} = \{X_1, \ldots, X_n\}$, and the actually observed or measured data by $D = \left\{x_1^{\text{obs}}, \ldots, x_n^{\text{obs}}\right\}$. The likelihood function is defined as $\mathcal{L}(D; \theta) = \prod_{i=1}^n p(x_i^{\text{obs}} \mid \theta)$, where $p(\cdot \mid \theta)$ is the density of F_θ with respect to a fixed dominating measure ν, which could be the Lebesgue measure.

implicitly encodes the likelihood, and that we observe data $\mathcal{D}$ of finite sample size n. We address two open challenges in LFI:

i) The first challenge is finding practical procedures for constructing a $(1-\alpha)$ confidence set $R(\mathcal{D})$ with nominal coverage[2]

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \mathcal{R}(\mathcal{D})\right) = 1-\alpha, \tag{2.1}$$

where $\alpha \in (0,1)$, *regardless of* the true value of the unknown parameter $\theta \in \Theta$ and of the number of observations n. Monte Carlo and bootstrap procedures are computationally infeasible for continuous parameter spaces Θ, and large-sample theory does not apply when, e.g., $n = 1$. The latter $n = 1$ scenario is very common in, e.g., large astronomical surveys where each object (e.g., galaxy or star) has a different parameter value θ and may only be measured once.

ii) The second challenge is finding practical and interpretable procedures to check that the empirical coverage of the constructed sets $\mathcal{R}(\mathcal{D})$ is indeed close to (and no smaller than) $1-\alpha$ for *any* $\theta \in \Theta$ (again, without resorting to costly Monte Carlo simulations at fixed parameter settings on a fine grid in parameter space Θ (Cousins, 2018, Section 13)). Local validity across the entire parameter space is essential for reliable scientific inference because the scientist does not actually know what the true value of θ is for the object of interest.

Novelty and significance. In this chapter, we introduce a fully modular statistical framework that addresses both problems above. We refer to the general approach as *likelihood-free frequentist inference* (LF2I)[3]. LF2I is fully nonparametric and targets modern scientific applications, involving, e.g, high-dimensional data of different modalities, intractable likelihood models, and/or small sample sizes. Section 2.7.1 describes how LF2I is related to other work in this area.

At the heart of LF2I is the *Neyman construction of confidence sets*, albeit applied to a setting where the test statistic's distribution is unknown. Frequentist confidence sets and their equivalence to hypothesis tests have a long history in statistics (Fisher, 1925; Neyman, 1935a, 1937a). While classical statistical procedures have significantly impacted fields like high-energy physics (see Section 2.7.1), most simulator-based methods lack theoretical guarantees for confidence sets beyond low-dimensional data and large-sample assumptions (Feldman and Cousins, 1998). Implementing the Neyman construction for LFI is challenging not only because one cannot evaluate the likelihood, but also because one needs to test null hypotheses across the entire parameter space. While Monte Carlo and bootstrap methods estimate critical values and p-values from a batch of simulations at each null value θ_0 (MacKinnon, 2009; Ventura, 2010), they become computationally infeasible for high-dimensional parameters. As a result, practical implementations might rely on parametric assumptions or asymptotic theory (Neyman and Pearson, 1928; Wilks, 1938). For instance, it is often assumed that the likelihood-ratio (LR) statistic follows a χ^2 distribution, but this does not hold for irregular models or small sample sizes Algeri et al. (2019); Kieseler et al. (2022); Ho et al. (2021). This work seeks to quickly and accurately estimate critical values and coverage

[2]We use the notation $\mathbb{P}_{\mathcal{D}|\theta}(\cdot)$ to emphasize the fact that $\mathcal{D}$ is random, but θ is fixed.

[3]Code is available as a Python package at https://github.com/lee-group-cmu/lf2i.

across the parameter space without knowing the test statistic distribution or relying on large-sample approximations.

The key insight behind LF2I is that the main quantities of interest in frequentist statistical inference — test statistics, critical values, p-values and coverage of the confidence set — are *distribution functions indexed by the (unknown) parameter* θ, which generally vary smoothly over the parameter space Θ. As a result, one can leverage machine learning methods and data simulated in the neighborhood of a parameter to improve estimates of quantities of interest with fewer total simulations. Figure 2.1 illustrates the general LF2I inference machinery, which is composed of three modular branches with separate functionalities:

i) The test statistic branch (Figure 2.1 center and Section 2.3.2) uses a simulated set $\mathcal{T}$ to estimate a test statistic $\lambda(\mathcal{D};\theta_0)$ for testing $H_{0,\theta_0} : \theta = \theta_0$ versus $H_{1,\theta_0} : \theta \neq \theta_0$. We study the theoretical and empirical performance of LF2I confidence sets derived from likelihood-based test statistics learned via the odds function $\mathbb{O}(X;\theta)$ of Equation (2.7).

ii) The calibration branch (Figure 2.1 left and Section 2.3.3) uses a left-out set $\mathcal{T}'$ to estimate critical values C_{θ_0} for every level-α test of H_{0,θ_0} via quantile regression of the estimated test statistic $\lambda(\mathcal{D};\theta_0)$ on $\theta_0 \in \Theta$. Once we have estimated the quantile function $\widehat{C}_{\theta_0}$ indexed by θ_0, we can directly construct Neyman confidence sets

$$\widehat{\mathcal{R}}_\alpha(\mathcal{D}) := \left\{\theta \in \Theta : \lambda(\mathcal{D};\theta) \geqslant \widehat{C}_{\theta,\alpha}\right\} \tag{2.2}$$

that have approximate $(1-\alpha)$ finite-n coverage for every value of $\theta \in \Theta$. LF2I with critical values is amortized, meaning that once trained it can be evaluated on an arbitrary number of observations D. Alternatively, we can estimate p-values $p(D;\theta_0)$ for every test at $\theta = \theta_0$ with observed data D.

iii) The diagnostics branch (Figure 2.1 right and Section 2.3.4) uses a validation set $\mathcal{T}''$ to assess the empirical coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ of the constructed confidence sets $\widehat{\mathcal{R}}(\mathcal{D})$ across the parameter space by regressing the indicator variable $W := \mathbb{1}(\lambda(\mathcal{D};\theta) \geqslant \widehat{C}_\theta)$ on θ. The diagnostics branch is not part of the inference procedure itself. Its purpose is to provide an independent assessment of local (instance-wise) coverage of the final constructed confidence sets.

The LF2I approach was first introduced in a conference proceeding Dalmasso et al. (2020). This preliminary version — `ACORE` (Approximate Computation via Odds Ratio Estimation) — uses a test statistic that maximizes odds over the parameter space. In this chapter, we analyze the statistical and computational properties of LF2I, while also introducing a new test statistic — the Bayesian Frequentist Factor (`BFF`) — which is the Bayes Factor (Jeffreys, 1935, 1961) treated as a frequentist test statistic. We show that the validity of LF2I only depends on calibration, whereas its power depends on the test statistic's definition and its estimation quality. In addition to new theoretical results in Section 2.4, we compare LF2I with approaches using Monte Carlo methods or Wilks' theorem (Section 2.6.1), and we illustrate how our diagnostics can help scientists in choosing the best tool to handle nuisance parameters (Section 2.6.2). Finally, we construct confidence sets given a high-dimensional particle physics simulation where ABC approaches are neither computationally feasible nor valid (Section 2.6.3).

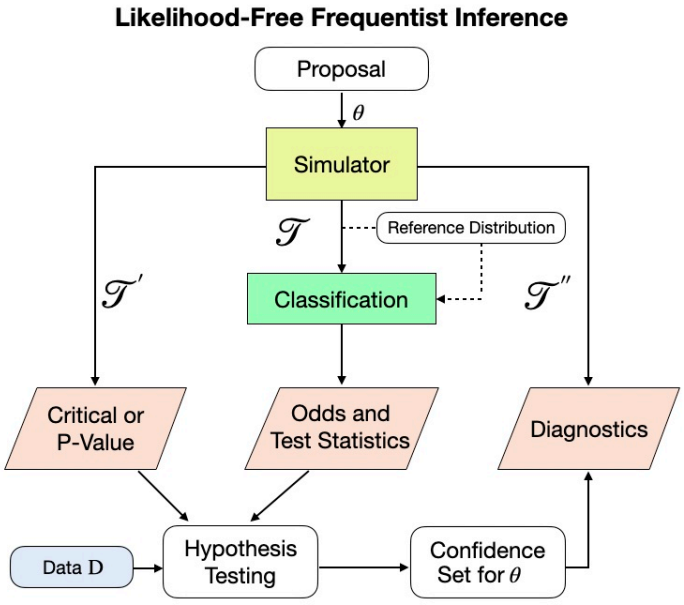


Figure 2.1: **The three-branch fully modular framework for likelihood-free frequentist inference (LF2I). Center branch:** Draw a sample $\mathcal{T}$ of size B from the simulator to estimate an arbitrary test statistic $\lambda(\mathcal{D};\theta)$. Here we show how to do so by estimating the likelihood via the odds function $\mathbb{O}(X;\theta)$. **Left branch:** Draw a second sample $\mathcal{T}'$ of size B' to estimate the critical values C_θ or p-values $p(\mathcal{D};\theta)$ for all $\theta \in \Theta$. **Left + Center:** Once data D are observed, we can construct confidence sets $\widehat{\mathcal{R}}(D)$ with finite-n validity according to Equation (2.12). **Right branch:** The LF2I diagnostics branch independently checks whether the coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ of the confidence set is indeed correct across the entire parameter space.

2.2 Statistical Inference in a Traditional Setting

We now review the Neyman construction of confidence sets and the definitions of likelihood ratio and Bayes factor, before moving on to the details of the LF2I framework and its two instances, `ACORE` and `BFF`.

Equivalence of tests and confidence sets. A classical approach to constructing a confidence set for an unknown parameter $\theta \in \Theta$ is to invert a series of hypothesis tests (Neyman, 1937a). Suppose that for each possible value $\theta_0 \in \Theta$, there exists a level-α test δ_{θ_0} of

$$H_{0,\theta_0} : \theta = \theta_0 \quad \text{versus} \quad H_{1,\theta_0} : \theta \neq \theta_0. \tag{2.3}$$

That is, a test δ_{θ_0} where the type-I error (the probability of erroneously rejecting a true null hypothesis H_{0,θ_0}) is no larger than α. For observed data $\mathcal{D} = D$, let $\mathcal{R}(D)$ be the set of all parameter values $\theta_0 \in \Theta$ for which the test δ_{θ_0} does not reject H_{0,θ_0}. Then, by construction, the random set $\mathcal{R}(\mathcal{D})$ satisfies

$$\mathbb{P}_{\mathcal{D}|\theta}\left(\theta \in \mathcal{R}(\mathcal{D})\right) \geqslant 1 - \alpha \quad \forall \theta \in \Theta,$$

which makes it a $(1-\alpha)$ *confidence set* for θ. Similarly, we can define tests with a desired significance level by inverting a confidence set with a certain coverage.

Likelihood ratio test. A general form of hypothesis tests that often leads to high power is the likelihood ratio test (LRT). Consider testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1, \tag{2.4}$$

where $\Theta_1 = \Theta \backslash \Theta_0$. For the *likelihood ratio (LR) statistic*,

$$\text{LR}(\mathcal{D}; \Theta_0) = \log \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\mathcal{D}; \theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\mathcal{D}; \theta)}, \tag{2.5}$$

the LRT of the hypotheses in Equation (2.4) rejects H_0 when $\text{LR}(D; \Theta_0) < C$ for some constant C. Figure 2.2 illustrates the construction of confidence sets for θ from the level-α likelihood ratio tests of Equation (2.3). The critical value for each such test δ_{θ_0} is $C_{\theta_0} = \sup \left\{ C : \mathbb{P}_{\mathcal{D}|\theta_0} \left(\text{LR}(\mathcal{D}; \theta_0) < C \right) \leqslant \alpha \right\}$.

Bayes factor. Let π be a probability measure over the parameter space Θ. The Bayes factor (Jeffreys, 1935, 1961) for comparing the hypothesis $H_0 : \theta \in \Theta_0$ to its complement, the alternative H_1, is the ratio of the marginal likelihood of the two hypotheses:

$$\text{BF}(\mathcal{D}; \Theta_0) \equiv \frac{\mathbb{P}(\mathcal{D} \mid H_0)}{\mathbb{P}(\mathcal{D} \mid H_1)} = \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D}; \theta) \mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D}; \theta) \mathrm{d}\pi_1(\theta)}, \tag{2.6}$$

where π_0 and π_1 are the restrictions of π to the parameter regions Θ_0 and $\Theta_1 = \Theta_0^c$, respectively. The Bayes factor is often used as a Bayesian alternative to significance testing, as it quantifies the change in the odds in favor of H_0 when going from the prior to the posterior: $\frac{\mathbb{P}(H_0|\mathcal{D})}{\mathbb{P}(H_1|\mathcal{D})} = \text{BF}(\mathcal{D}; \Theta_0) \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$.

2.3 Likelihood-Free Frequentist Inference via Odds Estimation

In the typical LFI setting, we cannot directly evaluate the likelihood ratio $\text{LR}(\mathcal{D}; \Theta_0)$ or even the likelihood $\mathcal{L}(\mathcal{D}; \theta)$. In this work, we describe a version of LF2I that is based on odds estimation. We assume that we have access to *(i)* a forward simulator F_θ to draw observable data, *ii* a reference distribution G that does not depend on θ, with larger support than F_θ for all $\theta \in \Theta$, and *(iii)* a probabilistic classifier to discriminate samples from F_θ and G.

2.3.1 Estimating an Odds Function across the Parameter Space

We start by generating a labeled sample $\mathcal{T} = \{(\theta_i, X_i, Y_i)\}_{i=1}^B$ to compare data from F_θ with data from the reference distribution G. Here, $\theta \sim \pi_\Theta$ (a proposal distribution over Θ), the

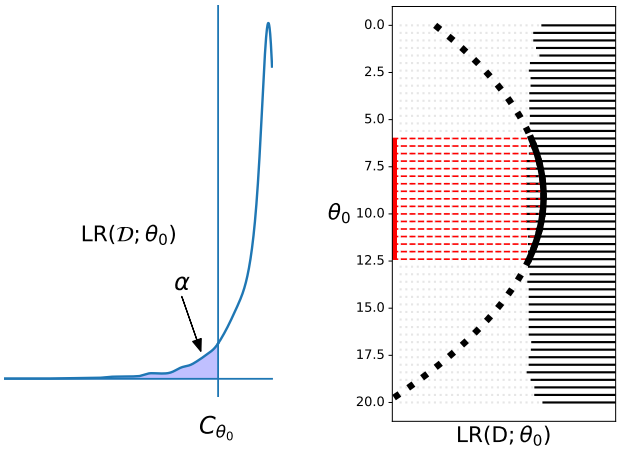


Figure 2.2: **Neyman construction of confidence sets by inverting hypothesis tests. Left:** For each $\theta_0 \in \Theta$, we find the critical value C_{θ_0} that rejects the null hypothesis H_{0,θ_0} at level α; that is, C_{θ_0} is the α-quantile of the distribution of the test statistic under the null (a likelihood ratio $\mathsf{LR}(\mathcal{D}; \theta_0)$ in this case). **Right:** The horizontal solid lines represent acceptance regions for each $\theta_0 \in \Theta$. Suppose we observe data D. The confidence set for θ (red vertical solid line) consists of all θ_0-values for which the observed test statistic $\mathsf{LR}(D; \theta_0)$ (black curve) falls in the acceptance region.

"label" $Y \sim \text{Bernoulli}(p)$, $X \mid (\theta, Y = 1) \sim F_\theta$ and $X \mid (\theta, Y = 0) \sim G$. We then define the odds at θ and fixed x as

$$\mathbb{O}(x; \theta) := \frac{\mathbb{P}(Y = 1 \mid \theta, x)}{\mathbb{P}(Y = 0 \mid \theta, x)}. \tag{2.7}$$

One way of interpreting $\mathbb{O}(x; \theta)$ is to regard it as a measure of the chance that x was generated from F_θ rather than from G. That is, a large odds $\mathbb{O}(x; \theta)$ reflects the fact that it is plausible that x was generated from F_θ (instead of G). We call G a "reference distribution" as we are comparing F_θ for different θ with this distribution. Equation (2.7) is equivalent to the likelihood $p(x \mid \theta)$ up to a normalization constant, as shown in Dalmasso et al. (2020, Proposition 3.1). The odds function $\mathbb{O}(X; \theta)$ with $\theta \in \Theta$ as a parameter can be estimated with a probabilistic classifier, such as a neural network with a softmax layer, suitable for the data at hand. Algorithm A.1 in Appendix A.1 summarizes our procedure for simulating a labeled sample $\mathcal{T}$. For all experiments in this chapter, we use $p = 1/2$ and $G = F_X$, where F_X is the (empirical) marginal distribution of F_θ with respect to π_Θ.

2.3.2 Test Statistics based on Odds

For testing $H_{0,\Theta_0} : \theta \in \Theta_0$ versus all alternatives $H_{1,\Theta_0} : \theta \notin \Theta_0$, we consider two test statistics: `ACORE` and `BFF`. Both statistics are based on $\mathbb{O}(X; \theta)$, but whereas `ACORE` eliminates the parameter θ by maximization, `BFF` averages over the parameter space.

`ACORE` by Maximization

The `ACORE` statistic (Dalmasso et al., 2020) for testing Equation (2.3) is given by

$$\begin{aligned}\Lambda(\mathcal{D};\Theta_0) &:= \log \frac{\sup_{\theta\in\Theta_0}\prod_{i=1}^{n}\mathbb{O}(X_i;\theta)}{\sup_{\theta\in\Theta}\prod_{i=1}^{n}\mathbb{O}(X_i;\theta)} \\ &= \sup_{\theta_0\in\Theta_0}\inf_{\theta_1\in\Theta}\sum_{i=1}^{n}\log\left(\mathbb{OR}(X_i;\theta_0,\theta_1)\right), \end{aligned} \tag{2.8}$$

where the odds ratio

$$\mathbb{OR}(x;\theta_0,\theta_1) := \frac{\mathbb{O}(x;\theta_0)}{\mathbb{O}(x;\theta_1)} \tag{2.9}$$

at $\theta_0,\theta_1\in\Theta$ measures the plausibility that a fixed x was generated from θ_0 rather than θ_1. We use $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ to denote the `ACORE` statistic based on $\mathcal{T}$ and estimated odds $\widehat{\mathbb{O}}(X;\theta_0)$. When $\widehat{\mathbb{O}}(X;\theta_0)$ is well-estimated for every θ and X, $\widehat{\Lambda}(\mathcal{D};\Theta_0)$ is the same as the $\mathrm{LR}(\mathcal{D};\Theta_0)$ in Equation (2.5) (Dalmasso et al., 2020, Proposition 3.1).

`BFF` by Averaging

Because the `ACORE` statistics in Equation (2.8) involves taking the supremum (or infimum) over Θ, it may not be practical in high dimensions. Hence, in this work, we propose an alternative statistic for testing (2.3) based on averaged odds:

$$\tau(\mathcal{D};\Theta_0) := \frac{\int_{\Theta_0}\prod_{i=1}^{n}\mathbb{O}(X_i;\theta_0)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_0^c}\prod_{i=1}^{n}\mathbb{O}(X_i;\theta)\mathrm{d}\pi_1(\theta)}, \tag{2.10}$$

where π_0 and π_1 are the restrictions of the proposal distribution π to the parameter regions Θ_0 and Θ_0^c, respectively. Let $\widehat{\tau}(\mathcal{D};\Theta_0)$ denote estimates based on $\mathcal{T}$ and $\widehat{\mathbb{O}}(\theta_0;x)$. If the probabilities learned by the classifier are well estimated, then the estimated averaged odds statistic $\widehat{\tau}(\mathcal{D};\Theta_0)$ is exactly the Bayes factor:

Proposition 2.1 (Fisher consistency)**.**

Assume that, for every $\theta\in\Theta$, G dominates ν. If $\widehat{\mathbb{P}}(Y=1\mid\theta,x)=\mathbb{P}(Y=1\mid\theta,x)$ for every θ and x, then $\widehat{\tau}(\mathcal{D};\Theta_0)$ is the Bayes factor $BF(\mathcal{D};\Theta_0)$.

In this chapter, we are using the Bayes factor as a frequentist test statistic. Hence, our term Bayes Frequentist Factor (`BFF`) statistic for τ and $\widehat{\tau}$.

2.3.3 Fast Construction of Neyman Confidence Sets

Instead of a costly MC or bootstrap hypothesis test of $H_0:\theta=\theta_0$ at each θ_0 on a fine grid (see, e.g., MacKinnon (2009) and Ventura (2010)), we draw only one sample $\mathcal{T}'$ of size B'. We then estimate either the critical value C_{θ_0} via quantile regression (Section 2.3.3), or the p-value $p(D;\theta_0)$ via probabilistic classification (Section 2.3.3), for all $\theta_0\in\Theta$ simultaneously. In Supplementary Material H[4], we outline a practical strategy to choose the number of simulations B' and the learning algorithm.

[4]Available at https://lucamasserano.github.io/data/LF2I_supplementary_material.pdf.

Algorithm 2.1 Estimate critical values C_{θ_0} for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$ for all $\theta_0 \in \Theta$ simultaneously

Input: simulator F_θ; number of simulations B'; π_Θ (fixed proposal distribution over the parameter space); test statistic λ; quantile regression estimator; level $\alpha \in (0, 1)$
Output: estimated critical values $\widehat{C}_{\theta_0}$ for all $\theta_0 \in \Theta$

1: Set $\mathcal{T}' \leftarrow \varnothing$
2: **for** i in $\{1, \ldots, B'\}$ **do**
3: Draw parameter $\theta_i \sim \pi_\Theta$
4: Draw sample $X_{i,1}, \ldots, X_{i,n} \overset{\text{iid}}{\sim} F_{\theta_i}$
5: Compute test statistic $\lambda_i \leftarrow \lambda((X_{i,1}, \ldots, X_{i,n}); \theta_i)$
6: $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, \lambda_i)\}$
7: Use $\mathcal{T}'$ to learn the conditional quantile function $\widehat{C}_\theta := \widehat{F}^{-1}_{\lambda|\theta}(\alpha \mid \theta)$ via quantile regression of λ on θ
8: **return** $\widehat{C}_{\theta_0}$

The Critical Value via Quantile Regression

Algorithm 2.1 describes how to use quantile regression (e.g., Meinshausen (2006); Koenker et al. (2017)) to estimate the critical value C_{θ_0} for the level-α test of Equation (2.3) as a function of $\theta_0 \in \Theta$. To test a composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, we use the cutoff $\widehat{C}_{\Theta_0} := \inf_{\theta \in \Theta_0} \widehat{C}_\theta$. Although the calibration procedure was originally proposed for `ACORE`, the same scheme leads to a valid test (control of type-I error as the number of simulations $B' \to \infty$) for *any* test statistic λ (Theorem A.4). Remarkably, this holds even if the test statistic is not well estimated. Note that in practice, we observe that the number of simulations B' needed to achieve correct coverage is usually much lower relative to B, the number of simulations needed to estimate the test statistic. In addition, Algorithm 2.1 does not rely on the observed data D and is therefore amortized, meaning that once the test statistic and critical values have been estimated, we can compute confidence sets for any new datapoint without the need to retrain the model.

The P-Value via Probabilistic Classification

If the data D are observed beforehand, then given any test statistic λ we can alternatively compute p-values for each hypothesis $H_{0,\theta_0} : \theta = \theta_0$, that is,

$$p(D; \theta_0) := \mathbb{P}_{\mathcal{D}|\theta_0} \left(\lambda(\mathcal{D}; \theta_0) < \lambda(D; \theta_0) \right). \tag{2.11}$$

The p-value $p(D; \theta_0)$ can be used to test hypothesis and create confidence sets for any desired level α. As detailed in Algorithm A.3, we can estimate it simultaneously for all $\theta \in \Theta$ by drawing a training sample $\mathcal{T}' = \{(Z_1, \theta_1), \ldots, (Z_{B'}, \theta_{B'})\}$ and using the random variable $Z := \mathbb{1}\left(\lambda(\mathcal{D}; \theta) < \lambda(D; \theta)\right)$ as a label for each θ. To test the composite null hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, we use

$$\hat{p}(D; \Theta_0) := \sup_{\theta \in \Theta_0} \hat{p}(D; \theta).$$

Note that there is a key computational difference between estimating p-values versus estimating critical values. The p-value is a function of both θ and the observed sample

Algorithm 2.2 Estimate empirical coverage $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$, for all $\theta \in \Theta$.

Input: simulator F_θ; number of simulations B''; π_Θ (fixed proposal distribution over parameter space); test statistic λ; level α; critical values $\widehat{C}_\theta$; probabilistic classifier
Output: estimated coverage $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ for all $\theta \in \Theta$

Set $\mathcal{T}'' \leftarrow \varnothing$
for i in $\{1, \ldots, B''\}$ **do**
    Draw parameter $\theta_i \sim \pi_\Theta$
    Draw sample $\mathcal{D}_i := \{X_{i,1}, \ldots, X_{i,n}\} \overset{\text{iid}}{\sim} F_{\theta_i}$
    Compute test statistic $\lambda_i \leftarrow \lambda(\mathcal{D}_i; \theta_i)$
    Compute indicator variable $W_i \leftarrow \mathbb{1}\left(\lambda_i \geqslant \widehat{C}_{\theta_i}\right)$
    $\mathcal{T}'' \leftarrow \mathcal{T}'' \cup \{(\theta_i, W_i)\}$
Use $\mathcal{T}''$ to learn $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ across Θ by regressing W on θ
return $\widehat{\mathbb{P}}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$

D itself. As a result, Algorithm A.3 has to be repeated for each observed D, making the computation of p-values non-amortized. In Chapter 4, we will see how to generalize this method to allow for seamless amortization with respect to both D and α.

Amortized Confidence Sets

Finally, we construct an approximate confidence region for θ by taking the set

$$\widehat{\mathcal{R}}(D) = \left\{\theta \in \Theta : \lambda(D; \theta) \geqslant \hat{C}_\theta\right\}, \tag{2.12}$$

or, alternatively,

$$\widehat{\mathcal{R}}(D) = \{\theta \in \Theta : \hat{p}(D; \theta) > \alpha\}. \tag{2.13}$$

See Algorithm A.4 in Appendix A.3 for details. As shown in Dalmasso et al. (2020, Theorem 3.3), the random set $\widehat{\mathcal{R}}(\mathcal{D})$ has nominal $(1 - \alpha)$ coverage as $B' \to \infty$ regardless of the observed sample size n. As noted in Section 2.3.3, the confidence set in Equation (2.12) is fully *amortized*, meaning that once we have $\lambda(\mathcal{D}; \theta)$ and $\widehat{C}_\theta$ as a function of $\theta \in \Theta$, we can perform inference on new data without retraining.

2.3.4 Diagnostics: Checking Coverage across the Parameter Space

The LF2I framework has a separate module ("Diagnostics" in Figure 2.1) for evaluating "local" goodness-of-fit in different regions of the parameter space Θ. This estimates the coverage probability $\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \widehat{\mathcal{R}}(\mathcal{D}))$ of confidence sets $\widehat{\mathcal{R}}(\mathcal{D})$ across the parameter space via probabilistic classification. As detailed in Algorithm 2.2, we first generate a set of size B'' from the simulator: $\mathcal{T}'' = \{(\theta_1, \mathcal{D}_1), \ldots, (\theta_{B''}, \mathcal{D}_{B''})\}$. Then, for each sample $\mathcal{D}_i$, we check whether or not the test statistic λ_i is larger than the estimated critical value $\widehat{C}_{\theta_i}$ (the output from Algorithm 2.1). This is equivalent to computing a binary variable W_i for whether or not the "true" value θ_i falls within the confidence set $\widehat{\mathcal{R}}(\mathcal{D}_i)$ of Equation (2.12). Recall that the computations of the test statistic and the critical value are amortized, meaning that we do not retrain algorithms to estimate these two quantities. The final step is to

estimate empirical coverage as a function of θ by using W as a label for each θ. This estimation requires a new fit, but after training the probabilistic classifier, we can evaluate the estimated coverage anywhere in parameter space Θ.

This diagnostic procedure locates regions in parameter space where estimated confidence sets might under- or over-cover; see Figures 2.3, 2.4 and 2.6 for examples. Note that standard goodness-of-fit techniques for conditional densities (Cook et al., 2006b; Bordoloi et al., 2010; Talts et al., 2018; Schmidt et al., 2020) only check for marginal coverage over Θ.

2.4 Theoretical Guarantees

We now prove consistency of the critical value and p-value estimation methods (Algorithms 2.1 and A.3, respectively) and provide theoretical guarantees for the power of `BFF`. We refer the reader to Appendix A.4 for a proof for finite Θ that the power of `ACORE` converges to the power of LRT as B grows (Theorem A.1).

In this section, $\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}$ denotes the probability integrated over both $\mathcal{D} \sim F_\theta$ and $\mathcal{T}'$, whereas $\mathbb{P}_{\mathcal{D}|\theta}$ denotes integration over $\mathcal{D} \sim F_\theta$ only. For notational ease, we do not explicitly state again (inside the parentheses of the same expression) that we condition on θ.

2.4.1 Critical Value Estimation

We start by showing that our procedure for choosing critical values leads to valid hypothesis tests (that is, tests that control the type-I error probability), as long as the number of simulations B' in Algorithm 2.1 is sufficiently large. We assume that the null hypothesis is simple, that is, $\Theta_0 = \{\theta_0\}$ — which is the relevant setting for the Neyman construction of confidence sets in the absence of nuisance parameters. See Theorem A.4 in Appendix A.6 for results for composite null hypotheses.

We assume that the quantile regression estimator described in Section 2.3.3 is consistent in the following sense:

Assumption 2.2 (Uniform consistency)**.** *Let $F(\cdot \mid \theta)$ be the cumulative distribution function of the test statistic $\lambda(\mathcal{D};\theta_0)$ conditional on θ, where $\mathcal{D} \sim F_\theta$. Let $\widehat{F}_{B'}(\cdot \mid \theta)$ be the estimated distribution function indexed by θ, implied by a quantile regression with a sample $\mathcal{T}'$ of B' simulations $\mathcal{D} \sim F_\theta$. Assume that the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\hat{F}_{B'}(\lambda \mid \theta_0) - F(\lambda \mid \theta_0)| \xrightarrow[B' \longrightarrow \infty]{P} 0.$$

Assumption 2.2 holds, for instance, for quantile regression forests (Meinshausen, 2006). Next, we show that Algorithm 2.1 yields a valid hypothesis test as $B' \to \infty$.

Theorem 2.3. *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on an absolutely continuous statistic $\lambda(\mathcal{D};\theta_0)$ chosen according to Algorithm 2.1 for a fixed $\alpha \in (0,1)$. If the quantile estimator satisfies Assumption 2.2, then for every $\theta_0, \theta \in \Theta$*

$$\mathbb{P}_{\mathcal{D}|\theta_0, C_{B'}}(\lambda(\mathcal{D};\theta_0) \leqslant C_{B'}) \xrightarrow[B' \longrightarrow \infty]{a.s.} \alpha,$$

where $\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}$ *denotes the probability integrated over* $\mathcal{D} \sim F_{\theta_0}$ *and conditional on the random variable* $C_{B'}$.

If the convergence rate of the quantile regression estimator is known (Assumption 2.4), Theorem 2.5 provides a finite-B' guarantee on how far the type-I error of the test will be from the nominal level.

Assumption 2.4 (Convergence rate of the quantile regression estimator)**.** *Using the notation of Assumption 2.2, assume that the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}} |\hat{F}_{B'}(\lambda \mid \theta_0) - F(\lambda \mid \theta_0)| = \mathcal{O}_P\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

Theorem 2.5. *With the notation and assumptions of Theorem 2.3, and if Assumption 2.4 also holds, then*

$$|\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leqslant C_{B'}) - \alpha| = \mathcal{O}_P\left(\left(\frac{1}{B'}\right)^r\right).$$

2.4.2 P-Value Estimation

Next we show that the p-value estimation method described in Section 2.3.3 is consistent. The results shown here apply to any test statistic λ. That is, these results are not restricted to `ACORE` nor `BFF`. We assume consistency in the sup norm of the regression method used to estimate the p-values:

Assumption 2.6 (Uniform consistency)**.** *The regression estimator used in Equation* (2.11) *is such that*

$$\sup_{\theta\in\Theta_0} |\widehat{\mathbb{E}}_{B'}[Z \mid \theta] - \mathbb{E}[Z \mid \theta]| \xrightarrow[B'\longrightarrow\infty]{a.s.} 0.$$

Examples of estimators that satisfy Assumption 2.6 include Bierens (1983); Hardle et al. (1984); Liero (1989); Girard et al. (2014).
The next theorem shows that the p-values obtained according to Algorithm A.3 converge to the true p-values. Moreover, the power of the tests obtained using the estimated p-values converges to the power one would obtain if the true p-values could be computed.

Theorem 2.7. *Under Assumption 2.6 and if* $p(\mathcal{D};\Theta_0)$ *is an absolutely continuous random variable then,for every* $\theta \in \Theta$

$$\widehat{p}(D;\Theta_0) \xrightarrow[B'\longrightarrow\infty]{a.s.} p(D;\Theta_0)$$

and

$$\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\hat{p}(\mathcal{D};\Theta_0) \leqslant \alpha) \xrightarrow[B'\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\Theta_0) \leqslant \alpha).$$

The next corollary shows that as $B' \longrightarrow \infty$, the tests obtained using the p-values from Algorithm A.3 have size α.

Corollary 2.8. *Under Assumption 2.6 and if F_θ is continuous for every $\theta \in \Theta$ and $p(\mathcal{D}; \Theta_0)$ is an absolutely continuous random variable, then*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\mathcal{D}, \mathcal{T}' | \theta}(\hat{p}(\mathcal{D}; \Theta_0) \leqslant \alpha) \xrightarrow[B' \longrightarrow \infty]{} \alpha.$$

Under stronger assumptions on the regression method, it is also possible to derive rates of convergence for the estimated p-values.

Assumption 2.9 (Convergence rate of the regression estimator)**.** *The regression estimator is such that*

$$\sup_{\theta \in \Theta_0} |\widehat{\mathbb{E}}[Z \mid \theta] - \mathbb{E}[Z \mid \theta]| = \mathcal{O}_P\left(\left(\frac{1}{B'}\right)^r\right).$$

for some $r > 0$.

Examples of regression estimators that satisfy Assumption 2.9 can be found in Stone (1982); Hardle et al. (1984); Donoho (1994); Yang et al. (2017).

Theorem 2.10. *Under Assumption 2.9,*

$$|p(D; \Theta_0) - \hat{p}(D; \Theta_0)| = \mathcal{O}_P\left(\left(\frac{1}{B'}\right)^r\right).$$

2.4.3 Power of `BFF`

In this section, we provide convergence rates for `BFF` and show that its power relates to the integrated squared error

$$\mathcal{L}(\widehat{\mathbb{O}}, \mathbb{O}) := \int \left(\hat{\mathbb{O}}(x; \theta) - \mathbb{O}(x; \theta)\right)^2 \mathrm{d}G(x)\mathrm{d}\pi(\theta), \tag{2.14}$$

which measures how well we are able to estimate the odds function. We assume that we are testing a simple hypothesis $H_{0,\theta_0} : \theta = \theta_0$, where θ_0 is fixed, and that $G(x)$ is the marginal distribution of $X \sim F_\theta(X)$ with respect to $\pi(\theta)$. We also assume that x contains all observations; that is, $X = \mathcal{D}$. In this case, the denominator of the average odds is

$$\begin{aligned} \int_\Theta \mathbb{O}(x, \theta)\mathrm{d}\pi(\theta) &= \int_{\Theta_1} \frac{p \cdot p(x \mid \theta)}{(1-p)g(x)}\mathrm{d}\pi(\theta) \\ &= \frac{p}{1-p}\int_\Theta \frac{p(x \mid \theta)}{\int_\Theta p(x \mid \theta)\mathrm{d}\pi(\theta)}\mathrm{d}\pi(\theta) = \frac{p}{1-p}, \end{aligned} \tag{2.15}$$

where g is the density of G with respect to ν and therefore there is no need to estimate the denominator in Equation (2.10). We also assume that the odds and estimated odds are both bounded away from zero and infinity:

Assumption 2.11 (Bounded odds and estimated odds)**.** *There exists $0 < m, M < \infty$ such that for every $\theta \in \Theta$ and $x \in \mathcal{X}$, $m \leqslant \mathbb{O}(x; \theta), \widehat{\mathbb{O}}(x; \theta) \leqslant M$.*

Finally, we assume that the CDF of the power function of the test based on the `BFF` statistic τ in Equation (2.10) is smooth in a Lipschitz sense:

Assumption 2.12 (Smooth power function)**.** *For every $\theta_0 \in \Theta$, the cumulative distribution function of $\tau(\mathcal{D};\theta_0)$, F_τ, is Lipschitz with constant C_L, i.e., for every $x_1, x_2 \in \mathbb{R}$, $|F_\tau(x_1) - F_\tau(x_2)| \leqslant C_L|x_1 - x_2|$.*

With these assumptions, we can relate the odds loss with the probability that the outcome of `BFF` is different from the outcome of the test based on the Bayes factor:

Theorem 2.13. *For fixed $c \in \mathbb{R}$, let $\phi_{\tau;\theta_0}(\mathcal{D}) = \mathbb{1}\left(\tau(\mathcal{D};\theta_0) < c\right)$ and $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = \mathbb{1}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ be the testing procedures for testing $H_{0,\theta_0} : \theta = \theta_0$ based on τ and $\widehat{\tau}_B$, respectively. Under Assumptions 2.11-2.12, for every $0 < \epsilon < 1$ and $\theta \in \Theta$,*

$$\int \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\pi(\theta_0) \leqslant \frac{2MC_L \cdot \sqrt{L(\widehat{\mathbb{O}}, \mathbb{O})}}{\epsilon} + \epsilon,$$

where T denotes the realized training sample $\mathcal{T}$ and $\mathbb{P}_{\mathcal{D}|\theta,T}$ is the probability measure integrated over the observable data $\mathcal{D} \sim F_\theta$, but conditional on the train sample used to create the test statistic.

Theorem 2.13 demonstrates that, on average (over $\theta_0 \sim \pi$), the probability that hypothesis tests based on the `BFF` statistic versus the Bayes factor lead to different conclusions is bounded by the integrated odds loss. This result is valuable because the integrated odds loss is easy to estimate in practice, and hence provides us with a practically useful metric. For instance, the integrated odds loss can serve as a natural criterion for selecting the "best" statistical model out of a set of candidate models with different classifiers, for tuning model hyperparameters, and for evaluating model fit.

Next, we provide rates of convergence of the test based on `BFF` to the test based on the Bayes factor. We assume that the chosen probabilistic classifier has the following rate of convergence:

Assumption 2.14 (Convergence rate of the probabilistic classifier)**.** *The probabilistic classifier trained with $\mathcal{T}$, $\widehat{\mathbb{P}}(Y = 1 \mid x, \theta)$ is such that*

$$\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y = 1 \mid x, \theta) - \mathbb{P}(Y = 1 \mid x, \theta)\right)^2 \mathrm{d}H(x, \theta)\right] = \mathcal{O}\left(B^{-\kappa/(\kappa+d+p)}\right),$$

for some $\kappa, d, p > 0$, where $H(x, \theta)$ is a measure over $\mathcal{X} \times \Theta$.

Typically, κ relates to the smoothness of $\mathbb{P}$, while d and p relate to the number of covariates of the classifier — in our case, the number of parameters plus the number of features. In Supplementary Material I, we provide some examples where Assumption 2.14 holds. We also assume that the density of the product measure $G \times \pi$ is bounded away from infinity.

Assumption 2.15 (Bounded density)**.** *$H(x, \theta)$ dominates $H' := G \times \pi$, and the density of H' with respect to H, denoted by h', is such that there exists $\gamma > 0$ with $h'(x, \theta) < \gamma$, $\forall x \in \mathcal{X}, \theta \in \Theta$.*

If the probabilistic classifier has the convergence rate given by Assumption 2.14, then the average probability that hypothesis tests based on the `BFF` statistic versus the Bayes factor goes to zero has the rate given by the following theorem.

Theorem 2.16. *Let $\phi_{\tau;\theta_0}(\mathcal{D})$ and $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})$ be as in Theorem 2.13. Under Assumptions 2.11-2.15, there exists $K' > 0$ such that, for any $\theta \in \Theta$,*

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\pi(\theta_0) \leqslant K' B^{-\kappa/(4(\kappa+d+p))}.$$

Corollary 2.17. *Under Assumptions 2.11-2.15, there exists $K' > 0$ such that, for any $\theta \in \Theta$,*

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = 1)\, d\theta_0 \geqslant \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1)\mathrm{d}\theta_0 - K' B^{-\kappa/(4(\kappa+d+p))}.$$

Corollary 2.17 tells us that the average power of the `BFF` test is close to the average power of the exact Bayes factor test. This result also implies that `BFF` converges to the most powerful test in the Neyman-Person setting, where the Bayes factor test is equivalent to the LRT.

2.5 Handling Nuisance Parameters

In most applications, we only have a small number of parameters that are of primary interest. The other parameters in the model are usually referred to as nuisance parameters. In this setting, we decompose the parameter space as $\Theta = \mathcal{M} \times \mathcal{N}$, where $\mathcal{M}$ contains the parameters of interest, and $\mathcal{N}$ contains nuisance parameters. Our goal is to construct a confidence set for $\mu \in \mathcal{M}$. To guarantee frequentist coverage by Neyman's inversion technique, however, one needs to test null hypotheses of the form $H_{0,\mu} : \mu = \mu_0$ by comparing the test statistics to the cutoffs $\widehat{C}_{\mu_0} := \inf_{\nu\in\mathcal{N}} \widehat{C}_{(\mu_0,\nu)}$ (Section 2.3.3). That is, one needs to control the type-I error at each μ_0 for *all* possible values of the nuisance parameters. Computing such infimum can be numerically unwieldy, especially if the number of nuisance parameters is large (van den Boom et al., 2020; Zhu et al., 2020). Below we propose approximate schemes for handling nuisance parameters.

In `ACORE`, we use a hybrid resampling or "likelihood profiling" method (Chuang and Lai, 2000; Feldman, 2000; Sen et al., 2009) to circumvent unwieldy numerical calculations as well as to reduce computational cost. For each μ (on a fine grid over $\mathcal{M}$), we first compute the "profiled" value

$$\widehat{\nu}_\mu = \arg\max_{\nu\in\mathcal{N}} \prod_{i=1}^{n} \widehat{\mathbb{O}}\left(x_i^{\mathrm{obs}}; (\mu, \nu)\right),$$

which (because of the odds estimation) is an approximation of the maximum likelihood estimate of ν at the parameter value μ for observed data D. By definition, the estimated `ACORE` test statistic for the hypothesis $H_{0,\mu_0} : \mu = \mu_0$ is exactly given by $\widehat{\Lambda}(\mathcal{D}; \mu_0) = \widehat{\Lambda}(\mathcal{D}; (\mu_0, \widehat{\nu}_{\mu_0}))$. However, rather than comparing this statistic to $\widehat{C}_{\mu_0}$, we use the hybrid cutoff

$$\widehat{C}'_{\mu_0} := \widehat{F}^{-1}_{\widehat{\Lambda}(\mathcal{D};\mu_0)\big|(\mu_0,\widehat{\nu}_{\mu_0})}(\alpha \mid \mu_0, \widehat{\nu}_{\mu_0}), \tag{2.16}$$

where $\widehat{F}^{-1}$ is obtained via a quantile regression as in Algorithm 2.1, but using a training sample $\mathcal{T}'$ generated at *fixed* $\widehat{\nu}_{\mu_0}$ (that is, we run Algorithm 2.1 with the proposal distribution

$\pi'((\mu,\nu)) \propto \pi(\mu) \times \delta_{\widehat{\nu}_\mu}(\nu)$, where $\delta_{\widehat{\nu}_\mu}(\nu)$ is a point mass distribution at $\widehat{\nu}_\mu$). Alternatively, one can compute the p-value

$$\hat{p}(D;\mu_0) := \widehat{\mathbb{E}}\left[\mathbb{1}\left(\widehat{\Lambda}\left(\mathcal{D};\mu_0\right) < \widehat{\Lambda}\left(D;\mu_0\right)\right) \mid \mu_0, \widehat{\nu}_{\mu_0}\right] \tag{2.17}$$

via probabilistic classification as in Algorithm A.3, but with $\mathcal{T}'$ simulated at fixed $\widehat{\nu}_{\mu_0}$ (that is, we run Algorithm A.3 with the proposal distribution $\pi'((\mu,\nu)) \propto \pi(\mu) \times \delta_{\widehat{\nu}_\mu}(\nu)$. Hybrid methods do not always control α, but they are often a good approximation that leads to robust results (Aad et al., 2012a; Qian et al., 2016). We refer to `ACORE` approaches based on Equation 2.16 or Equation 2.17 as "`h-ACORE`" approaches.

In contrast to `ACORE`, the `BFF` test statistic averages (rather than maximizes) over nuisance parameters. Hence, instead of adopting a hybrid resampling scheme to handle nuisance parameters, we approximate p-values and critical values, in what we refer to as "`h-BFF`", by using the marginal model of the data $\mathcal{D}$ at a parameter of interest μ:

$$\widetilde{\mathcal{L}}(D;\mu) = \int_{\nu \in \mathcal{N}} \mathcal{L}(D;\theta)\, \mathrm{d}\pi(\nu).$$

We implement such a scheme by first drawing the train sample $\mathcal{T}'$ from the entire parameter space $\Theta = \mathcal{M} \times \mathcal{N}$, and then applying quantile regression (or probabilistic classification) using μ only. Algorithm A.5 details our construction of `ACORE` and `BFF` confidence sets when calibrating critical values under the presence of nuisance parameter (construction via p-value estimation is analogous). In Section 2.6.2, we demonstrate how our diagnostics branch can shed light on whether or not the final results have adequate frequentist coverage.

2.6 Experiments

We analyze the empirical performance of the LF2I framework under different problem settings: unknown null distribution of (known) test statistic (Section 2.6.1); nuisance parameters (Section 2.6.2); intractable likelihood and high-dimensional data (Section 2.6.3). We use the cross-entropy loss (Equation (A.7)) when estimating the odds function in Equation (2.7) and the empirical coverage probability as in Section 2.3.4 via probabilistic classification. Moreover, we use the pinball loss (Koenker et al., 2017) when estimating critical values as in Section 2.3.3 via quantile regression.

2.6.1 Gaussian Mixture Model: Unknown Null Distribution

A common practice in LFI is to first estimate the likelihood and then assume that the LR statistic is approximately χ^2 distributed according to Wilks' theorem (Drton, 2009). However, in settings with small sample sizes or irregular statistical models, such approaches may lead to confidence sets with incorrect coverage; it is often difficult to identify exactly when that happens, and then know how to recalibrate the confidence sets. (See Algeri et al. (2019) for a discussion of all conditions needed for Wilks' theorem to apply, which are often not realized in practice.)

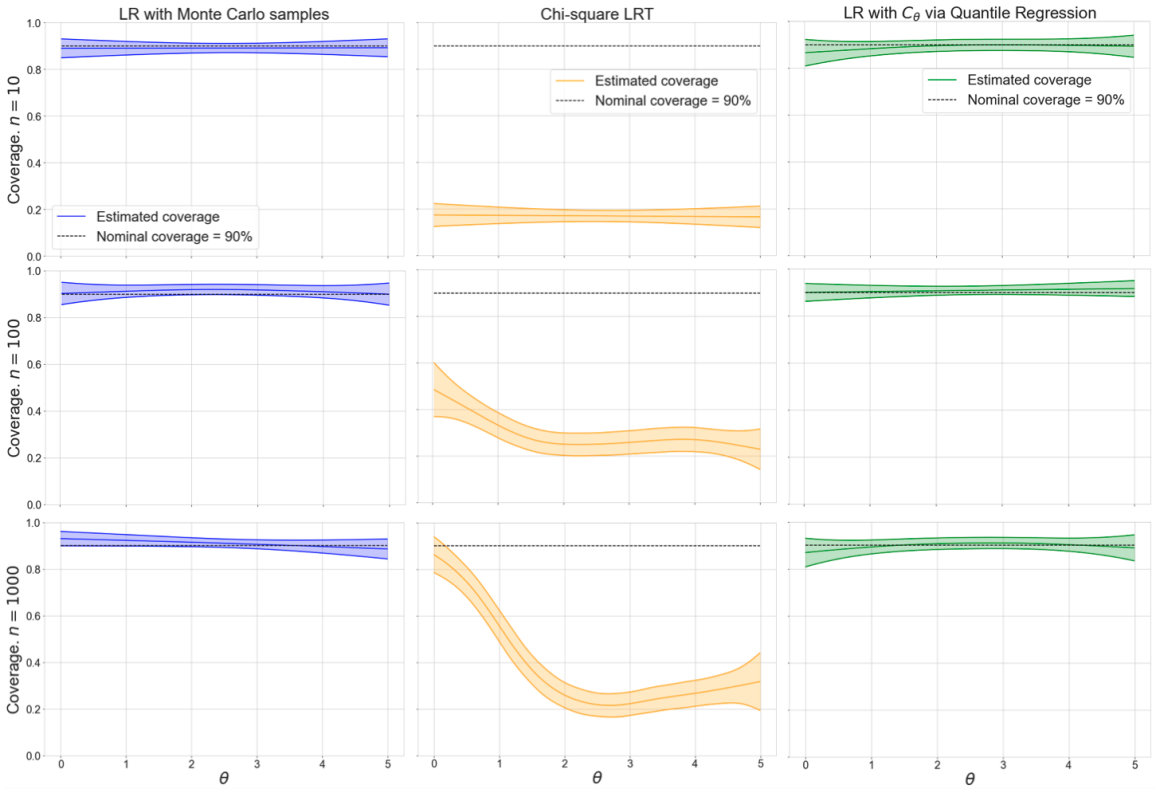


Figure 2.3: **GMM with unknown null distribution**. Each panel shows the estimated coverage across the parameter space of 90% confidence sets for θ. Rows represent experiments with different observed sample sizes: $n = 10, 100, 1000$ (top, center, bottom). Columns represent three different approaches. **Left:** "LR with Monte Carlo samples" achieves nominal coverage everywhere but is computationally expensive, especially in higher dimensions. **Center:** "Chi-square LRT" clearly under-covers, i.e. confidence sets are not valid even for large n, other than at $\theta = 0$ where the mixture collapses to one Gaussian. **Right:** "LR with C_{θ_0} via quantile regression" returns finite-sample confidence sets with the nominal coverage of 90% for all values of θ, but using a total of 1000 simulations, instead of a MC sample of 1000 simulations at each grid point.

The Gaussian mixture model (GMM) is a classical example where the LR statistic is known but its null distribution is unknown in finite samples. Indeed, the development of valid statistical methods for GMM is an active area of research (Redner, 1981; McLachlan, 1987; Dacunha-Castelle and Gassiat, 1997; Chen and Li, 2009; Wasserman et al., 2020). Here we consider a one-dimensional Normal mixture with unknown mean but known unit variance:

$$X \sim 0.5\mathcal{N}(\theta, 1) + 0.5\mathcal{N}(-\theta, 1),$$

where the parameter of interest $\theta \in \Theta = [0, 5]$. In this example, the LRT statistic is not estimated but computed exactly. The goal is to analyze three different approaches for estimating the critical value C_{θ_0} of a level-α LRT of the hypothesis test $H_{0,\theta_0} : \theta = \theta_0$, for different $\theta_0 \in \Theta$, in a setting where we have removed potential effects of estimation errors in the test statistic:

- "LR with Monte Carlo samples", where we draw 1000 simulations at each point θ_0 on a fine grid over Θ and take C_{θ_0} to be the $1-\alpha$ quantile of the distribution of the LR statistic, computed using the MC samples at each fixed θ_0. This approach is often just referred to as MC hypothesis testing.
- "Chi-square LRT", where we *assume* that $-2\text{LR}(\mathcal{D};\theta_0) \sim \chi^2_1$, and hence take $-2C_{\theta_0}$ to be the same as the upper α quantile of a χ^2_1 distribution.
- "LR with C_{θ_0} via quantile regression", where we estimate C_{θ_0} via quantile regression (Algorithm 2.1) based on a total of $B' = 1000$ simulations of size n sampled uniformly on Θ.

We then construct confidence sets by inverting the hypothesis tests, and finally assess their conditional coverage with the diagnostic branch of the LF2I framework (Algorithm 2.2 with $B'' = 1000$).

Figure 2.3 shows LF2I diagnostics for the three different approaches when the observed sample size (i.e., the number of observations from each unknown θ) is $n = 10, 100, 1000$. Confidence sets from "Chi-square LRT" are clearly not valid at any n, which shows that Wilks' theorem does not apply in this setting. The only exception arises when n is large enough and θ approaches 0, in which case the mixture reduces to a *unimodal* Gaussian whose LR statistic has a known limiting distribution (see bottom center panel of Figure 2.3). On the other hand, "LR with C_{θ_0} via quantile regression" returns valid finite-sample confidence sets with conditional coverage equivalent to "LR with Monte Carlo samples". A key difference between the LF2I and MC methods is that the LF2I results are based on 1000 samples in total, whereas the MC results are based on 1000 MC samples at each θ_0 on a grid. The latter approach quickly becomes intractable in higher parameter dimensions and larger scales.

In Appendix A.5, we show that critical values are clearly non-constant across the parameter space, which also provides insight as to why assumptions of a pivotal test statistic (e.g., a χ^2-distributed test statistic asymptotically, or calibration based on a single point in the parameter space Warne et al. (2024)) do not yield correct coverage. Supplementary Material J gives details on the specific quantile regressor (for Algorithm 2.1) and probabilistic classifier (for Algorithm 2.2) used in Figure 2.3, and presents extensions of the above experiments to confidence sets via p-value estimation and asymmetric mixtures.

2.6.2 Poisson Counting Experiment: Nuisance Parameters and Diagnostics

Hybrid methods, which maximize or marginalize over nuisance parameters, do not always control the type-I error of statistical tests. For small sample sizes, there is no theorem as to whether profiling or marginalization of nuisance parameters will give better frequentist coverage for the parameter of interest (Cousins, 2018, Section 12.5.1). In addition, most practitioners consider a thorough check of frequentist coverage to be impractical (Cousins, 2018, Section 13). In this example, we apply the hybrid schemes from Section 2.5 to a high-energy physics (HEP) counting experiment (Lyons, 2008; Cowan et al., 2011b; Cowan,

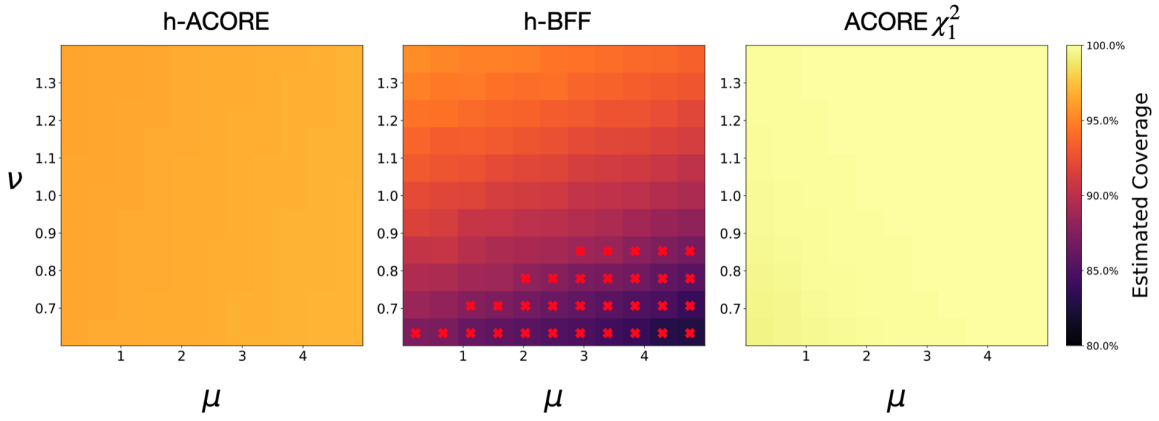


Figure 2.4: **Poisson counting experiment with nuisance parameters.** The diagnostics branch provides guidance as to which LFI approach to use for the problem at hand by pinpointing regions of the parameter space Θ where inference is unreliable. The panels show empirical coverage as a function of both μ, the parameter of interest, and ν, the nuisance parameter. Nominal coverage is 90%. **Left:** `h-ACORE`, which uses profiled likelihoods, is overly conservative in terms of actual coverage ($\approx 96\%$) across Θ. **Center:** `h-BFF`, which marginalizes over ν, under-covers in several regions (red crosses). **Right:** `ACORE` χ^2_1, which uses cutoffs from the chi-square distribution, has almost no constraining power, yielding empirical coverage close to 100% everywhere.

2012; Cousins et al., 2008; Heinrich, 2022) with nuisance parameters, which is a simplified version of a real particle physics experiment where the true likelihood function is not known. We illustrate how our diagnostics can guide the analyst and provide insight into which method to choose for the problem at hand.

Consider a "Poisson counting experiment" where particle collision events are counted under the presence of both an uncertain background process and a (new) signal process. The goal is to estimate the signal strength. To avoid identifiability issues, the background rate is estimated separately by counting the number of events in a control region where the signal is believed to be absent. Hence, the observable data $X = (N_b, N_s)$ contain two measurements, where $N_b \sim \text{Poisson}(\nu\tau b)$ is the number of events in the control region, and $N_s \sim \text{Poisson}(\nu b + \mu s)$ is the number of events in the signal region. Our parameter of interest is the signal strength μ, whereas the scaling factor for the background ν is a nuisance parameter. The hyper-parameters s and b indicate the nominally expected counts from signal and backgrounds, and τ describes the relationship in measurement time between the two processes. We treat the three hyper-parameters as known with values $s = 15$, $b = 70$, $\tau = 1$, respectively. The hyper-parameters move the model away from the Gaussian limiting regime and make the relationship between data and parameters more complicated Heinrich (2022).

We compare the hybrid methods `h-ACORE` and `h-BFF` with `ACORE` χ^2_1 (which uses cutoffs from the chi-square distribution). We learn the odds using a QDA classifier with $B = 100{,}000$ and estimate critical values for the hybrid methods via quantile gradient

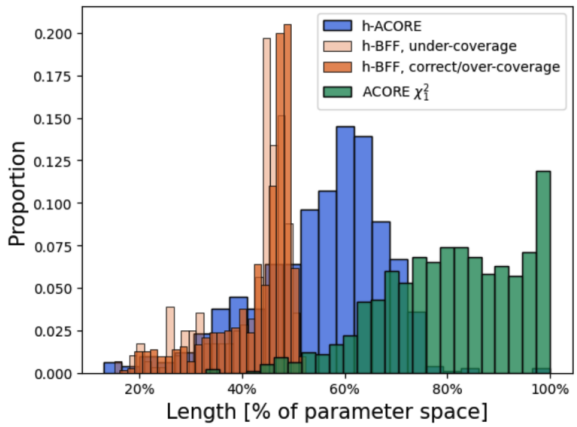


Figure 2.5: **Constraining power.** Relative size of the confidence sets constructed in Section 2.6.2. `ACORE` χ^2_1 and `h-ACORE` yield the widest intervals (they are indeed overly conservative according to Figure 2.4). `h-BFF` provides tighter confidence sets, but their size cannot be trusted when the method under-covers. LF2I diagnostics can identify the parameter regions where the approach is not valid (red crosses in Figure 2.4). The dark-orange histogram reports `h-BFF` results after removing those points.

boosted trees with $B' = 10{,}000$. We evaluate the different methods on a separate set of size $B'' = 1000$ by estimating coverage and measuring the length of confidence sets for each of the simulated samples.

Figure 2.4 shows the estimated coverage as a function of both μ and ν. Confidence sets are considered to be valid when they achieve the nominal coverage level regardless of the true value of *both* the parameter of interest and the nuisance parameters. Both `h-ACORE` and `ACORE` χ^2_1 are overly conservative across the whole parameter space, while `h-BFF` under-covers in regions of high signal strength and low background. These results are consistent with the length of the corresponding confidence sets shown in Figure 2.5: `h-ACORE` and `ACORE` χ^2_1 are overly conservative, with the former being almost uninformative for the majority of evaluation samples. On the other side, while `h-BFF` seems to provide tighter parameter constraints, their length can be trusted only in regions where the method has coverage at least equal to the nominal level. Our LF2I diagnostic branch can pinpoint the regions of the parameter space where inference is reliable or not.

2.6.3 Muon Energy Estimation: Intractable and High-Dimensional Likelihood

We now showcase LF2I on a high-energy physics application with intractable likelihood and very high-dimensional data. The goal is to estimate the energy of muons using a high-granularity calorimeter in a particle collider experiment. Muons are subatomic particles that have proven to be excellent probes of new physical phenomena: their detection and measurement has enabled several crucial discoveries in the last few decades, including the discovery of the Higgs boson Augustin et al. (1974); Herb et al. (1977); Collaboration et al. (1995); Aad et al. (2012b); Chatrchyan et al. (2012). Traditionally, the energy of a muon is determined from the curvature of its trajectory in a magnetic field, but curvature-based measurements have proven to be insufficiently precise at high energies. Recently, muon energy measurements based on their radiative losses in a dense, finely segmented calorimeter (Figure 2.6, left) have been shown to be a feasible alternative Kieseler et al. (2022); Dorigo et al. (2022).

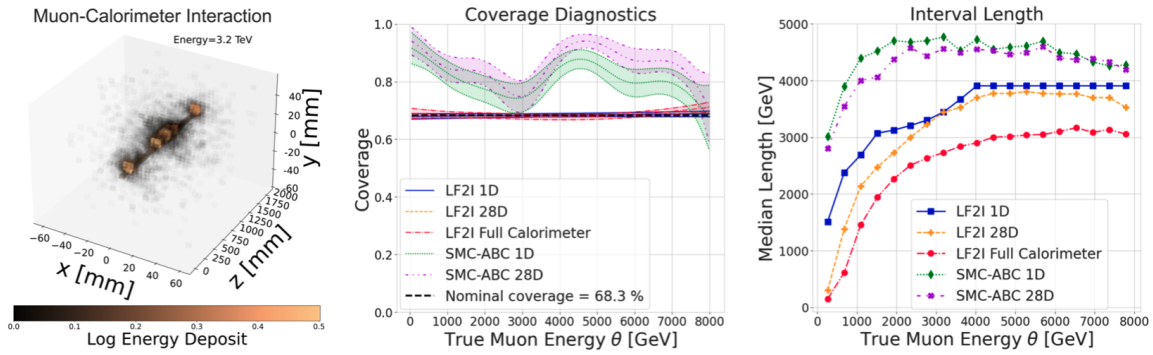


Figure 2.6: **Muon energy estimation.** LF2I guarantees nominal coverage and yields smaller confidence intervals relative to SMC-ABC. **Left:** Data point example of a muon with incoming energy $\theta \approx 3.2$ TeV entering a calorimeter with $32 \times 32 \times 50$ cells. **Center:** LF2I (blue, orange, red in the right two panels) achieves coverage at the nominal level (68.3%), whereas SMC-ABC (green and purple) is consistently over-covering across the parameter space. **Right:** Median lengths of constructed intervals. While being extremely computationally intensive, SMC-ABC has also the least constraining power regardless of the data set used. SMC-ABC on the full calorimeter data is not reported as it was computationally infeasible to run.

In this application, the dimensionality of one data point x (a 3D image) is of the order of $\approx$ 50,000 and the observed sample size is $n = 1$ (as each unique data point is the output of one experiment with a specific parameter of interest θ). In total, we have available 886,716 3D "image" inputs $\mathbf{x}$ with corresponding scalar muon energies θ. The data are obtained by accurately mimicking particle showers with `GEANT4` (Agostinelli et al., 2003), a high-fidelity simulator that has been calibrated for decades and is trusted to incorporate all the dynamics of the Standard Model of particle physics. The data are available at Kieseler et al. (2021).

The scientific goal of this experiment is to quantify whether a high-granularity calorimeter would better constrain the energy of a muon (that is, lead to smaller confidence sets) than, for example, a detector that only measures the total energy of the incoming particle. To answer this question, we consider nested versions of the same energy measurement, where the inputs to our algorithms are of increasing dimensionality: (i) a 1D input which is equal to the sum over all the cells of the calorimeter (for each muon with deposited energy $E > 0.1$ GeV); (ii) 28 custom features extracted from the spatial and energy information of the calorimeter cells (see Kieseler et al. (2022)); and (iii) the full calorimeter measurement, $\mathbf{x} \in \mathbb{R}^{51,200}$. We then construct LF2I confidence sets for each data point using `BFF`. On the full calorimeter data, we learn the odds function through a convolutional neural network classifier derived from the regressor proposed in Kieseler et al. (2022), and estimate critical values via quantile gradient boosted trees. For the 1D and 28D data sets, we instead learn odds through a gradient boosting classifier. In both cases, we use approximately 83% of the data to learn the odds function ($B = 738{,}930$) and 14% to estimate critical values ($B' = 123{,}155$). For comparison, we also include results from SMC-ABC (Sisson et al., 2007), a popular LFI algorithm from the Approximate Bayesian Computation literature. To provide a fair assessment of the results,

SMC-ABC uses all the simulations that LF2I exploits separately (i.e., $B+B' = 862{,}085$). The remaining data points ($B'' = 24{,}631$) are used for validation and diagnostics of both methods.

Figure 2.6 (center) shows that LF2I with the `BFF` test statistic achieves the nominal level of coverage (68.3%) regardless of the data set used. This is consistent with Theorem 2.3: as long as the quantile regression is well estimated, LF2I confidence sets are guaranteed to be valid at the nominal $(1-\alpha)$ level regardless of how well the test statistic is estimated. On the other hand, SMC-ABC is overly conservative with credible intervals that strongly over-cover across the whole parameter space. As to constraining power (interval length), Figure 2.6 (right) shows that SMC-ABC credible intervals are significantly wider than LF2I confidence sets for both the 1D and 28D data sets (running SMC-ABC on the 51,200-dimensional full calorimeter data was computationally infeasible, and we were not able to report the results). Finally, note how the amount of information in the data directly influences the size of LF2I confidence sets: going from the 1D data set to the full calorimeter leads to noticeably smaller confidence intervals, and hence higher constraining power.

Remark on validity and computational cost SMC-ABC does not have the right coverage, because the goal of ABC is to construct Bayesian credible regions and not valid confidence sets; see, e.g., Hermans et al. (2021) for other examples of SMC-ABC under- or over-covering. Furthermore, note that (i) LF2I is amortized: once training is done, confidence sets can be efficiently computed on an arbitrary number of observations without having to retrain the algorithms; and (ii) there is no need for a prior dimension reduction of the data (that is, we can directly input the three-dimensional image). Specifically, LF2I required approximately 10 and 5 CPU minutes on an AMD's EPYC 7763 machine to train the odds classifier and the quantile regressor respectively, and less than a second to obtain confidence intervals all at once for all observations (in this example, unique 24,631 "test" muons) regardless of their dimensionality. In contrast, SMC-ABC required approximately 1 CPU hour for *each* observation even for the lower-dimensional 1D and 28D data sets.

2.7 Conclusions and Discussion

Validity. Our proposed LF2I methodology leads to frequentist confidence sets and hypothesis tests with finite-sample guarantees (when there are no nuisance parameters). *Any* existing or new test statistic – that is, not only estimates of the LR or BF statistics – can be plugged into our framework to create tests that control type I error. The implicit assumption is that the null distribution of the test statistic varies smoothly in parameter space. If that condition holds, then we can efficiently leverage quantile regression methods to construct valid confidence sets by a Neyman inversion of simple hypothesis tests, without having to rely on asymptotic results.

Nuisance parameters and diagnostics. For small sample sizes, no theorem guarantees whether profiling or marginalizing nuisance parameters will provide better frequentist coverage for the parameter of interest (Cousins, 2018, Section 12.5.1). It is generally believed that hybrid resampling methods return approximately valid confidence sets, but that a

rigorous check of validity is infeasible when the true solution is not known. Our diagnostic branch presents practical tools for assessing empirical coverage across the entire parameter space (including nuisance parameters). After seeing the results, one can decide which method is most appropriate for the application at hand. For example, in the Poisson counting experiment of Section 2.6.2, LF2I diagnostics revealed that `h-BFF` (which averages the estimated odds over nuisance parameters) returned smaller confidence intervals, but at the cost of under-covering in some regions of the parameter space.

Power. Statistical power is the hardest property to achieve in practice in LFI. This is the area where we foresee that most statistical and computational advances will take place. As shown theoretically in Theorem 2.13 and empirically in Supplementary Material K, the power (or size) of LF2I confidence sets depends not only on the theoretical properties of the (exact) test statistics, but is also influenced by how precisely we are able to estimate it. In the case of `ACORE` and `BFF`, the latter can be divided in (i) how well we are able to estimate the likelihood or odds function (a statistical estimation error), and (ii) how accurate are the integration or maximization procedures we use (a purely numerical error); see Supplementary Material H for a more precise breakdown of the sources of error in LF2I confidence sets, particularly for `ACORE` and `BFF`. Machine learning offers exciting possibilities on both fronts. For example, with regards to (i), Brehmer et al. (2020) offers compelling evidence that one can can dramatically improve estimates of the likelihood $p(x \mid \theta)$ for $\theta \in \Theta$, or the likelihood ratio $p(x \mid \theta_1, \theta_2)$ for $\theta_1, \theta_2 \in \Theta$, by a "mining gold" approach that extracts additional information from the simulator about the latent process. Future work could incorporate such an approach into the LF2I framework, with the calibration and diagnostic branches as separate modules.

Other test statistics. Our work presents also another new direction for LF2I: So far frequentist LFI methods have been estimating either likelihoods or likelihood ratios, and then often relying on asymptotic properties of the LR statistic. We note that there are settings where it may be easier to either estimate the posterior $p(\theta \mid x)$ rather than the likelihood $p(x \mid \theta)$, or alternatively to obtain point estimates for parameters directly via predictions algorithms. Because the LF2I framework is agnostic to which algorithms we use to construct the test statistic itself, we can potentially leverage methods that estimate the conditional mean $\mathbb{E}[\theta \mid x]$ and variance $\mathbb{V}[\theta \mid x]$ to construct frequentist confidence sets and hypothesis tests for θ with finite-sample guarantees. For example, Masserano et al. (2023) — whose content we cover in Chapter 3 — uses $T = \frac{(\mathbb{E}[\theta|x]-\theta_0)^2}{\mathbb{V}[\theta|x]}$, which in some scenarios corresponds to the Wald statistic for testing $H_{0,\theta_0} : \theta = \theta_0$ against $H_{1,\theta_0} : \theta \neq \theta_0$ Wald (1943), as an attractive alternative to get LF2I confidence sets from prediction algorithms and posterior estimators.

See Appendices A-F for proofs and details on the algorithms, and refer to the separate Supplementary Material file[5] for additional experiments and results referenced in the main text.

[5]Available at https://lucamasserano.github.io/data/LF2I_supplementary_material.pdf.

2.7.1 Related Work

Classical statistical inference in high-energy physics (HEP) LF2I is inspired by pioneering work in HEP that adopted classical hypothesis tests and Neyman confidence sets for the discovery of new physics (Feldman and Cousins, 1998; Cowan et al., 2011b; Aad et al., 2012a; Chatrchyan et al., 2012; Cranmer, 2015). Our work grew from the discussion in HEP regarding theory and practice, and open problems such as how to efficiently construct Neyman confidence sets for general settings (Cowan et al., 2011b), how to assess coverage across the parameter space without costly Monte Carlo simulations (Cousins, 2018), and how to choose hybrid techniques in practice (Cousins, 2006). This paper proposes a general approach to solve the above-mentioned open problems with a modular framework that can be adapted to fit the data at hand.

Universal inference. Recently, Wasserman et al. (2020) proposed a "universal" inference test statistic for constructing valid confidence sets and hypothesis tests with finite-sample guarantees without regularity conditions. The assumptions are that the likelihood $\mathcal{L}(\mathcal{D};\theta)$ is known and that one can compute the maximum likelihood estimator (MLE). Our LF2I framework does *not* require a tractable likelihood, but it assumes that we have regression methods that can estimate the chosen test statistic and its critical values. In tractable likelihood settings where both universal inference and LF2I apply, the LF2I approach leads to more powerful tests than universal inference (see, e.g., Figure 11 in Supplementary Material).

Simulation-based calibration of Bayesian posterior distributions. In Bayesian inference, the posterior distribution $\pi(\theta \mid x)$ is fundamental for quantifying uncertainty about the parameter θ given the data x. Recent methods have been developed to assess the quality of estimated posterior distributions; that is, assessing whether an estimate $\widehat{\pi}(\theta \mid x)$ is consistent with the posterior distribution $\pi(\theta \mid x)$ implied by the assumed prior and likelihood (Dey et al., 2021; Zhao et al., 2021; Dey et al., 2022; Linhart et al., 2023; Lemos et al., 2023). *The calibration in LF2I is fundamentally different:* Even if posteriors are calibrated in the sense that $\widehat{\pi}(\theta \mid x) = \pi(\theta \mid x)$ for every x and θ, confidence sets derived from it will not necessarily have the correct empirical coverage (according to Equation (2.1)). LF2I is agnostic to the choice of the test statistic (for instance, whether the test statistic is formed from likelihoods or posteriors (Masserano et al., 2023)), and provides guarantees of how well we are able to constrain the true parameters of interest regardless of the choice of the prior or proposal distribution $\pi(\theta)$.

Likelihood-free inference via machine learning. Recent LFI methods have been using simulators output as training data to learn surrogate models for inference; see Cranmer et al. (2020) for a review. These techniques use synthetic data simulated across the parameter space to directly estimate key quantities, such as:

1. *posteriors* $p(\theta \mid x)$ (Blum and François, 2010; Marin et al., 2016; Papamakarios and Murray, 2016; Lueckmann et al., 2017; Greenberg et al., 2019; Chen and Gutmann, 2019; Izbicki et al., 2019; Radev et al., 2020);

2. *likelihoods* $p(x \mid \theta)$ (Wood, 2010; Meeds and Welling, 2014; Wilkinson, 2014; Gutmann and Corander, 2016; Fasiolo et al., 2018; Lueckmann et al., 2019; Papamakarios et al., 2019; Picchini et al., 2020; Järvenpää et al., 2021); or

3. *density ratios*, such as the likelihood-to-marginal ratio $p(x \mid \theta)/p(x)$ (Izbicki et al., 2014; Thomas et al., 2021; Hermans et al., 2020; Durkan et al., 2020b),[6] the likelihood ratio $p(x \mid \theta_1)/p(x \mid \theta_2)$ for $\theta_1, \theta_2 \in \Theta$ (Cranmer et al., 2015; Brehmer et al., 2020) or the profile-likelihood ratio (Heinrich, 2022).[7]

Recently, there have also been works that directly predict parameters θ of intractable models using neural networks Gerber and Nychka (2021); Lenzi et al. (2021) (that is, they do not estimate posteriors, likelihoods or density ratios). In addition, new methods such as normalizing flows (Papamakarios et al., 2021) and other neural density estimators are revolutionizing LFI in terms of sample efficiency and capacity, and will continue to do so.

Nonetheless, although the goal of LFI is inference on the unknown parameters θ, it remains an open question whether a given LFI algorithm produces reliable measures of uncertainty, as current methods lack guarantees of local (instance-wise) validity and power for a finite number of observations. They also have no practical diagnostics to assess local coverage across the parameter space. Our framework can be used in combination with any LFI approach that relies on a test statistic (such as the LRT) to provide both local coverage and diagnostics. Finally, thanks to the modular structure of LF2I, the diagnostic branch can be used separately to evaluate whether other approaches (like ABC and posterior methods that return credible regions) have good frequentist coverage, and in cases where they do not, LF2I can identify regions of the parameter space of over- or under-confidence.

[6] In 2014, Izbicki et al. approximate likelihoods for high-dimensional data (such as 2D images) via density ratios (Izbicki et al., 2014, Equation 3) and kernel methods, building on Izbicki's PhD thesis work on spectral series approaches to high-dimensional nonparametric inference. The kernel approximate likelihood approach was later superseded by neural SBI approaches.

[7] `ACORE` and `BFF` are based on estimating the odds $\mathbb{O}(X;\theta)$ at $\theta \in \Theta$ (Equation (2.7)); this is a "likelihood-to-marginal ratio" approach, which estimates a one-parameter function as in the original paper by Izbicki et al. (2014). The likelihood ratio $\mathbb{OR}(X;\theta_0,\theta_1)$ at $\theta_0, \theta_1 \in \Theta$ (Equation (2.9)) is then computed from the odds function, without the need for an extra estimation step.

3

Confidence Sets from Prediction Algorithms and Posterior Estimators

3.1 Introduction

The vast majority of modern machine learning targets prediction problems, with algorithms such as Deep Neural Networks (DNNs) being particularly successful with point predictions of a target variable $Y \in \mathbb{R}$ when the input vectors $x \in \mathcal{X}$ represent complex high-dimensional data. In many science applications, however, one is often interested in the "inverse" problem of estimating the internal parameters of a data-generating process with reliable measures of uncertainty. The parameters of interest, which we denote by θ, are then not directly observed but are the "causes" of the observed data x.

In order to make inference on internal parameters, one needs a statistical model that relates the (unknown) parameters to the observed data. In science and engineering, simulations are often used to model the behavior of complex systems in lieu of an analytical likelihood, when the latter is too complicated to be evaluated explicitly. Let $\mathcal{D} \coloneqq (X_1, \ldots, X_n)$ denote observable data, where the "sample size" n refers to the number of observations at a fixed configuration of the parameters θ. *Likelihood-free inference* (LFI), which is a form of simulator-based inference (SBI; Cranmer et al. (2020)), refers to parameter estimation in a setting where the likelihood function $\mathcal{L}(\theta; \mathcal{D}) \coloneqq p(\mathcal{D} \mid \theta)$ itself is intractable, but the scientist, in lieu of an explicit likelihood, has access to a simulator that can generate $\mathcal{D}$ given any $\theta \in \Theta$.

LFI has undergone a revolution in terms of the complexity of problems that can be tackled, both because of faster and more realistic simulators that can generate a large number of examples $\mathcal{T} = \{(\theta_i, \mathcal{D}_i)\}_{i=1}^{B}$, and because of more powerful AI techniques that can learn various quantities of interest from these simulations. DNNs — such as convolutional neural networks (CNNs) (LeCun et al., 1995) — are now used in many domain sciences to directly *predict* internal parameters of interest in statistical models, especially in settings where X represents images or other high-dimensional data. Recent examples include estimating the energy (θ) of muons that radiate photons when traversing a finely segmented calorimeter (X) (Kieseler et al., 2022); estimating the mass of a galaxy cluster (θ) from velocities and projected radial distances (X) for a particular line-of-sight of the observer relative to the

galaxy cluster (Ho et al., 2019); and estimating the range and noise-to-signal covariance parameters (θ) of spatial Gaussian processes from spatial fields or variograms (X) (Gerber and Nychka, 2021). In parallel, modern neural density estimators, such as normalizing flows, are becoming increasingly popular for uncertainty quantification, especially when both parameters θ and observations X are high-dimensional. Recent examples include Boyda et al. (2021); Mishra-Sharma and Cranmer (2022); Lueckmann et al. (2021).

Purely predictive approaches are known to suffer from prediction bias in inverse problems, as the point prediction — e.g., $\mathbb{E}[\theta \mid x]$ under squared error loss — is generally different from the true (unknown) parameter θ. Concrete examples include Dorigo et al. (2022); Ho et al. (2019); Kiel et al. (2019), where attempts are made to correct for the observed bias post-hoc. At the same time, many posterior estimation methods are known to be overly confident, meaning that they yield confidence sets with empirical coverage lower than the desired nominal level (Hermans et al., 2021), hence leading to potentially misleading results. At the heart of the matter is the fact that both predictive and posterior approaches in SBI rely heavily on how the values of θ in the training set $\mathcal{T}$ are sampled. For reliable inference, however, the coverage guarantees of the confidence sets should be independent of the choice of prior π_θ, thereby allowing the user to design priors that can lead to tighter, *but guaranteed to be valid*, confidence sets. In this chapter, we present a solution without relying on large-sample theory or computationally intensive Monte Carlo sampling.

`Waldo` is a new LFI procedure that can leverage any prediction algorithm or neural posterior estimator to construct confidence regions for θ with correct *conditional coverage*; that is, sets $\mathcal{R}(\mathcal{D})$ satisfying

$$\mathbb{P}_{\mathcal{D}|\theta}(\theta \in \mathcal{R}(\mathcal{D})) = 1 - \alpha, \quad \forall \theta \in \Theta, \tag{3.1}$$

regardless of the size n of the observed sample, where $(1 - \alpha) \in (0, 1)$ is a prespecified confidence level. Note that this is the same definition we gave in Equation (2.1) in Chapter 2. Correct conditional coverage implies correct *marginal coverage*, $\mathbb{P}(\theta \in \mathcal{R}(\mathcal{D})) = 1 - \alpha$, but the former is a stronger requirement that checks that the confidence set is calibrated no matter what the true parameter is, whereas marginal coverage only requires the set to be calibrated on average over the parameter space Θ. `Waldo` reframes the Wald test (Wald, 1943) and leverages existing prediction or posterior algorithms to first compute a test statistic (Equation (3.4)) based on estimates of the conditional mean $\mathbb{E}[\theta \mid \mathcal{D}]$ and conditional variance $\mathbb{V}[\theta \mid \mathcal{D}]$. It then uses a recent approach (Dalmasso* et al., 2024) to the Neyman construction (Neyman, 1937b), which estimates critical values via quantile regression and converts hypothesis tests into a confidence region with finite-n conditional coverage. `Waldo` also includes an independent diagnostics module to check that the constructed confidence sets achieve the correct nominal level of empirical coverage across the parameter space, analogously to what we introduce in Section 2.3.4. Section 3.3.2 describes our methodology in detail, and Figure 3.1 summarizes its different components.

`Waldo` embraces the best sides of both the Bayesian and frequentist perspectives to statistical inference by providing confidence sets that *(i)* can effectively exploit available domain-specific knowledge, further constraining parameters when the prior is consistent with the data, and *(ii)* are guaranteed to have the nominal conditional coverage even in finite samples as long

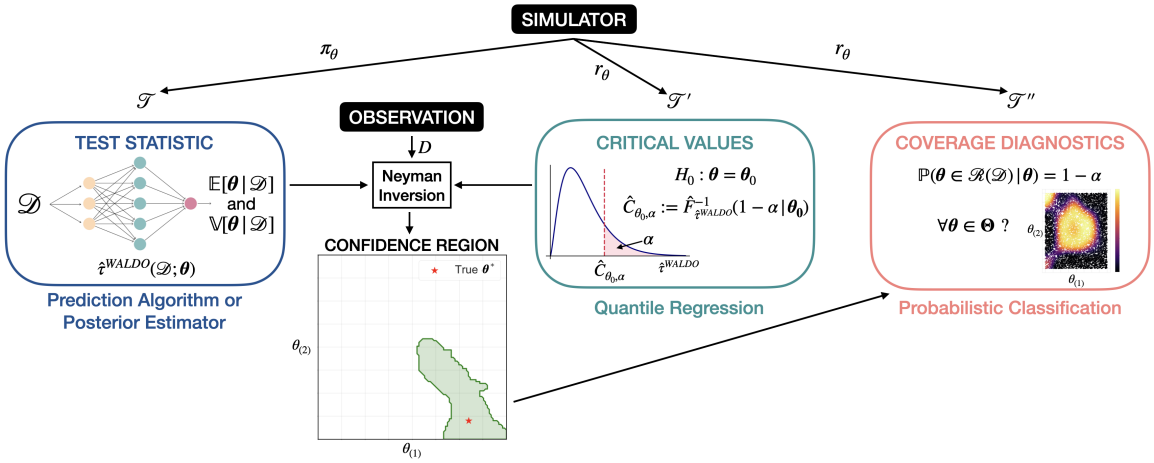


Figure 3.1: **Schematic diagram of `Waldo`. Left (blue)**: For a training set $\mathcal{T}$, we estimate the conditional mean $\mathbb{E}[\theta \mid \mathcal{D}]$ and variance $\mathbb{V}[\theta \mid \mathcal{D}]$ using a prediction algorithm (e.g., DNN) or posterior estimator (e.g., normalizing flows). This gives us the `Waldo` test statistic $\hat{\tau}_{\texttt{Waldo}}$ in Equation (3.4). **Center (green):** For a calibration set $\mathcal{T}'$, we estimate critical values $\widehat{C}_{\theta_0,\alpha}$ for all tests $H_0 : \theta = \theta_0$ across the parameter space Θ via a quantile regression of $\hat{\tau}_{\texttt{Waldo}}$ on θ. **Bottom:** Given an observation D, Neyman inversion converts the tests (which compare test statistics with critical values) into a confidence region for θ. **Right (red):** For a validation set $\mathcal{T}''$, we provide an independent assessment of the conditional validity of constructed confidence regions by computing coverage diagnostics across the entire parameter space. See Section 3.3.2 and Algorithm 3.1 for details.

as the quantile regressor is well estimated, regardless of the correctness of the prior. `Waldo` is also amortized, meaning that once the procedure has been trained, it can be evaluated on any number of observations. We lay out the statistical and computational properties of `Waldo`, providing synthetic examples with analytical solutions to verify and support our claims (see Section 3.3.3 and Section 3.3.4). We then show its effectiveness on two complex applications, which confirm the results we obtained on the synthetic examples: the first one (Section 3.4.1) uses an established benchmark in SBI and leverages posterior distributions to construct valid confidence sets regardless of the prior distribution. The second application (Section 3.4.2) deals with a current problem in high-energy physics: inferring the energy of muons from a particle detector exploiting predictions from a custom CNN and an innovative source of information, i.e., the pattern of energy deposits left by muons in a finely segmented calorimeter. The results we obtain for this problem, which are closely connected to those presented in Section 2.6.3, are of scientific interest by themselves, as a rigorous estimate of the uncertainty around estimated muon energies is essential in the search of new physics. A ready-to-use and flexible implementation of `Waldo` is available at `https://github.com/lee-group-cmu/lf2i`.

Notation We refer to parameters of interest as $\theta \in \Theta \subset \mathbb{R}^d$ and to a sample of size n of observable input data as $\mathcal{D} = (X_1, \ldots, X_n)$, with $x_i \in \mathcal{X} \subset \mathbb{R}^p$ and possibly $d \neq p$. Note that n is distinct from B, B' and B'', i.e., the number of simulations required at different steps of our method. We distinguish between observable data and actual observations by denoting the latter as $D = (x_1^{\text{obs}}, \ldots, x_n^{\text{obs}})$. We refer to confidence regions as $\mathcal{R}(\mathcal{D})$. The terms "set", "region" and (when $p = 1$) "interval" are used interchangeably.

3.2 Related Work

There exist many approaches for calibrating predictive distributions $p(y \mid x)$ to achieve marginal or conditional validity in "forward" $x \to y$ problems; examples include conformal inference (Vovk et al., 2005a; Lei et al., 2018; Chernozhukov et al., 2021) and the calibration procedures of Bordoloi et al. (2010); Dey et al. (2022). In the Bayesian inference domain, such calibration procedures correspond to ensuring that an estimate $\hat{p}(\theta \mid x)$ of the posterior $p(\theta \mid \mathbf{x})$ indeed corresponds to the true posterior implied by the prior that was used. This chapter, on the other hand, deals with the question of constructing *confidence sets* with correct conditional coverage for internal unknown parameters θ in so-called "inverse problems" (recall Equation (3.1)), which is not the same as achieving conditional coverage for prediction sets, or recalibrating posteriors.

Similarly, existing approaches for deep learning uncertainty quantification (see Gawlikowski et al. (2021) for a recent review), such as Monte Carlo drop out (Gal and Ghahramani, 2016) and conformal inference DNNs (Papadopoulos et al., 2007; Angelopoulos et al., 2023b), construct prediction sets instead of confidence sets. Before `Waldo`, there has been no straightforward way to obtain confidence sets from point predictions or estimated posteriors obtained from deep neural networks and other predictive ML algorithms.

For example, various domain science applications have developed post-hoc corrections to predictive or posterior inferences to reduce observed biases and to improve the calibration of uncertainties. Such corrections are common in areas ranging from particle physics (Dorigo et al., 2022) to cosmology (Ho et al., 2019) and remote sensing (Kiel et al., 2019). Usually the goal of the corrections is to reduce the impact of the prior specification, but in contrast to `Waldo`, post-hoc correction approaches do not provide formal coverage guarantees. Similarly, in some settings, priors can be designed so that credible regions achieve correct conditional coverage (Bayarri and Berger, 2004; Berger, 2006; Kass and Wasserman, 1996; Scricciolo, 1999; Datta and Sweeting, 2005). However, this technique requires knowledge of the likelihood function (which is not available in LFI). Moreover, such prior distributions often do not encode actual prior information, a limitation that is not present in `Waldo`.

Finally, posterior inferences do not control conditional coverage even for correctly specified priors (Patil et al., 2022). `Waldo` addresses this problem using Neyman inversion via an efficient regression-based approach proposed in Dalmasso* et al. (2024), which we covered in Chapter 2. In the latter, however, we construct likelihood-based test statistics (the Bayes factor or likelihood ratio) which require an extra numerical integration or optimization step that can lead to a loss of power of the resulting confidence sets. `Waldo`, on the other hand,

directly leverages flexible prediction algorithms and posterior estimators to construct valid and potentially more precise finite-n confidence sets.

3.3 Methodology

`Waldo` leverages a regression-based approach to the Neyman construction, reframing the Wald test to use the output of common LFI prediction algorithms and posterior estimators. After outlining its statistical foundations, we describe our procedure and its properties using synthetic examples.

3.3.1 Foundational Tools from Classical Statistics

Neyman construction. A key ingredient of `Waldo` is the equivalence between hypothesis tests and confidence sets, which was formalized by Neyman (1937b). The basic idea is to invert a series of level-α hypothesis tests of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0, \tag{3.2}$$

for all $\theta_0 \in \Theta$. After observing a sample D, one constructs a confidence region $\mathcal{R}(D)$ for θ by taking all θ_0 values that were not rejected by the series of tests above. By design, the set $\mathcal{R}(\mathcal{D})$ satisfies Equation (3.1), i.e., it has the correct $1 - \alpha$ coverage across the *entire* parameter space Θ. Albeit simple, the Neyman construction requires one to control the type-I error for every $\theta \in \Theta$. It is therefore hard to implement in practice within an LFI setting, without resorting to large-n approximations like Wilks' theorem (Wilks, 1938), or to Monte Carlo approaches, which become computationally prohibitive as the dimensionality of the parameter space increases (Cousins (2018); see also Section 3.3.4).

Wald test. Since any test that controls the type-I error at level α can be used for the Neyman construction, we base `Waldo` on the classical Wald test (Wald, 1943), which is uniformly most powerful in many settings (Ghosh, 1991; Lehmann et al., 2005). The Wald test measures the agreement of the data with the null hypothesis for θ, and it has the following form for $d = 1$:

$$\tau_{\texttt{Wald}}(\mathcal{D}; \theta_0) := \frac{(\widehat{\theta}_{\texttt{MLE}} - \theta_0)^2}{\mathbb{V}(\widehat{\theta}_{\texttt{MLE}})}, \tag{3.3}$$

where $\widehat{\theta}_{\texttt{MLE}}$ is the maximum-likelihood estimator of θ and $\widehat{\mathbb{V}}(\widehat{\theta}_{\texttt{MLE}})$ can be any consistent estimator of its variance. However, in our setting, we do not have access to the likelihood and we cannot resort to assumptions on the distribution of $\tau_{\texttt{Wald}}(\mathcal{D}; \theta_0)$, nor to asymptotic regimes, which makes it difficult to directly compute the Wald test statistic.

3.3.2 Confidence Sets from Predictions and Posteriors

From Wald to `Waldo`. `Waldo` reframes the Wald test by replacing $\widehat{\theta}_{\texttt{MLE}}$ and its variance with quantities that are easily computable with prediction algorithms or posterior estimators

commonly used in LFI. We define the `Waldo` test statistic for parameters of arbitrary dimensionality d as

$$\tau_{\texttt{Waldo}}(\mathcal{D};\theta_0) = (\mathbb{E}[\theta \mid \mathcal{D}] - \theta_0)^T \mathbb{V}[\theta \mid \mathcal{D}]^{-1} (\mathbb{E}[\theta \mid \mathcal{D}] - \theta_0), \tag{3.4}$$

where $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$ are, respectively, the conditional mean and covariance matrix of θ given $\mathcal{D}$. The connection to the Wald test follows from the asymptotic behavior of Bayes estimators (e.g., Chao (1970); Ghosh and Ramamoorthi (2003); Ghosh et al. (1982); Li et al. (2020)):

$$\mathbb{E}[\theta \mid \mathcal{D}] - \widehat{\theta}_{\texttt{MLE}} = \mathcal{O}_p(n^{-1/2}) \quad \text{and} \quad \mathbb{V}[\theta \mid \mathcal{D}] - \frac{1}{n} H^{-1}(\widehat{\theta}_{\texttt{MLE}}) = \mathcal{O}_p(n^{-1}),$$

where $H^{-1}(\widehat{\theta}_{\texttt{MLE}})$ is the negative inverse Fisher information matrix evaluated at $\widehat{\theta}_{\texttt{MLE}}$. The above result implies that `Waldo` would enjoy the same asymptotic properties typical of the Wald test, making it a pivotal test statistic. On the other hand, this does not mean that Wald and `Waldo` will give the same results for small n: indeed, in Section 3.3.3 and Appendix B.1.2, we demonstrate that `Waldo` can benefit from a prior over θ that is consistent with the data to achieve smaller confidence sets, whereas the Wald test statistic only depends on the likelihood.

Likelihood-Free Frequentist Inference (LF2I). `Waldo` expands on the LF2I framework formalized in Dalmasso* et al. (2024) — see Chapter 2 — which proposed a fast construction of Neyman confidence sets using quantile regression to bypass large-sample approximations or expensive Monte-Carlo simulations. In its original formulation, the LF2I machinery includes three modular procedures which, respectively, **(i)** estimate a likelihood-based test statistic via odds ratios, **(ii)** estimate critical values $C_{\theta,\alpha}$ via quantile regression, and **(iii)** check that the constructed confidence sets achieve the desired coverage level for all $\theta \in \Theta$. Each module is based on an independent dataset sampled from a high-fidelity simulator F_θ. `Waldo` replaces **(i)** and instead uses posteriors or predictions to compute $\tau_{\texttt{Waldo}}$ as in Equation (3.4). We break down the construction of a confidence set (including diagnostics) in the following steps, as outlined in Figure 3.1 and Algorithm 3.1:

(i) Estimate the test statistic via prediction algorithms or neural posterior estimators. Use the dataset $\mathcal{T} = \{(\theta_i, \mathcal{D}_i)\}_{i=1}^{B}$, where θ can be drawn from any prior distribution π_θ, to estimate $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$. This can be done by choosing between two methods: if using a prediction algorithm, we can leverage the fact that they approximate the conditional mean of the outcome variable given the inputs $\mathcal{D}$, when minimizing the squared error loss (lines 4-6 in Algorithm 3.1). Conversely, if using modern neural posterior estimators (such as normalizing flows (Papamakarios et al., 2021)), we can approximate $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$ via Monte Carlo sampling from the estimated posterior distribution (lines 16-18 in Algorithm 3.1);

(ii) Estimate critical values via quantile regression. Estimate $C_{\theta,\alpha} \coloneqq F^{-1}_{\widehat{\tau}_{\texttt{Waldo}}}(1-\alpha \mid \theta)$ by learning the conditional $(1-\alpha)$-quantile of $\widehat{\tau}_{\texttt{Waldo}}(\mathcal{D};\theta)$ using quantile regression over a calibration set $\mathcal{T}' = \{(\theta_i, \mathcal{D}_i)\}_{i=1}^{B'}$, where θ is drawn from a distribution with density

Algorithm 3.1 Confidence set for θ via `Waldo`

Input: Datasets $\mathcal{T}, \mathcal{T}', \mathcal{T}''$; observed sample D; prediction algorithm or posterior estimator; quantile regressor; grid of parameter values $\Theta_{N_{\text{grid}}}$; desired coverage level $1-\alpha$

Output: Confidence set $\widehat{\mathcal{R}}D$

// **Estimate building blocks of test statistic**
Draw $\mathcal{T} = \{(\theta_i, \mathcal{D}_i\}_{i=1}^{B}$
if prediction algorithm **then**
    Estimate $\mathbb{E}[\theta \mid \mathcal{D}]$ on $\mathcal{T}$ under squared error loss
    Compute $\{z_i := (\theta_i - \mathbb{E}[\theta \mid \mathcal{D}_i])^2\}_{i=1}^{B}$
    Estimate $\mathbb{V}[\theta \mid \mathcal{D}] = \mathbb{E}[z \mid \mathcal{D}]$ under squared error loss
else if posterior estimator **then**
    Estimate posterior distribution $p(\theta \mid \mathcal{D})$ on $\mathcal{T}$

// **Estimate critical values**
Simulate $\mathcal{T}' = \{(\theta_i, \mathcal{D}_i)\}_{i=1}^{B'}$
if prediction algorithm **then**
    Predict $\{\widehat{\mathbb{E}}[\theta \mid \mathcal{D}_i], \widehat{\mathbb{V}}[\theta \mid \mathcal{D}_i]\}_{i=1}^{B'}$
else if posterior estimator **then**
    for each $\mathcal{D}$ that appears in $\mathcal{T}'$ **do**
        Draw $N_{\hat{p}}$ samples from $\hat{p}(\theta \mid \mathcal{D})$
        $\widehat{\mathbb{E}}[\theta \mid \mathcal{D}] \approx \frac{\sum_i \theta_i}{N_{\hat{p}}}$
        $\widehat{\mathbb{V}}[\theta \mid \mathcal{D}] \approx \frac{\sum_i (\theta_i - \widehat{\mathbb{E}}[\theta|\mathcal{D}])(\theta_i - \widehat{\mathbb{E}}[\theta|\mathcal{D}])^T}{N_{\hat{p}} - 1}$
Compute $\{\hat{\tau}^{\texttt{Waldo}}(\mathcal{D}_i; \theta_i)\}_{i=1}^{B'}$
Estimate critical values $C_{\theta,\alpha}$ via quantile regression of $\hat{\tau}_{\texttt{Waldo}}(\mathcal{D}; \theta)$ on θ

// **Neyman inversion**
if prediction algorithm **then**
    Predict $\widehat{\mathbb{E}}[\theta \mid D]$ and $\widehat{\mathbb{V}}[\theta \mid D]$
else if posterior estimator **then**
    Draw $N_{\hat{p}}$ samples from $\hat{p}(\theta \mid D)$
    $\widehat{\mathbb{E}}[\theta \mid D] \approx \frac{\sum_i \theta_i}{N_{\hat{p}}}$
    $\widehat{\mathbb{V}}[\theta \mid D] \approx \frac{\sum_i (\theta_i - \widehat{\mathbb{E}}[\theta|D])(\theta_i - \widehat{\mathbb{E}}[\theta|D])^T}{N_{\hat{p}} - 1}$
Predict $\widehat{C}_{\theta_0,\alpha}\ \forall \theta_0 \in \Theta_{\text{grid}}$
Initialize $\widehat{\mathcal{R}}(D) \leftarrow \varnothing$
for $\theta_0 \in \Theta_{\text{grid}}$ **do**
    if $\hat{\tau}_{\texttt{Waldo}}(D; \theta_0) \leqslant \widehat{C}_{\theta_0;\alpha}$ **then**
        $\widehat{\mathcal{R}}(D) \leftarrow \widehat{\mathcal{R}}(D) \cup \{\theta_0\}$
return confidence set $\widehat{\mathcal{R}}(D)$

$r_\theta > 0, \forall \theta \in \Theta$ to allow for effective calibration across the entire parameter space;

(i) + (ii) Neyman inversion. Once D is observed, evaluate $\widehat{\tau}_{\texttt{Waldo}}(D; \theta_0)$ and $\widehat{C}_{\theta_0;\alpha}$ over a fine grid of parameters $\theta_0 \in \Theta$, and retain all θ_0 for which the corresponding test does not reject the null:

$$\widehat{\mathcal{R}}(D) = \{\theta_0 \in \Theta : \widehat{\tau}_{\texttt{Waldo}}(D; \theta_0) \leqslant \widehat{C}_{\theta_0,\alpha}\}. \tag{3.5}$$

As we showed in Section 2.4, step ***(ii)*** leads to valid level-α hypothesis tests as long as the quantile regressor is well estimated, which then implies that $\widehat{\mathcal{R}}(D)$ satisfies conditional coverage (Equation (3.1)) at level $1 - \alpha$, regardless of the true value of θ and of the size n of the observed sample D;

(iii) Coverage diagnostics. To check that the constructed confidence sets indeed achieve the desired level of conditional coverage, we leverage the diagnostics procedure introduced in Dalmasso* et al. (2024) and covered in Section 2.3.4. In detail: simulate a set $\mathcal{T}'' = \{(\theta_i, \mathcal{D}_i)\}_{i=1}^{B''}$ and construct a confidence region for each $\mathcal{D}_i \in \mathcal{T}''$. Then model $\mathbb{1}\{\theta_i \in \mathcal{R}(\mathcal{D}_i)\}$ as a function of θ_i adopting a suitable probabilistic classification method. By definition, this will estimate $\mathbb{E}[\mathbb{1}\{\theta \in \mathcal{R}(\mathcal{D}\}|\theta] = \mathbb{P}[\theta \in \mathcal{R}(\mathcal{D})|\theta]$ across the whole parameter space. Note that this module is completely *independent* from **(i)** and **(ii)**. As such, it can be used to to check the empirical conditional coverage of any uncertainty estimate, as illustrated in Section 3.3.4 for Neyman confidence sets where critical values are estimated via Monte Carlo sampling, in Section 3.4.1 for posterior credible regions, and in Section 3.4.2 for prediction sets from the output of a CNN.

3.3.3 Statistical Properties: Coverage and Power

We now show that the coverage guarantees of `Waldo` are independent from the prior distribution, which can also be chosen to increase power. We do so through univariate Gaussian examples with analytically computable solutions. Since $d = 1$, we use simple prediction algorithms to estimate $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$. See Appendix B.2.1 for details.

Property I: `Waldo` guarantees conditional coverage across Θ, regardless of the specified prior. Scientists sometimes have domain-specific knowledge that can guide inference through the elicitation of a prior distribution over the parameters of interest. The goal is to introduce a bias to help quantifying the uncertainty, but if the prior happens to be at odds with the data, then this bias can be harmful and cause posteriors to be overconfident and smaller than they should be (Hermans et al., 2021). Ideally, we would want the coverage guarantees of any estimated parameter region to be preserved under this bias. In this example, we assume $\theta \sim \mathcal{N}(0, 2),\ \mathcal{D} \mid \theta \sim \mathcal{N}(\theta, 1)$. As Figure 3.2 shows, confidence sets for θ (left panel) constructed through Neyman inversion of a series of Wald tests guarantee the correct conditional coverage (right panel), since Wald tests are only influenced by the likelihood. Conversely, prediction sets ($\mathbb{E}[\theta \mid \mathcal{D}] \pm z_{\alpha/2}\sqrt{\mathbb{V}[\theta \mid \mathcal{D}]}$) are influenced by the prior through the bias induced in the point predictions, which increases with the distance from the prior mean and results in strong under-coverage. `Waldo` exploits the same inputs of prediction sets ($\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$), but corrects this problem by calibrating the critical

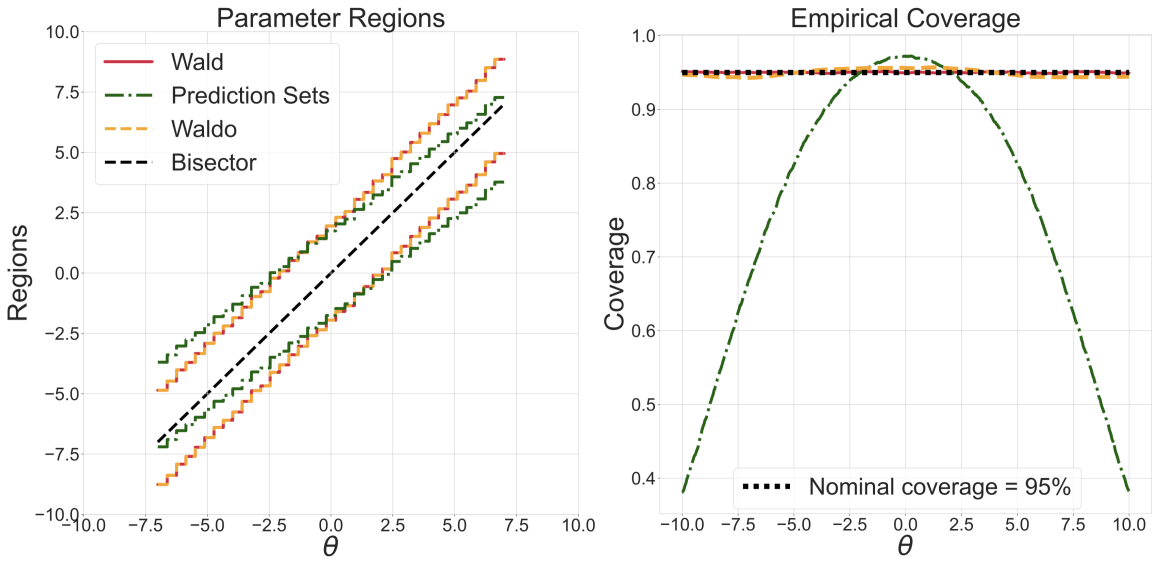


Figure 3.2: **Property I: `Waldo` guarantees conditional coverage across Θ, regardless of the specified prior.** Prior: $\theta \sim \mathcal{N}(0, 2)$. Likelihood: $\mathcal{D} \mid \theta \sim \mathcal{N}(\theta, 1)$. **Left:** median of upper/lower bounds of constructed parameter regions. **Right:** empirical coverage computed numerically using 100,000 samples for each θ over a fine grid in Θ (i.e., not using coverage diagnostics).

values via quantile regression, hence guaranteeing conditional coverage. Note that we only use a single observation ($n = 1$) for each confidence set.

Property II: `Waldo` exploits prior information and achieves higher statistical power. When the prior is correctly specified, we would like to leverage the induced bias to increase the power of the inverted tests and produce tighter constraints on the parameters, while retaining conditional coverage. Here we simulate data from a unique "true" Gaussian likelihood $\mathcal{D} \mid \theta \sim \mathcal{N}(\theta = 40, 1)$, and investigate the effect that the informativeness of the prior has on the power of the resulting tests. As Figure 3.3 shows, `Waldo` and Wald coincide when the prior is uninformative ($\theta \sim \mathcal{U}(35, 45)$; left panel), but the former has higher power when the prior is instead correctly specified ($\theta \sim \mathcal{N}(40, 1)$; right panel), thereby leading to smaller confidence sets. In Chapter 4, we will make this property more rigorous by proving that a class of posterior-based confidence sets is provably optimal (i.e., as precise as possible) with respect to the prior distribution.

3.3.4 Computational Properties

Scaling with high-dimensional parameters. As mentioned in Section 3.3.2, `Waldo` exploits a dataset sampled over Θ to estimate critical values via quantile regression and guarantee coverage across the whole parameter space[1]. While this might seem a daunting requirement, the only alternative to guarantee conditional coverage is to resort to Monte

[1]Technically, we only need to sample from a distribution that places mass on all Θ.

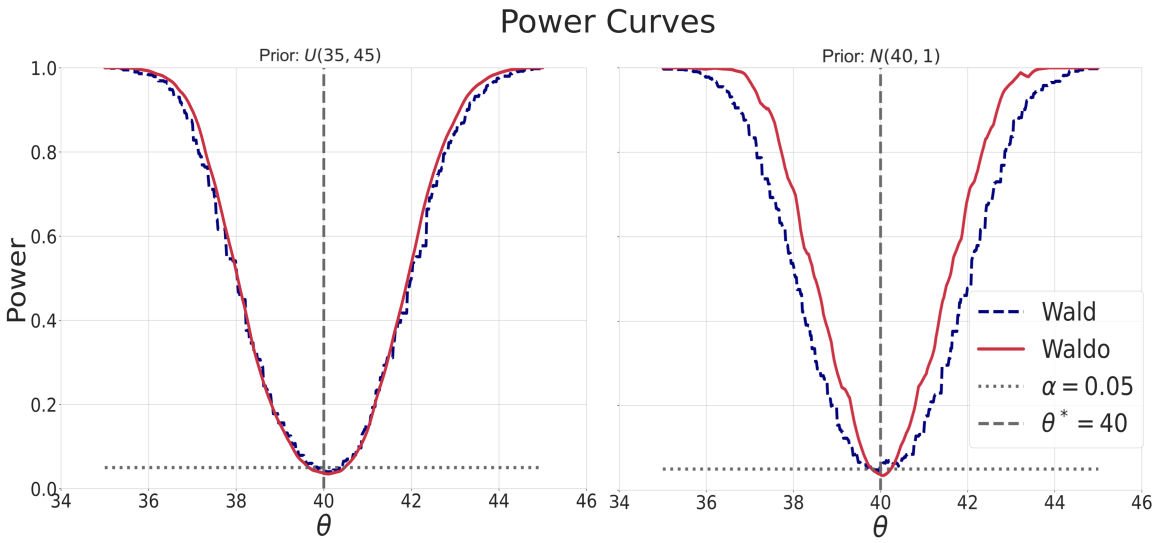


Figure 3.3: **Property II: `Waldo` exploits prior information and achieves higher power.** Power curves computed by recording the number of times a wrong value of θ is correctly outside the confidence set over 1,000 repetitions. Likelihood: $\mathcal{D} \sim \mathcal{N}(40, 1)$. **Left:** Wald and `Waldo` are equivalent when $\theta \sim \mathcal{U}(35, 45)$. **Right:** `Waldo` has higher power when $\theta \sim \mathcal{N}(40, 1)$.

Carlo approaches that sample many times *at each* $\theta \in \Theta$. As Figure 3.4 shows, `Waldo` requires several orders of magnitude *less* simulations to achieve the correct calibration. This is true already when $d = 1$, and is even more evident when $d = 10$.

Quality of models. `Waldo` relies on two estimation procedures (**(i)** and **(ii)** below) to construct the confidence set itself. The accuracy of the results relies on the estimation quality of these models and on the number of simulations B and B' that are available. In addition, there is a diagnostics procedure **(iii)** to estimate the conditional coverage of the final confidence sets, as a separate check that Equation (3.1) indeed holds.

(i) Test statistic. The quality of prediction algorithms and posterior estimators is positively correlated with the power of the resulting tests. As the precision in the estimates of $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$ decreases, the variance of the test statistics increases, which implies more conservative critical values and larger confidence regions. A good prior distribution will clearly help in achieving more precise estimates in regions of interest in the parameter space.

(ii) Critical values. As we proved in Section 2.4, conditional coverage is achieved as long as the quantile regressor is well estimated. In practice, we observe that little hyper-parameter optimization is needed and that the number of simulations required to achieve well-calibrated critical values is usually a small fraction of those needed for the test statistic.

(iii) Diagnostics. The quality of the probabilistic classifier used to check the empirical

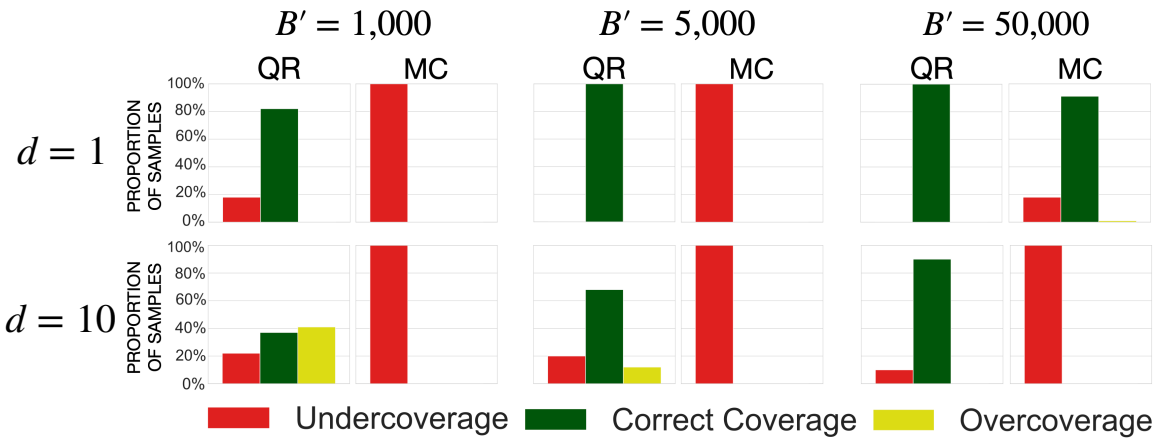


Figure 3.4: **Quantile regression (QR) is orders of magnitude more efficient than Monte Carlo (MC) in terms of the number of simulations B' required to achieve correct coverage.** Each panel shows the fraction of samples (out of 1,000 total) for which the selected method to estimate critical values achieves approximately correct coverage ($\mathbb{P}(\theta \in \mathcal{R}(\mathcal{D}) \mid \theta) \in [0.95 \pm 0.03]$). Prior: $\theta \sim \mathcal{N}(0, 0.1 \cdot I)$. Likelihood: $\mathcal{D} \mid \theta \sim \mathcal{N}(\theta, 0.1 \cdot I)$. In both cases, we used normalizing flows to estimate the posterior.

coverage probability affects only the reliability of the diagnostics. Note that this module is completely independent of the others, and we can check its quality by inspecting the cross-entropy loss, and the standard errors and confidence bands on the estimates that common statistical packages provide (e.g., `MGCV` (Wood, 2015) in `R`).

3.4 Experiments

We assess the performance of `Waldo` on two challenging experiments. In the first example (Section 3.4.1), we show how to use a posterior distribution estimated via normalizing flows to compute valid confidence regions, and how prior information can improve precision. The second example (Section 3.4.2) tackles a complex particle energy reconstruction problem in high-energy physics: we leverage predictions from a custom convolutional neural network (CNN) to construct confidence intervals with correct coverage and high power.

3.4.1 Confidence Sets from Neural Posteriors

This inference task was introduced in Sisson et al. (2007) and has become a standard benchmark in the SBI literature (Clarté et al., 2021; Toni et al., 2009; Simola et al., 2021; Lueckmann et al., 2021). It consists of estimating the (common) mean of the components of a two-dimensional Gaussian mixture, with one component having much broader covariance: $\mathcal{D} \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, I) + \frac{1}{2}\mathcal{N}(\theta, 0.01 \cdot I)$, where $\theta \in \mathbb{R}^2$ and $n = 1$[2]. We estimate $p(\theta \mid \mathcal{D})$ using a Neural Posterior Estimator (NPE) based on masked autoregressive flows (Papamakarios et al., 2017) as implemented by the `nflows` library (Durkan et al., 2020a) through the `SBI`

[2] `Waldo` works for an observed sample of any size, but we had to use $n = 1$ because the `SBI` Python library we used to estimate the posterior does not yet support larger sample sizes for NPE.

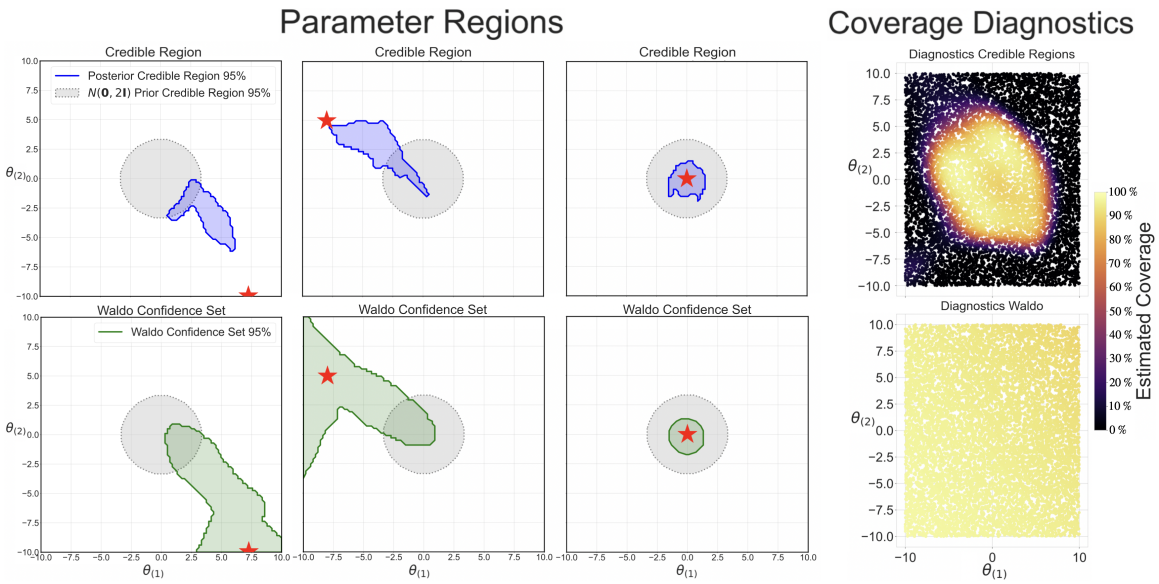


Figure 3.5: `Waldo` **converts posterior distributions into confidence regions with correct conditional coverage and high power**. **Left Panel - Top:** Examples of 95% credible regions (blue) from posteriors estimated with normalizing flows and a Gaussian $\mathcal{N}(0, 2 \cdot I)$ prior (gray) for different values of the true unknown parameter θ^* (red star). **Right Panel - Top:** Credible regions have conditional coverage close to the nominal level only in a neighborhood of the prior, and severely undercover everywhere else. **Left Panel - Bottom:** Corresponding 95% `Waldo` confidence sets (green), derived from the same posterior estimates used for the top row. **Right Panel - Bottom:** Conditional coverage for `Waldo` confidence sets achieves the nominal 1-α level everywhere, where $\alpha = 0.05$.

package (Tejero-Cantero et al., 2020), and report results obtained with two different priors: $\theta \sim \mathcal{N}(0, 2 \cdot I)$ and $\theta \sim \mathcal{U}([-10, 10]^2)$ (the latter in Appendix B.1.2). We estimate the critical values with a 2-layer feedforward neural network minimizing the quantile loss. Simulated datasets used for training are of the following sizes: $B = 100{,}000$, $B' = 30{,}000$ when using a Gaussian prior. Conditional mean and variance were approximated with 50,000 Monte Carlo samples from the learned neural posterior.

The first four panels on the left of Figure 3.5 show examples of 95% credible regions (top) and `Waldo` confidence sets (bottom) obtained from the same posterior distribution, when the true parameter is far from the prior. If the data is at odds with the prior, then the induced bias leads to credible regions that severely undercover across the parameter space, as it is shown at the top of the rightmost panel, where the coverage probability for credible regions reaches values as low as 0-10%. `Waldo` can correct for this bias and output larger confidence sets which account for the added uncertainty, thereby leading to correct conditional coverage everywhere (bottom of rightmost panel). This is the same behavior seen in the first example of Section 3.3.3, although for a more complex setting and for a posterior estimator.

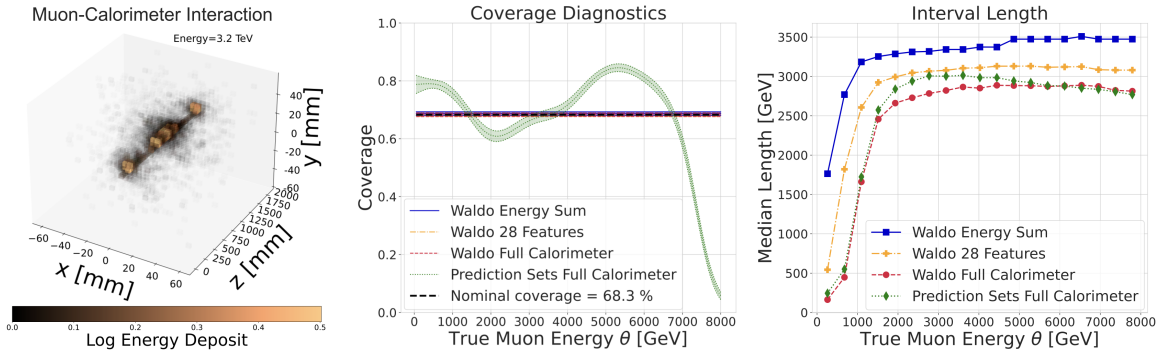


Figure 3.6: `Waldo` **guarantees the nominal coverage level, and yields smaller confidence intervals (more precise estimates of muon energy) with the higher-granularity ("full") calorimeter data. Left:** Energy deposited by a $\theta \approx 3.2$ TeV muon entering a calorimeter with $32 \times 32 \times 50$ cells. **Center:** `Waldo` (blue, orange, red in the right two panels) guarantees nominal coverage (68.3%), while 1σ prediction intervals (green) under- or over-cover in different regions of Θ. **Right:** Median lengths of constructed intervals: shorter intervals imply higher precision in the estimates. Prediction sets are on average wider than the corresponding confidence sets, using the same data.

Conversely, when the prior is consistent with the data (Figure 3.5, right two panels of "Parameter Regions"), `Waldo` is not overly conservative and leverages the additional information to tighten the constraints on the parameters, closely tracking the size of the posterior credible region. In Appendix B.1.2, we also show that, over many independent observations, the average size of `Waldo` confidence sets is indeed smaller when using an informative prior than when using a Uniform over Θ. These results closely mimic those seen in the second example of Section 3.3.3. In Chapter 4, we will make this property more rigorous by proving that a class of posterior-based confidence sets is provably optimal (i.e., as precise as possible) with respect to the prior distribution.

3.4.2 Confidence Sets for Muon Energies using CNN Predictions

We now discuss the performance of `Waldo` on an application of interest to fundamental research: estimating the energy of muons at a future particle collider. Muons are a heavier replica of electrons; they are produced in sub-nuclear reactions involving electroweak interactions. Muons are also excellent probes of new phenomena: in fact, their detection and measurement has been key to several crucial discoveries in the past decades, including the Higgs boson (Augustin et al., 1974; Herb et al., 1977; Collaboration et al., 1995; Aad et al., 2012b; Chatrchyan et al., 2012). Traditionally, the energy of a muon is determined from the curvature of its trajectory in a magnetic field, but at energies above a few TeV these methods breaks down as trajectories become indistinguishable from straight paths even within the strongest practically achievable fields. Searching for viable alternatives, it has been observed (Kieseler et al., 2022; Dorigo et al., 2022) that both the pattern and the magnitude of small radiative energy losses that muons withstand in traversing dense and finely segmented calorimeters can be used to infer the incident muon energy, leveraging the capacity of modern deep learning architectures. Nonetheless, the above work also clearly

showed that predictions of θ suffered from a strong bias, mainly due to the high nonlinearity of the response at very high energies. Motivated by this problem, we pose two questions: **(i)** Can we construct confidence sets with correct coverage of the true energy of muons using the information contained in the pattern and magnitude of radiative deposits in a dense calorimeter? **(ii)** Is it possible to extract additional information from finer segmentations of the calorimeter to allow for tighter constraints (i.e., smaller confidence sets with correct coverage) on muon energy estimates? Quantifying the latter would allow scientists to optimize their detector designs, since manufacturing very small calorimeter cells is expensive.

We collected 886,716 3D input "images" X and scalar muon energies θ obtained through `GEANT4` (Agostinelli et al., 2003), a high-fidelity stochastic simulator. See Figure 3.6 (left panel) for an illustration of one simulated X_i for a particular θ_i. The data are available at Kieseler et al. (2021). As the interest is on constraining muon energies as much as possible while guaranteeing conditional coverage, we use three versions of the same dataset with increasing dimensionality: a 1D input equal to the sum over all calorimeter cells with deposited energy $E > 0.1$ GeV, for each muon; 28 custom features extracted from the spatial and energy information of the calorimeter cells (see Kieseler et al. (2022)); and the full calorimeter measurements ($X_i \in \mathbb{R}^{51,200}$). For the first two datasets, we estimate $\mathbb{E}[\theta \mid \mathcal{D}]$ and $\mathbb{V}[\theta \mid \mathcal{D}]$ via gradient boosted trees as implemented in `XGBoost`(Chen and Guestrin, 2016). For the full calorimeter data, we rely on the CNN developed by Kieseler et al. (2022). We use quantile gradient boosted trees for quantile regression, as implemented in `scikit-learn` (Pedregosa et al., 2011).

Answering **(i)** affirmatively, Figure 3.6 (center) shows that confidence sets constructed with `Waldo` achieve exact conditional coverage (68.3%) regardless of the dataset used. The corresponding 1σ prediction intervals ($\mathbb{E}[\theta \mid \mathcal{D}] \pm \sqrt{\mathbb{V}[\theta \mid \mathcal{D}]}$) using full calorimeter data, instead, exhibit over- or under-coverage in different regions over Θ, which in the latter case means that prediction sets contain the true value with much lower probability than anticipated. As for question **(ii)**, we make two observations (see Figure 3.6; right panel): First, using the raw higher-dimensional energy deposits with `Waldo` allows to reduce the uncertainty around muon energies. Second, confidence sets constructed with `Waldo` are even shorter than the corresponding prediction intervals, while also guaranteeing conditional coverage.

3.5 Conclusions and Discussion

In this Chapter, we presented `Waldo`, a novel method to construct confidence sets with correct finite-n conditional coverage by leveraging prediction algorithms and posterior estimators for inverse problems. `Waldo` relies on a regression-based Neyman construction, which requires orders of magnitude fewer simulations than traditional Monte Carlo approaches to be well calibrated across the parameter space (see Section 3.3.4). Nonetheless, our method still needs a simulator that is both high-fidelity — to draw inferences that reflect the true data-generating process — and fast — to simulate sufficiently large training sets to accurately learn the key quantities of `Waldo`: the test statistics, the critical values, and the coverage diagnostics, as discussed in Section 3.3.4. `Waldo` disentangles the *coverage*

guarantees of the confidence region from the choice of the prior distribution. To increase *power*, one may be able to leverage domain-specific knowledge (see Sections 3.3.3 and 3.4.1), or take advantage of the internal structure of the simulator (Brehmer et al., 2020), with the guarantee that the confidence sets always contain the true parameter with the desired proability. One could also adaptively simulate more data in specific regions of interest in the parameter space. Active learning strategies, and a more formal treatment of the relation between power and priors, are promising areas for future studies.

Domain sciences, especially the physical sciences, routinely seek to constrain parameters of interest using both theoretical (or simulation) models and experimental data. `Waldo` provides reliable constraints that can be used to deduce trustworthy scientific conclusions when other uncertainty quantification methods are either unavailable, unreliable or inefficient.

4

Optimal Confidence Sets from Generative Models

4.1 Introduction and Problem Setting

Modern science relies on complex models of physical, biological and chemical phenomena. Yet, inferring internal parameters of a scientific model from observed data when the likelihood is intractable remains a major statistical challenge. This *inverse* inference task — where the likelihood is only implicitly encoded by a simulator or, alternatively, by labeled data[1] from auxiliary measurements — lies at the heart of many pressing questions across the natural sciences. In high-energy physics, for instance, the ATLAS and CMS experiments at the Large Hadron Collider have used the extremely complex outcome of proton-proton collisions to accurately measure parameters of the Standard Model (Glashow, 1959; Salam, 1959; Cabibbo, 1963; Weinberg, 1967), as well as to constrain its possible extensions, such as supersymmetry (see, e.g., ATLAS Collaboration, 2024). In astronomy, space telescopes like *Gaia* are used to infer stellar properties from low-resolution spectra, often with the aid of auxiliary high-resolution surveys Collaboration et al. (2016). In environmental science, complex Earth system models (e.g., UKESM; Sellar et al. 2019) are used to constrain the parameters of multiple physical processes for, e.g., land surface, atmosphere, ocean and ice sheet dynamics simultaneously.

Traditionally, statistical inference is based on evaluating likelihood functions that model the probability $p(X \mid \theta)$ of observing data X for different parameter values — i.e. instances — of θ. However, this approach becomes infeasible for modern scientific problems involving next-generation precision data. Likelihoods are often intractable, either due to the complexity of the data-generating process — frequently involving simulation-based processes or complicated integrals on large latent spaces — or due to prohibitive computational costs when evaluating them on massive data. To address these challenges, the scientific community has adopted a new class of inference methods, here referred to as *neural likelihood-free inference* (NLFI; see, e.g., Cranmer et al. (2020); Lueckmann et al. (2021)). The most popular NLFI approaches in astronomy, biology, and environmental sciences completely bypass likelihood computations, and instead rely on AI-based generative models — such as

[1] In what follows, we use the term labels to indicate the values of parameters (not necessarily discrete) associated with different objects; e.g., stellar labels for the properties of different stars (age, mass, etc.).

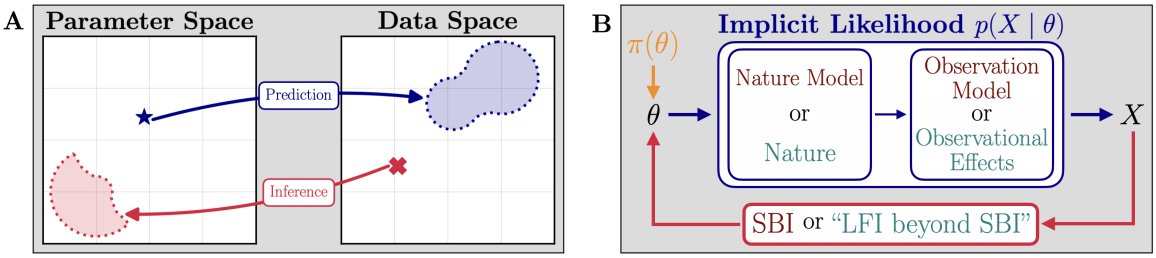


Figure 4.1: **The likelihood-free inference setting. Panel A**: With a forward model, we can make predictions on data X given parameters θ. The inverse problem is to infer the parameters θ of a model given observed data X. **Panel B**: In likelihood-free inference (LFI), the likelihood $p(X \mid \theta)$ is intractable. We consider two LFI scenarios, where the likelihood is implicitly encoded either by (i) a simulator (the inverse problem is then known as simulator-based inference or SBI; brown), or by (ii) labeled data from observational studies (we refer to the latter inverse problem as "LFI beyond SBI", green).

normalizing flows, diffusion models, and flow matching — to estimate posterior distributions $\pi(\theta|X)$.[2] These flexible posterior-based methods are attractive because they avoid the need for computationally tractable likelihoods, and they scale to massive data sets with the accuracy of traditional methods and several orders of magnitude faster inference; see, e.g., Wang et al. 2023 and Sainsbury-Dale et al. 2024 for examples with ultra-fast neural Bayesian inference with James Webb Space Telescope survey data and irregularly spaced remote sensing ocean data. However, despite the promise of neural inference methods, a fundamental question remains:

> *How can we make trustworthy inferences in inverse problems with posteriors learned via generative models?*

In forward problems, where the goal is to predict observable data X for fixed θ, generative models often perform well: their predictions can be validated against held-out data from the implicit likelihood, and the quality of the predictions can be directly assessed. *Inverse* problems are more challenging: internal parameters (such as, for example, the age and distance of a galaxy) are *inferred* from data (the output of the forward model), rather than *caused* by data. As we shall see, this distinction turns out to make a key difference in ensuring reliable scientific inference. To draw conclusions that adhere to the rigor of the scientific method, scientists need to reliably constrain unknown parameters given the data they can collect with valid measures of uncertainties: We say that a $100(1-\alpha)\%$ confidence region $R(X)$ for θ is valid if there is at least a $100(1-\alpha)\%$ chance that the region contains the true value of θ, no matter what that unknown value is. For these regions to be useful in parameter estimation, we also need them to have high constraining power; that is, to be small.

[2]The "posterior" distribution $\pi(\theta|X)$ can be interpreted as the uncertainty in our knowledge of θ *a posteriori* (after the fact) of observing data X. In this chapter, we will use the terms "posteriors" and "priors" beyond the traditional subjective Bayesian view (Gelman et al., 2013) to also apply to probabilities that can be indirectly determined by the observed population of physical entities, such as stars, galaxies, and so on.

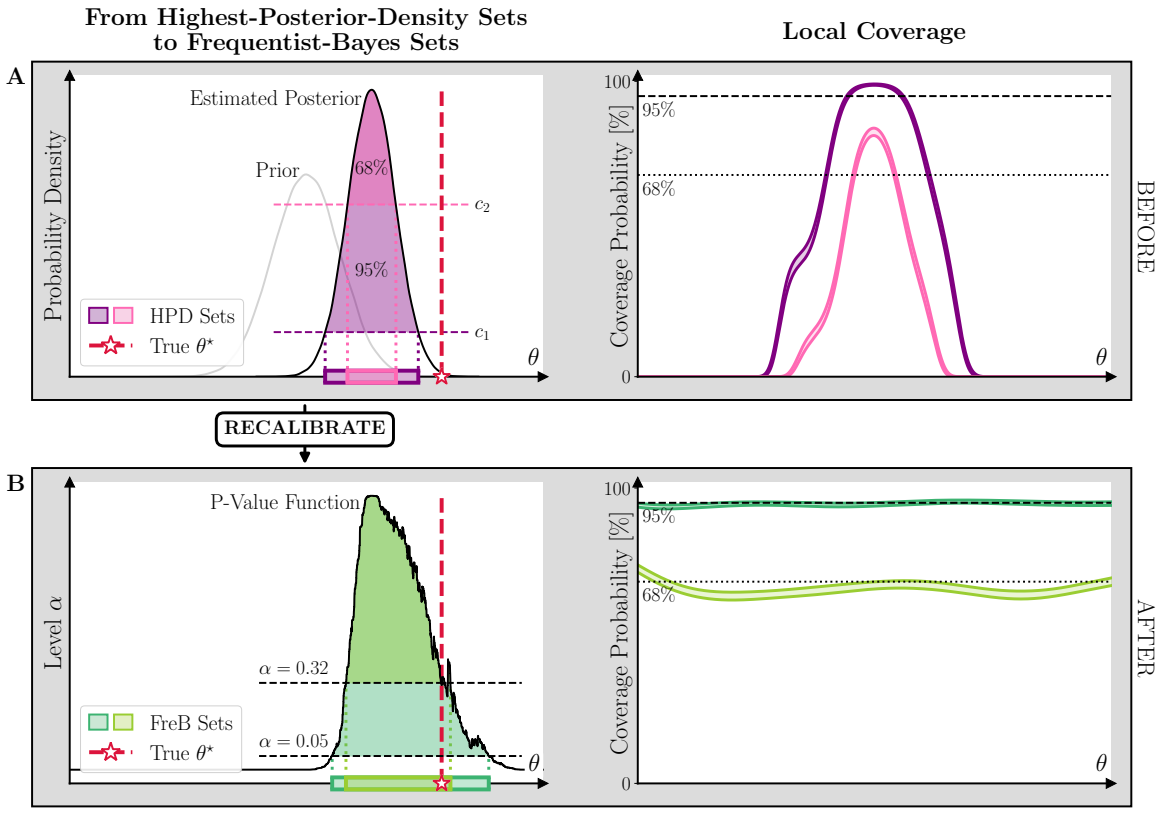


Figure 4.2: **Our proposed approach to valid scientific inference. Panel A:** *(Left)* The typical workflow for inferring parameters with neural density estimators is to first learn the posterior, $\widehat{\pi}(\theta|X)$, from train data. Then, for new observed data X_{obs}, one slices $\widehat{\pi}(\theta|X_{obs})$ to compute a highest-posterior density (HPD) set. The purple and pink intervals at the bottom depict 95% and 68% HPD sets, respectively, for an observation whose true parameter (indicated by a red star) lies in the tail of the prior $\pi(\theta)$. *(Right)* The actual chance (coverage probability, y-axis) that the two HPD sets contain the true parameter value can be far less than what the nominal coverage of 95% and 68%, respectively, suggest, for a wide range of different θ-values (x-axis). **Panel B**: *(Left)* Recalibration — our approach effectively transforms the posterior to a p-value function, which we then slice to obtain valid ("Frequentist-Bayes"; FreB) confidence sets. *(Right)* The actual chance (coverage probability, y-axis) that FreB sets contain the true parameter value is indeed close to the desired coverage probability for every instance of θ (x-axis).

Validity guarantees are critical in exact sciences. In high-energy physics, billions of dollars and thousands of person-years are devoted to ensuring that confidence statements about the existence of newly discovered subatomic particles or the values of fundamental constants are statistically valid. Yet, even with perfectly estimated posterior and ideal modeling conditions, standard *credible* intervals can fail in two key ways:

1. **Local coverage.** Even if a 90% credible region covers the true parameter 90% of the time on average (over the entire population of labeled objects), it provides no guarantees for individual instances. Scientists often need to constrain specific

parameters (e.g., determine the properties of a specific star, galaxy, event), and local coverage failures can lead to misleading conclusions.

2. **Robustness to label or prior probability shift.** In practice, labeled train data rarely reflect the parameter distribution of target data. Selection bias, observational limitations and different sampling strategies all lead to shifts between the training distribution $\pi(\theta)$ (effectively the "working" prior) and the distribution $p_{\text{target}}(\theta)$ of the target population. Similarly, different theories of natural phenomena — which in turn elicit (working) prior distributions — can lead to discrepancies, or tension, in the estimates of key physical parameters. These mismatches limit the usability of credible regions. Label/prior probability shift is challenging because a domain scientist rarely knows the true distribution $p_{\text{target}}(\theta)$, even when the likelihood $p(X|\theta)$ is perfectly known. She can base her sampling strategy and choice of $\pi(\theta)$ on existing knowledge of the underlying physical phenomenon, but this does not guarantee that $\pi(\theta)$ is close to $p_{\text{target}}(\theta)$, especially if the target represents a new physical source not yet observed.

These limitations highlight a core vulnerability of posterior-based inference: its lack of uncertainty estimates with frequentist coverage for individual objects, and its reliance on training priors that may not match the target data.

A framework for trustworthy scientific inference with biased training data. To overcome these limitations, we propose a new framework for constraining parameters that retains the advantages of neural posteriors while satisfying strict coverage guarantees. Our method transforms posteriors into statistically valid *Frequentist-Bayes* (FreB) confidence sets. These regions are calibrated to contain the true parameter with the desired probability regardless of the true (unknown) value of θ and across all levels α. The procedure works as follows: for a given posterior estimate $\widehat{\pi}(\theta|X)$, we learn a monotonic transformation of the posterior, using labeled calibration data (from a simulator or cross-matched catalogs). This transformation is effectively a p-value function: rather than slicing the posterior distribution at the required level to obtain a $(1-\alpha)100\%$ credible set of high posterior density, we slice the p-value function at the nominal level(s) α to construct confidence sets with desired local frequentist coverage. Crucially, the procedure is amortized — no additional training is needed at deployment, allowing for efficient inference for massive unlabeled data sets. This approach to constraining parameters offers several key advantages:

1. It provides **reliable inference with limited observations**, including the traditionally challenging case of just a single observation per object (that is, a sample size of $n = 1$).

2. It guarantees parameter constraints with **local validity** (that is, confidence sets with stated coverage probability for every parameter value), regardless of how the training data are collected, as long as the number of training examples is large enough and the (underlying) likelihood is the same for training and target data.

3. It achieves **optimal precision** (that is, small confidence sets) when prior knowledge aligns well with the target data.

Finally, we provide means to verify that the number of simulations in (2) is large enough to ensure that the results are trustworthy.

Outline and significance. The approach we propose — transforming neural posteriors to confidence sets with local frequentist guarantees — bridges simulation-based inference, classical statistics, and modern machine learning. It enables domain experts to perform principled inference using state-of-the-art generative models, even in settings with intractable likelihoods and label shift caused by selection or prior biases. By doing so, we provide a principled path towards reliable AI-driven scientific discovery, enabling advancements in fields such as astronomy, high-energy physics, biology, remote sensing, and beyond. In Section 4.2, we illustrate the practical value of the approach in three case studies from physical applications[3]:

I. Reconstructing gamma-ray showers from different astrophysical sources with ground-based detectors.

II. Inferring properties of Milky Way stars from spectra using different galaxy models.

III. Inferring stellar parameters with partially labeled data from cross-matched astronomical catalogs ("LFI beyond SBI").

Each case study addresses a specific statistical challenge (see Table 4.1 for more details). Finally, Section 4.3 and Figures 4.1 and 4.2 outline the practical implementation and assumptions behind our method, clarifying the details of our protocol for trustworthy scientific inference under intractable likelihoods.

4.2 Results

4.2.1 Case Study I: Reconstructing Gamma-Ray-Induced Air Showers with Ground-Based Detector Arrays

This case study illustrates how our framework enables the reliable identification and reconstruction of previously unknown physical sources — phenomena that would likely be missed or misinterpreted using standard generative models applied naively as inferential tools.

In astroparticle physics, high-energy gamma rays and cosmic rays yield crucial information on violent phenomena that take place in the cosmos. Unlike protons or light nuclei, which are deflected by cosmic magnetic fields, gamma rays travel in straight paths, allowing precise localization of their astrophysical sources. An important line of research is therefore the reconstruction of particle showers induced by such messengers in the atmosphere. Ground-based detector arrays (see the figure in the first row of Table 4.1) are commonly used to study these events by detecting the secondary particles reaching the ground (Chadwick, 2021). We consider the problem of estimating the parameter vector $\theta = (E, Z, A)$ — representing the energy (E), zenith angle (Z), and azimuthal angle (A) of the incoming gamma ray — from data X that include the identity (electrons, photons, etc.), count rate and density, and

[3]Code is avaliable at https://github.com/lee-group-cmu/vsi.

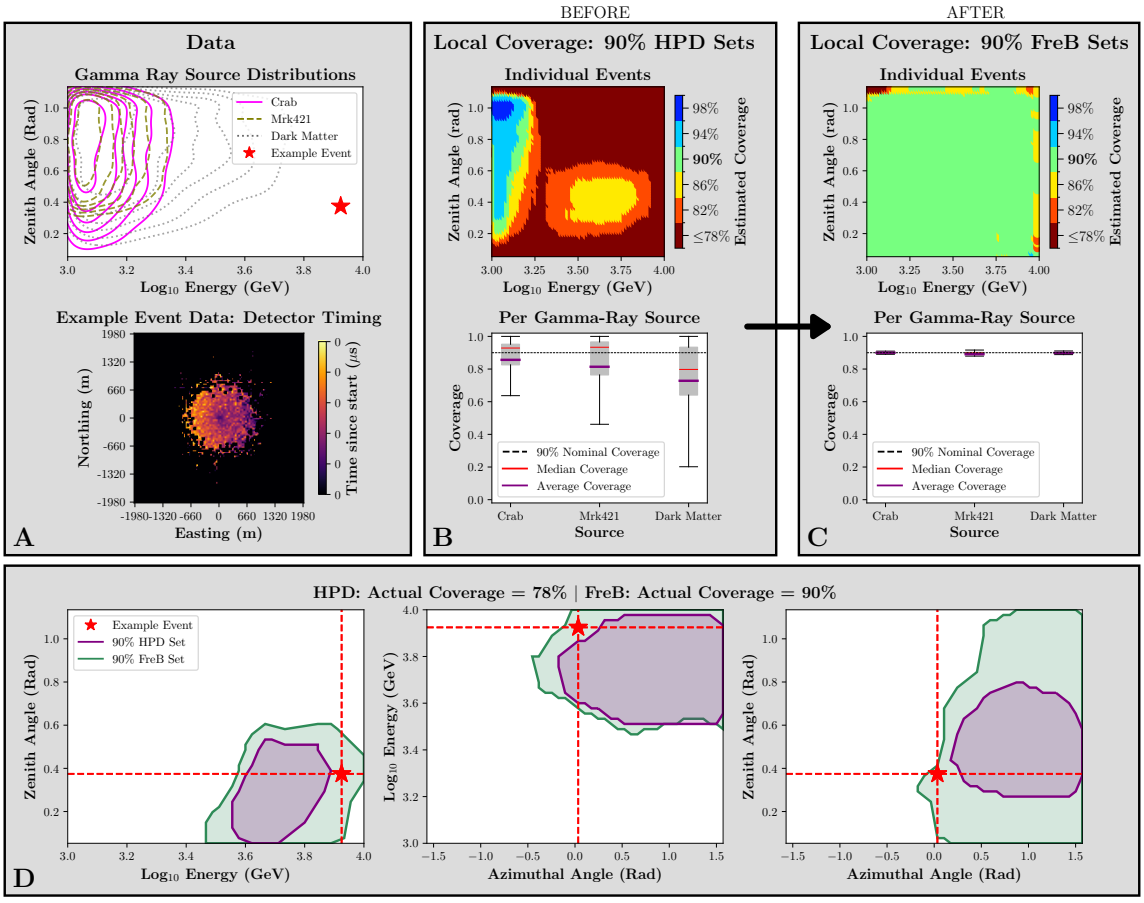


Figure 4.3: **Posterior-based methods lack local coverage guarantees and thus fail to reliably reconstruct gamma-ray showers from unfamiliar sources. Panel A:** *(Top)* Distribution of three gamma-ray sources in energy and zenith angle. An example gamma-ray event/shower at high energies is indicated by a red marker. *(Bottom)* Detector data for example event, showing arrival times at different locations. **Panel B:** *(Top)* Estimated local coverage of 90% HPD sets of individual events (averaged over azimuth) reveals undercoverage, especially at higher energies. *(Bottom)* Distribution of coverage across events from each gamma-ray source; coverage drops when training and target sources are different. **Panel C:** *(Top)* Local coverage of 90% FreB sets instead shows uniform validity across the parameter space. *(Bottom)* Coverage distribution per gamma-ray source confirms consistent validity regardless of source. **Panel D:** Comparison for a high-energy event from the Crab Nebula (for the same example event as in Panel A, Top): the 90% HPD set (purple) is overconfident and biased (actual coverage is 78%), while the 90% FreB set (green) provides valid and informative uncertainty.

various properties (e.g., energy, direction) of secondary particles detected on the ground. While the energy of gamma rays is typically distributed according to a target spectral shape (typically modeled as a power law or log-parabola) within the energy range of interest, the azimuth and zenith angles change over time, following the target's trajectory across the sky (see Appendix C.5 for more details). We assume the training set is drawn from sources resem-

bling the *Crab Nebula*[4], while test observations may originate from two benchmark sources: one mimicking *Markarian 421 (Mrk421)* — a well-studied blazar and among the brightest known gamma-ray sources (Abdo and Others, 2011) — and another resembling a potential *Dark Matter* signal, such as that expected from dark matter annihilation near the Galactic Center (Doro et al., 2024; Cirelli et al., 2024). All events are simulated using `Corsika` (Heck et al., 1998) with an idealized detector that perfectly records all particles on the ground.

We estimate the posterior distribution $\pi(\theta|X)$ via flow matching (Wildberger et al., 2024; Lipman et al., 2022) and construct HPD sets and FreB sets, both at the 90% confidence level. As detailed in Figure 4.3, we observe the following:

- HPD sets can mis-characterize unfamiliar gamma-ray sources due to the lack of local coverage and are only approximately valid when the true source parameters are similar to the *Crab Nebula*, from which the training set was constructed (Figure 4.3, Panel B). On the other hand, FreB sets ensure validity across all astrophysical sources for each value of the parameters (Figure 4.3, Panel C).
- FreB sets enable reliable reconstruction of gamma rays from unknown astrophysical sources by correctly quantifying the uncertainty around the truth. On the other hand, HPD sets tend to be biased and over-confident for (unknown) parameter values that were under-represented in the training set (Figure 4.3, Panel D).

4.2.2 Case Study II: Inferring Properties of Milky Way Stars in Simulation-Based Inference

Our framework resolves the paradox of conflicting scientific conclusions caused by differing models of nature. In this case study, two competing descriptions of our Milky Way galaxy are shown to be at odds when tasked with assigning labels to a newly discovered stellar object along the $(\ell, b) = (70^\circ, 30^\circ)$ line of sight. Inferring the properties of stars like this one from observational data is of crucial importance — in the current era of massive surveys equipped with next-generation instrumentation, notable discoveries are made regularly (Koposov et al., 2024), and they help drive our understanding of the structure and evolution of the Milky Way and the universe beyond.

We compare two galactic models, each reflecting different beliefs about the Milky Way:

- **Model H** asserts that the metallicity range on observed stellar objects from the Milky Way's halo underrepresents the true diversity of halo stars.
- **Model D** diminishes the contribution of the halo, instead emphasizing objects typically found within the galactic disk.

These models differ notably in their implied age-metallicity relationships, as depicted in Figure 4.4, Panel A, Left. Consequently, each model induces a distinct prior distribution on key stellar parameters θ. These include the gravitational constant, g; effective temperature,

[4]The Crab Nebula is a pulsar-wind nebula emitting the brightest and stable TeV signal in the northern hemisphere sky, for the past 970 years.

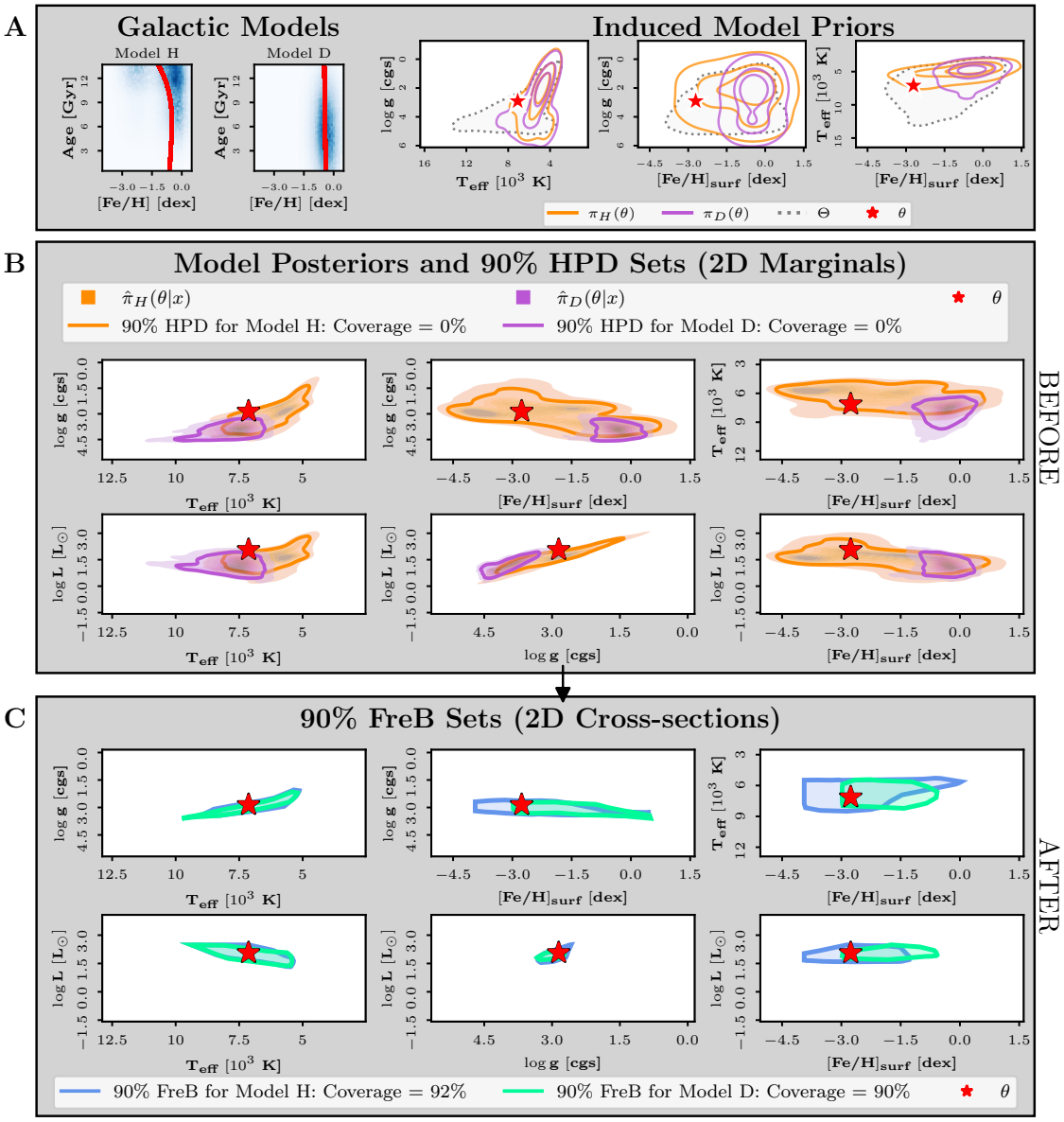


Figure 4.4: **FreB resolves tension between differing galactic models. Panel A:** *(Left)* The age-metallicity relationships implied by two Galactic models. The red curves indicate conditional means of metallicity given age. **Panel A:** *(Right)* Surface-level priors induced by the galactic models along line of sight $(70°, 30°)$. Log gravitational constant $(\log g)$, effective temperature (T_{eff}), and surface metallicity $([Fe/H]_{\text{surf}})$ are shown. The true label for a typical object is marked in red, unseen at inference time. **Panel B:** Tension between Models H and D's posteriors at $X \sim p(X|\theta)$. Solid contours for each show 90% credible regions of high posterior density, marginalized. The HPD regions feature 0% conditional coverage. **Panel C:** 90% FreB sets for θ for Models H and D. Each subplot shows cross-sections of the FreB sets at the true label. Conditional coverage for each FreB set is close to the nominal 90% level.

$T_{\text{effective}}$; surface metallicity, $[Fe/H]_{\text{surface}}$; and luminosity L.[5] The priors, seen in Panel A, Right, are derived according to stellar evolution theories using `brutus` (Speagle et al., 2025), an open-source Python package tailored for fast stellar characterization. We then simulate measurements $X \sim p(X \mid \theta)$ that replicate spectral observations from the 2MASS (Skrutskie et al., 2006) and PS (Bolden and Kervin, 2010) surveys.

A posterior-based approach produces contradictory results between our models. We estimate neural posteriors with Masked Autoregressive Flows (Greenberg et al., 2019; Tejero-Cantero et al., 2020) and show their stark disagreement — with each other, and with the truth — in Figure 4.4, Panel B. For example, under Model D, $\hat{\pi}_D(\theta \mid X)$ significantly overestimates $[Fe/H]_{\text{surface}}$ due to its metal-rich prior. Even Model H's posterior fails diagnostic tests, never covering all parameters at once (Panel B legend). In contrast, our FreB sets resolve this paradox by reconciling discrepancies between models while ensuring validity for any θ. Figure 4.4, Panel C displays cross-sections of the FreB sets which show simultaneous coverage for all parameters. Moreover, their compactness demonstrates their superior constraining power, particularly when some parameters can be independently constrained. Appendix C.3 provides further insights into FreB sets' statistical power when good prior information is available.

4.2.3 Case Study III: Inferring Stellar Parameters from Cross-Matched Astronomical Catalogs ("LFI Beyond SBI")

Our framework can handle observational studies with selection bias in the labels. In this section, we illustrate that as long as we have some labeled examples that sample the underlying likelihood for different parameter values, then we can achieve approximately valid confidence sets. In addition, with training data that are sampled with the same distribution as the unlabeled data — which may not be achievable in practice — the constraining power would increase.

Selection bias is a prevalent issue in astronomical surveys, as observations are often made deliberately and are not collected uniformly or randomly (Wang et al., 2023; Tak et al., 2024). Such intentional data collection inherently introduces biases: the training distribution, denoted $\pi(\theta)$, often deviates significantly from the underlying true distribution of parameters, $p_{\text{target}}(\theta)$. This selection bias is compounded when astronomers cross-match two or more survey catalogs to obtain high-quality and multi-wavelength data for making inferences about physical parameters, which are then used in down-stream astrophysical studies (Laroche and Speagle, 2024). As large-scale surveys like APOGEE (Majewski et al., 2017) and Gaia(Gaia Collaboration et al., 2023), and soon LSST (Ivezić et al., 2019), collect massive amounts of spectroscopic data, it is becoming increasingly important to develop and deploy estimation methods that are both scalable and trustworthy. For example, estimates of stellar labels — stellar properties (e.g., $\log g$ and T_{eff}) and elemental abundances (e.g., Fe/H) — are used in studies aimed at answering fundamental questions in astrophysics, from modeling stellar evolution (Minchev et al., 2018) and galaxy formation (Lagarde et al.,

[5] Refer to Table C.2 in Appendix C.6 for the true values of these parameters as well as the values of some intrinsic properties of the simulated star.

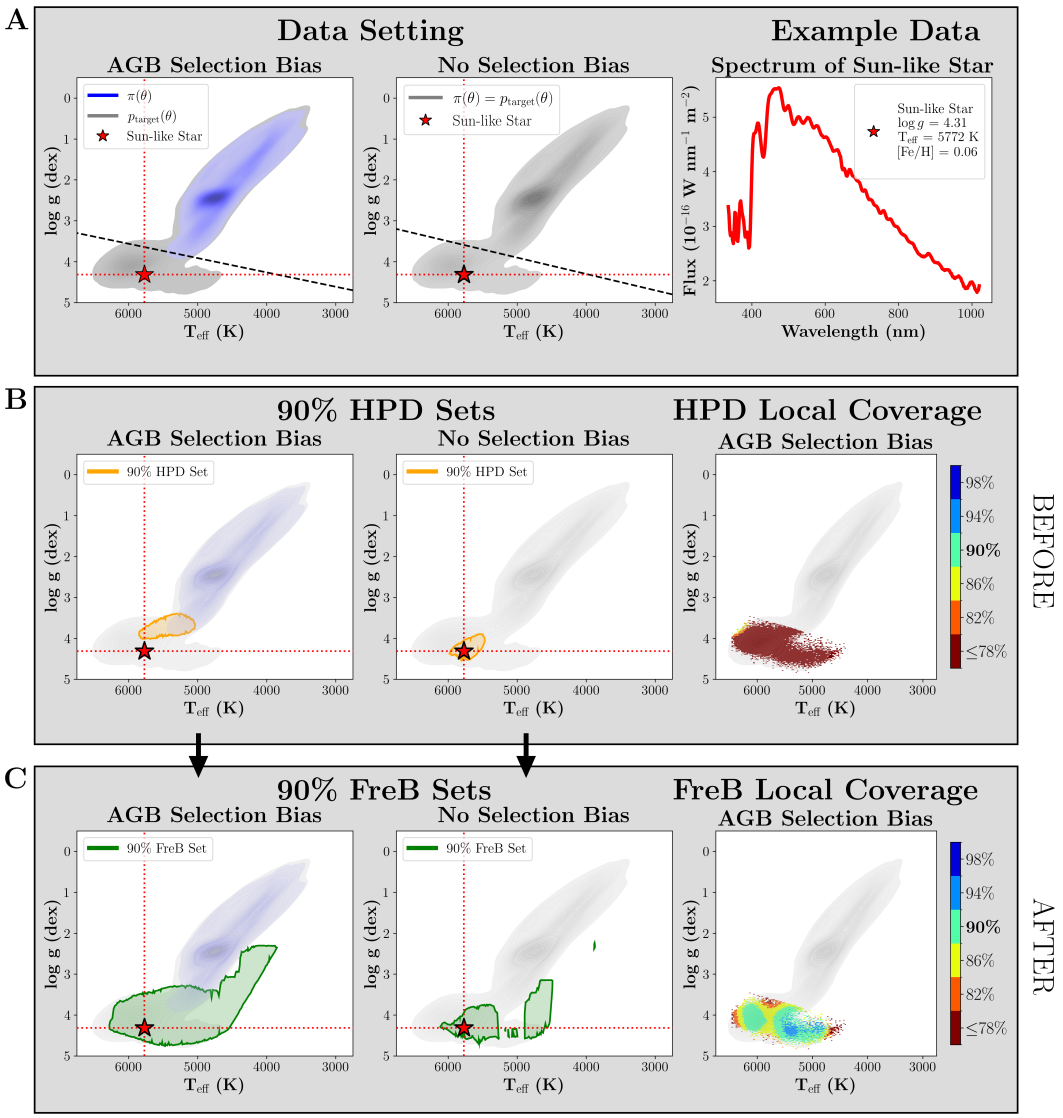


Figure 4.5: **FreB is robust to label bias in observational studies. Panel A:** Kiel diagrams displaying the training distribution of stellar gravities $\log g$ against the corresponding effective temperatures T_{eff}) for two data settings, where the labeled data are biased towards the asymptotic giant branch stars (*left*, "AGB Label Bias"), and where the labeled and unlabeled target data have the same distribution (*center*, "No Label Bias"). *(Right)* An example spectrum for a Sun-like star, for which the true label marked in red is unknown. **Panel B:** *(Left)* 90% HPD sets under the two selection settings, with the HPD set under the AGB selection bias not including the true label (red). *(Right)* Local coverage plot of 90% HPD sets in the held-out main sequence (MS) parameter space, showing under-coverage for all labels. **Panel C:** *(Left)* 90% FreB sets under the two selection settings, with the FreB set under both settings covering the true (red) label, but with higher constraining power with well-aligned training data. *(Right)* Local coverage plot of 90% FreB sets in the held-out main sequence (MS) parameter space, showing nearly nominal coverage for all labels.

2021) to characterizing stellar winds to understand the mass loss of stars (Carpenter et al., 1999).

Using stars from a Gaia/APOGEE cross-match, we estimate a parameter $\theta = (\log g,\ T_{\text{eff}},\ Fe/H)$ of stellar labels from data X consisting of 110 Gaia BP/RP spectra coefficients. We perform this estimation task in two settings:

- **No label bias**: The prior $\pi(\theta)$ aligns well with the true distribution $p_{\text{target}}(\theta)$.
- **Label bias**: The prior $\pi(\theta)$ is biased, specifically skewed towards AGB stars, differing significantly from $p_{\text{target}}(\theta)$, which predominantly consists of MS stars.

We estimate the posterior distribution $\pi(\theta \mid X)$ with Masked Autoregressive Flows (Greenberg et al., 2019; Tejero-Cantero et al., 2020) and construct HPD and FreB sets at the 90% confidence level in both selection bias settings (see Appendix C.7 for details). As detailed in Figure 4.5, we observe the following:

- FreB sets ensure valid local coverage even under selection bias, unlike traditional highest posterior density sets, which are overconfident and exhibit poor coverage under these conditions.
- FreB sets provide high constraining power when the training distribution closely matches the true target distribution.

4.3 Methods

This chapter proposes a new framework for reliable scientific inference under intractable likelihoods, which bridges classical (frequentist) statistics (Neyman, 1935b, 1937b) with state-of-the-art generative models and Bayesian inference.

4.3.1 Experimental Set-Up

Our key assumption is that labeled and unlabeled data stem from the same data-generating process and hence the *same likelihood* $p(X \mid \theta)$. However, the data could be sampled differently over the parameter space to reflect prior, observational, or experimental biases. The labeled data is then further categorized into a "universal set" (defined over the entire parameter space of interest) and a "train set" (which may be different for each use case). More specifically, we assume there are three distinct sets from the same likelihood:

- a labeled "universal" set,

$$\mathcal{T}_{\text{univ}} = \{(\theta'_1, X'_1) \ldots (\theta'_{B'}, X'_{B'})\} \sim r(\theta)p(X|\theta),$$

  where the reference distribution $r(\theta)$ covers the entire parameter space Θ of interest (this set could for example represent broad data from different sources);

- a labeled train set for learning the neural density estimator $\widehat{\pi}(\theta|X)$ for the problem at hand,

$$\mathcal{T}_{\text{train}} = \{(\theta_1, X_1) \ldots (\theta_B, X_B)\} \sim \pi(\theta)p(X|\theta),$$

  where $\pi(\theta)$ could be the same as $r(\theta)$, or it could be a distribution that reflects prior or selection biases (as in Case Studies II and III);

- an *unlabeled* target data set

$$\mathcal{T}_{\text{target}} = \left\{(\theta_1^*, X_1^{\text{target}}) \ldots (\theta_N^*, X_N^{\text{target}})\right\} \sim p_{\text{target}}(\theta)p(X|\theta),$$

  where neither the true parameters $\theta_1^*, \ldots, \theta_N^*$ nor the distribution $p_{\text{target}}(\theta)$ are known to the scientist.[6]

Our goal is to construct a confidence region $C(X)$ for θ that has correct frequentist coverage; that is, $\mathbb{P}_{X|\theta}(\theta \in C(X) \mid \theta) \geqslant 1 - \alpha$ for every θ. Since the conditional distribution $X \mid \theta$ is assumed to be the same across all the sets described above, we have the result that if $C(X)$ ensures valid coverage for the universal set, then it will also do so for the target data.

4.3.2 A Protocol for Valid Scientific Inference

Our proposed Frequentist-Bayes procedure mirrors the style of HPD level sets $H_c(X) = \{\theta : \widehat{\pi}(\theta|X) > c\}$ in Bayesian inference, while providing frequentist coverage properties for every $\theta \in \Theta$, regardless of $\pi(\theta)$ and the number of events per parameter. The main steps, with details described in the Supplementary Material C.2, are as follows:

1. **Learn the posterior distribution**: From training data $\mathcal{T}_{\text{train}}$, learn the posterior distribution $\pi(\theta|X)$ with, for example, a neural density estimator. The estimated posterior $\hat{\pi}(\theta|X)$, or a related function, is treated as a frequentist test statistic $\lambda(X;\theta)$. This statistic assigns a score $\lambda(X;\theta_0)$ that measures the degree to which a parameter value θ_0 is plausible given that X is observed. Examples of other posterior-based scores include the Bayes Frequentist Factor (BFF; Dalmasso* et al. (2024)) and the Waldo test statistics (Masserano et al., 2023).

2. **Transform the posterior into p-values**: From the universal set $\mathcal{T}_{\text{univ}}$, learn a family of monotonic transformations $F(\cdot;\theta)$ of the test statistic λ (Algorithm C.1 and Equation C.5). These functions are effectively "amortized p-values" that allow the construction of confidence sets at all miscoverage levels α simultaneously; see Figures 4.2B, 4.3D, 4.4C, 4.5 and 4.6B for some examples. Alternatively, if one is only interested in confidence sets at a prespecified level α (as in Section 4.2), then directly estimate "critical values" for λ, $F^{-1}(\alpha;\theta)$, at fixed α (Algorithm 2.1).

[6] From a classical statistics perspective, these parameters are perhaps best understood as "latent variables". Although each parameter θ_i^* is *fixed* and not random for each object i, we define a marginal distribution for θ that represents its prevalence in the target population. In addition, in some applications we only observe each target object once (that is, the sample size $n = 1$ for each parameter), whereas other applications allow for multiple observations ($n > 1$).

3. **Construct confidence sets**: Finally, compute Frequentist-Bayes sets $B_\alpha(X)$ by taking level sets of a transformation of $\hat{\pi}(\theta|X)$:

$$B_\alpha(X) = \{\theta \in \Theta \mid F(\hat{\pi}(\theta|X); \theta) > \alpha\} = \left\{\theta \in \Theta \mid \hat{\pi}(\theta|X) > F^{-1}(\alpha; \theta)\right\}.$$

   This computation is "amortized" with respect to X in the sense that once we have learned the posterior distribution (Step 1) and the monotonic transformation (Step 2), no further training is needed for new X: we can just evaluate the confidence set $B_\alpha(X)$.

4. **Check local coverage of constructed confidence sets**: After building confidence sets, check that the actual coverage probability $\mathbb{P}_{X|\theta}(\theta \in \hat{B}_\alpha(X))$ for data X generated at θ is indeed the same as the nominal value $(1-\alpha)$, for *every* θ in the parameter space. This check is not part of the construction of confidence sets per se, but provides the scientist with an independent diagnostic tool to assess her final results. See Algorithm 2.2 for an efficient way to compute such diagnostics. Figure 4.3 (Panels B and C, top) illustrates how these diagnostics can help domain scientists identify regions of the parameter space where the confidence sets might under- or over-cover, even when parameter distribution of the target source is unknown.

In Supplementary Materials, we prove the following key properties of our framework:

- **Correct local coverage across the parameter space**: The Frequentist-Bayes confidence procedure achieves $(1-\alpha)$ coverage for all parameter values regardless of the prior distribution (when the universal set used for recalibration is large enough); see Figure 4.6 (Panel B, right).

  See Appendix C.2.2 for theoretical results: specifically, Theorem 2.7 for guarantees on validity of the p-value approach as the number of simulations B' in the universal set increases, Theorem 2.10 for convergence rates, and Theorems 2.3 and 2.5 for the corresponding results under the critical value approach.

- **Efficiency with well-specified priors**: When the prior matches the target distribution, Frequentist-Bayes sets are optimal, with a smaller average size than other confidence sets with the same coverage properties; see Figure 4.6 (Panel B, left and center). This result is also consistent with our observations in Case Study III (see Section 4.2.3).

  See Theorem C.13 in Appendix C.3 for a formal proof that, among all valid confidence sets, Frequentist-Bayes sets $B_\alpha(X)$ are those with the smallest average size; that is, informally, $B_\alpha(X) = \arg\min_{A \in \mathcal{A}} \mathbb{E}_{p(X|\theta)\pi(\theta)}[|A(X)|]$, where $|A(X)|$ is the size of a set A and the expectation is taken over the distribution of the training data.

4.4 Conclusions

Neural posterior inference can lead to misleading scientific conclusions, even with an all-knowing simulator or perfectly labeled data. We presented a general amortized framework for

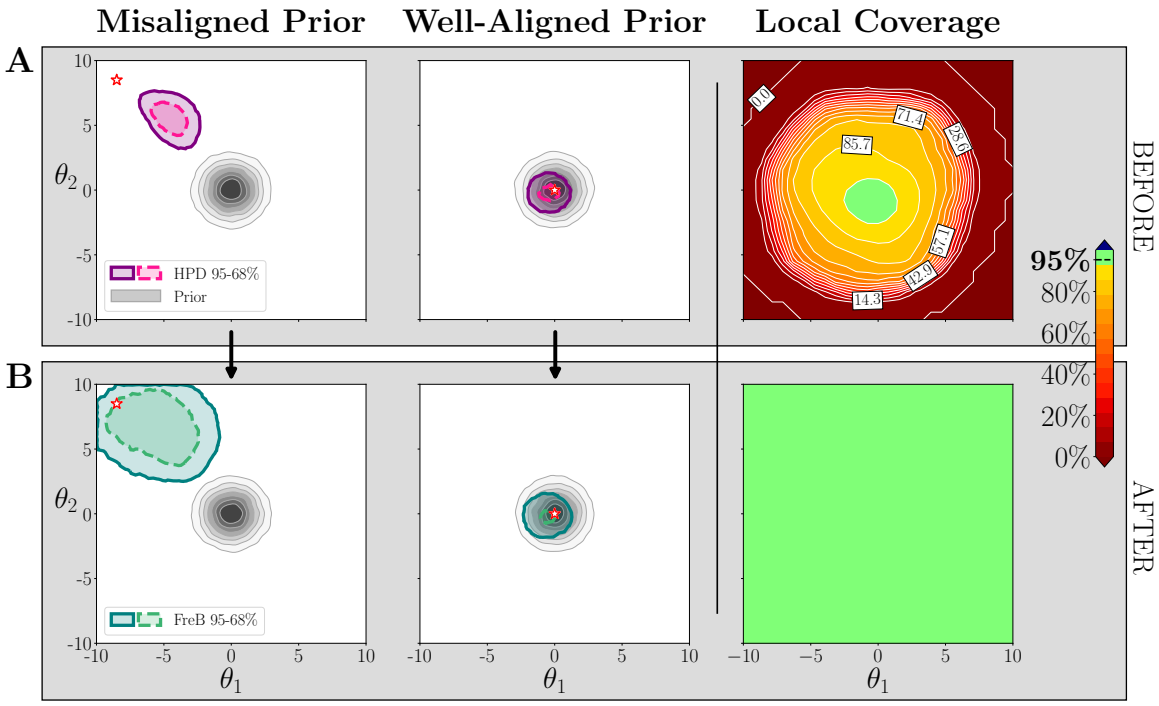


Figure 4.6: **FreB sets are simultaneously robust against misaligned priors and small in size for well-aligned priors.** Synthetic two-dimensional example where the task is to infer the location θ of a mixture of two Gaussians with different covariances, $X \sim \frac{1}{2}\mathcal{N}(\theta, \sigma_1^2 I) + \frac{1}{2}\mathcal{N}(\theta, \sigma_2^2 I)$, using a posterior learned with a Flow Matching generative model trained with a localized prior, $\pi(\theta) = \mathcal{N}(0, 2)$. **Panel A:** 95% and 68% HPD sets for two scenarios where the prior is misaligned (*left*) versus well-aligned (*center*) with the true θ. *(Right)* Local coverage plot of 95% HPD sets shows that the actual coverage of these sets can be very far from the nominal 95% level, when the truth is further away from the center where the prior is concentrated. **Panel B:** Corresponding FreB sets obtained from the same posterior estimated via the same generative model as in Panel A. For all instances of θ and for all levels of α, domain scientists are guaranteed to achieve the desired coverage level, here illustrated for the 95% case in the *right* plot. That is, FreB sets are robust against misaligned priors. Moreover, the size of FreB sets is smaller for well-aligned priors (compare *center* plot with the *left* plot).

transforming estimated posteriors into statistically valid Frequentist-Bayes (FreB) confidence sets. FreB sets contain the true parameters with the desired probability regardless of what the true parameter values are, as long as the train and target data arise from the same likelihood. However, if the domain scientist has good prior knowledge and is able to collect training data from a distribution aligned with the target data, then FreB sets become smaller than procedures that do not use prior distributions.

Our method applies broadly across several fields of science and equips researchers with a principled tool for leveraging generative AI for inverse problems in high-stakes contexts, from discovering new particles to tracking climate-driven environmental changes. Future work with FreB could explore pretraining large AI models to first learn likelihoods from multiple

sources, and then (in, e.g., a data fusion scenario) tune instrument priors for specific use cases to better constrain the main parameters of interest.

Scientific Inference Challenges Addressed in this Chapter

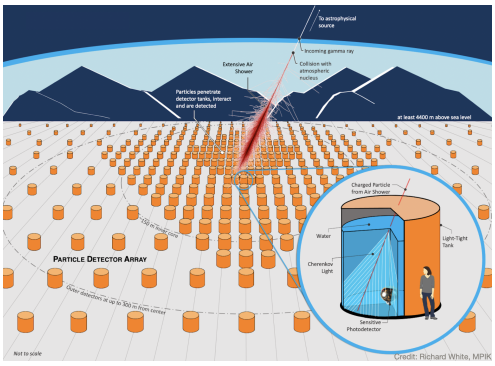
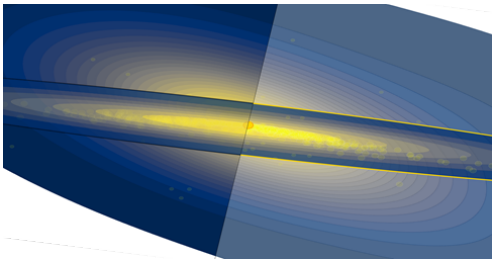
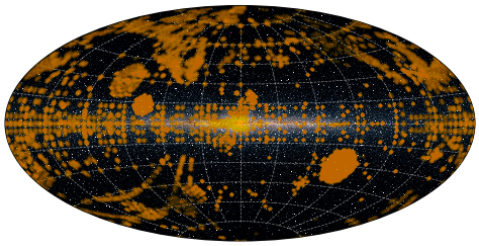
#	Inference Challenge	Case Study
I	Enable reliable identification and reconstruction of previously unknown physical sources	 Reconstructing gamma-ray showers from ground-based detectors with SBI
II	Resolve the paradox of conflicting scientific conclusions due to differing models of nature	 Inferring properties of Milky Way stars with SBI
III	Ensure trustworthy inference in the presence of selection bias in observational studies	 Inferring stellar parameters from cross-matched astronomical catalogs ("LFI beyond SBI")

Table 4.1: **Scientific inference challenges addressed in this chapter.** Each case study in Sections 4.2.1, 4.2.2 and 4.2.3 (with the set-up listed in the right column) illustrates a unique scientific challenge, which we resolve with our proposed approach. *Right Column,* **I:** Ground-based detector array for measuring atmospheric cosmic-ray showers (proposed SWGO experiment; Abreu et al. 2019). **II:** Two differing models of the galaxy, simulated using Brutus (Speagle et al., 2025). **III:** Galactic map displaying the stars included in a cross-match between Gaia Data Release 3 (Gaia Collaboration et al., 2023) and APOGEE Data Release 17 (Majewski et al., 2017).

5

Inference under Nuisance Parameters and Generalized Label Shift

5.1 Introduction

Problem Set-up. Likelihood-free inference refers to settings where the likelihood function $\mathcal{L}(x;\theta)$ — associated with a "theory" or model of the data-generating process — is intractable, but one is able to simulate relatively large data sets $\mathcal{T} = \{(\theta_1, X_1), \ldots, (\theta_B, X_B)\} \sim p_{\text{train}}(\theta)\mathcal{L}(x;\theta)$. These mechanistic models (or simulators) implicitly define the "causal" model $\theta \to X$ that encodes our knowledge of how internal parameters determine observable data, and are widely used in several domains of science.

While the likelihood $\mathcal{L}(x;\theta)$ stays the same under the assumed theory, the prior over parameters $p_{\text{train}}(\theta)$ is *chosen by design* and can be different from the true target distribution $p_{\text{target}}(\theta)$, thereby causing a potentially harmful bias when inferring θ given a new observation x_{target}. If the unknown parameter of interest is a categorical variable $Y \in \mathcal{Y} = \{0, 1, \ldots, K\}$ and the causal mechanistic model remains the same — that is, $p_{\text{train}}(X \mid Y) = p_{\text{target}}(X \mid Y)$ — the difference in the joint distribution of (θ, X) between train and target data is referred to as prior probability shift or label shift (Quinonero-Candela et al., 2008; Vaz et al., 2019; Polo et al., 2023; Storkey et al., 2009; Fawcett and Flach, 2005; Moreno-Torres et al., 2012). We refer to this setting as *standard label shift* (SLS).

In this paper, we consider a more general setup that reflects a richer mechanistic model: $\theta = (Y, \nu) \to X$, where $\nu \in \mathcal{N}$ are continuous or discrete nuisance parameters that are not of direct interest but critically influence the data-generating process. These nuisance parameters are available at the training stage, but are *not* observed at the inference stage when estimating Y from x_{target}. We refer to a shift that simultaneously affects Y and ν as *generalized label shift* (GLS), and assume that $p_{\text{train}}(X \mid Y, \nu) = p_{\text{target}}(X \mid Y, \nu)$. Within this setting, our goal is not just to do binary classification per se (that is, providing a 0 versus 1 response), but rather to do trustworthy uncertainty quantification for the classification output, even under GLS.

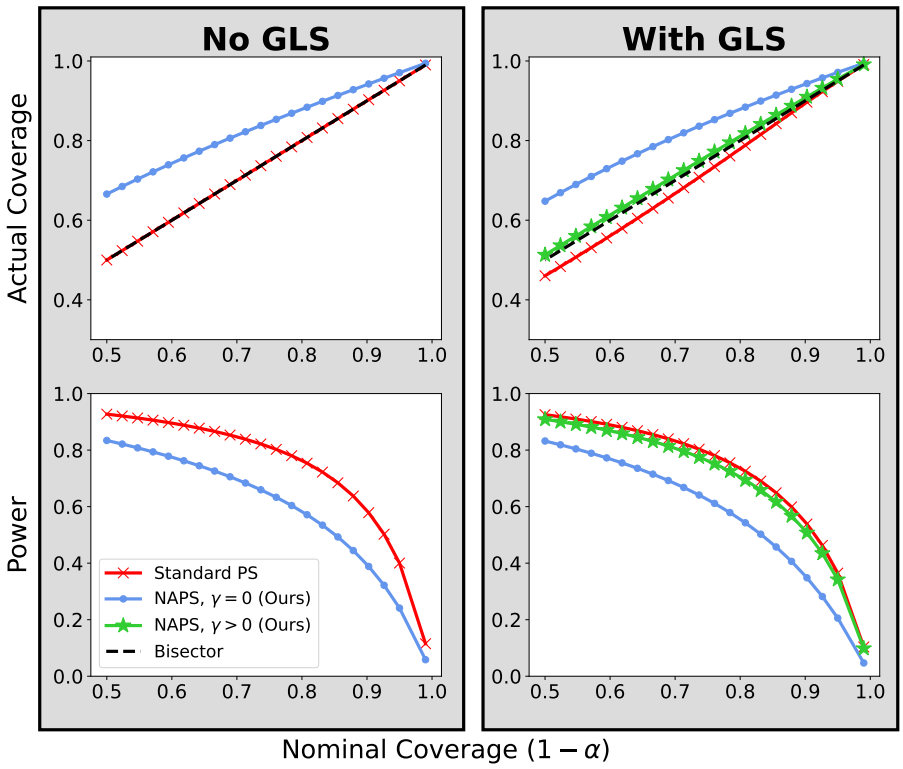


Figure 5.1: **Synthetic Example. Left (no GLS):** Standard prediction sets $\mathcal{R}_\alpha(x)$ (red) guarantee marginal coverage at the nominal level. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) are also marginally valid, but the "universality" of conditional validity across the entire nuisance parameter space comes at the price of more conservative prediction sets and lower power. **Right (with GLS):** Standard prediction sets are no longer valid and undercover for all α levels (red curve is below the black bisector), while NAPS are still valid. Furthermore, we can increase power while maintaining validity (NAPS $\gamma > 0$; green) by constructing $(1 - \gamma)$ confidence sets of the nuisance parameter ν and deriving less conservative cutoffs given an observation. Here $\gamma = \alpha \times 0.01$.

Scientific Motivation. Nuisance parameters can be seen as a way of accounting for model misspecifications. Statistical models are indeed rarely accurate in capturing the complexity of physical phenomena. To account for "known unknowns", such as calibration errors in the measuring device or inaccuracies and approximations in the theory, scientists usually resort to enlarging the mechanistic model with additional parameters that are not of direct relevance, but yet have to be considered during inference in order to make reliable statements about the parameters of interest. These additional parameters are commonly referred to as nuisance parameters (Kitching et al., 2009; Dorigo and de Castro, 2020; Pouget et al., 2013; HEP ML Community, 2025): they are necessary to achieve more faithful models of reality, but make correct inference much more challenging.

Statistical Challenges. We introduce a simplified example (see Section 5.5.1 for details) to illustrate the challenges of classification under the presence of nuisance parameters. Suppose $Y = 1$ represents a class with cases of interest (e.g., the presence of a medical condition) and $Y = 0$ a class with cases of no interest. We have good knowledge of the probability density function (PDF) of $Y = 1$, $f_1(x)$, but the shape of the distribution of $Y = 0$ is largely unknown. To accommodate different scenarios, we resort to a nuisance-parameterized PDF $f_0(x; \nu)$. Our goal is to discriminate between negative $Y = 0$ and positive $Y = 1$ cases based on potentially high-dimensional data $x \in \mathcal{X}$ and to provide valid measures of uncertainties on the true label Y under the presence of a nuisance parameter ν. However, directly classifying x_{target} based on $\mathbb{P}_{\text{train}}(Y = 1 \mid X)$ and a cutoff C derived from $\mathcal{T} = \{(Y_i, X_i)\}_{i=1}^{B}$ would lead to invalid uncertainty quantification. Indeed, under GLS (or even SLS), standard prediction sets (defined as in, e.g., Equation (5.10)) do not guarantee marginal validity:

$$\mathbb{P}_{\text{target}}(Y \in \mathcal{R}_\alpha(X)) \geqslant 1 - \alpha,$$

where Y and X are random and $\alpha \in (0, 1)$ is a pre-specified miscoverage level. Various solutions have been proposed for the SLS setting (see references in Section 5.2), whereas GLS is still a largely unexplored area in the machine learning literature. The key open challenge is to design general-purpose inference algorithms that can guarantee *valid* measures of uncertainty for all Y and ν while providing high constraining power on Y (that is, smaller prediction sets).

Returning to our simplified experiment, Figure 5.1 (top left) illustrates how standard prediction sets $\mathcal{R}_\alpha(x)$ are marginally valid when the train and target distributions are the same, while under GLS prediction sets are no longer valid even marginally (top right). Our nuisance-aware prediction sets (NAPS, $\gamma = 0$ in Figure 5.1), on the other hand, are valid in both settings. In addition, we can increase the constraining power (NAPS, $\gamma > 0$) once we observe data without the need to re-train the classifier, effectively endowing our method with domain adaptation capabilities.

Approach and Contributions. We categorize our main contributions as follows:

i) **TPR and FPR across $\mathcal{N}$.** By casting classification under GLS as a hypothesis testing problem with nuisance parameters, we propose a method to estimate the TPR and FPR curves across the nuisance parameter space via monotone regression. This allows us to compute the entire receiver-operating-characteristic (ROC) of the classifier for all $\nu \in \mathcal{N}$ (Section 5.3.2 and Algorithm D.1).

ii) **Nuisance-aware prediction sets (NAPS).** Rather than providing a point prediction based on an estimate of $\mathbb{P}_{\text{train}}(Y = 1 \mid X)$, we derive selection criteria that are valid under GLS and construct a *set-valued classifier* $\mathbf{H} : x \mapsto \{\varnothing, 0, 1, \{0, 1\}\}$ which guarantees that the true label is included in the set with probability at least $(1 - \alpha)$, regardless of the true class y and of the value of the nuisance parameters ν. That is, the prediction sets $\mathbf{H}_\alpha(X)$ guarantee conditional validity under GLS (Theorem 5.6):

$$\mathbb{P}_{\text{target}}(Y \in \mathbf{H}_\alpha(X) \mid y, \nu) \geqslant 1 - \alpha, \ \forall y \in \mathcal{Y}, \ \nu \in \mathcal{N}. \tag{5.1}$$

Standard point classifiers (e.g., the Bayes classifier; Appendix D.5) and prediction sets based on $\mathbb{P}_{\text{train}}(Y = 1 \mid X)$ are not conditionally valid across the nuisance parameter space, and hence are also not valid marginally under GLS. On the other hand, our algorithm returns valid NAPS for all levels $\alpha \in (0, 1)$ simultaneously given any new observation x_{target} without having to retrain the classifier. This also yields marginal validity under GLS (Theorem 5.6). Our results do *not* rely on asymptotic theory with the number of observations $n \to \infty$. We only assume to have a sufficient number of simulations B to train and calibrate the classifier.

iii) **NAPS with higher power.** We show how one can further increase power while maintaining validity by constraining nuisance parameters given an observed x_{target} through $(1 - \gamma)$ confidence sets of the nuisance parameters ν, where γ is a small pre-defined error level. This effectively allows to derive data-dependent cutoffs that decrease the average size of prediction sets given a specific observation.

We demonstrate our method using data from two high-fidelity scientific simulators: `scDesign3` (Song et al., 2023) which generates realistic single-cell RNA-sequencing data, and `CORSIKA` (Heck et al., 1998) which models the interactions of primary cosmic rays with the Earth's atmosphere. A flexible implementation of NAPS is available at https://github.com/lee-group-cmu/lf2i.

5.2 Related Work

To the best of our knowledge, this is the first work that estimates ROC curves across the entire parameter space $\Theta = \mathcal{Y} \times \mathcal{N}$. To construct *frequentist confidence sets*, we base our results directly on the class probability $\mathbb{P}_{\text{train}}(Y = 1 \mid X)$, rather than using a surrogate likelihood or likelihood ratio (see for example references in Cranmer et al. (2020)). The idea of *improving power* of NAPS with $\gamma > 0$ is similar to Berger and Boos (1994), and close in spirit to likelihood profiling, with the key difference that profiling does not guarantee validity (even for a large number of simulations B and under no GLS), and also requires an approximation of the likelihood and the maximum likelihood estimate of ν. The ROC *calibration* framework of Section 5.3.2 is related to Zhao et al. (2021) and Dey et al. (2022), which use monotone regression to estimate the CDF of probability integral transforms for calibrating posterior probabilities, but not for constructing valid prediction sets under GLS. When the prior distribution over y in the target data is known, $\mathbb{P}_{\text{train}}(Y = 1 \mid X)$ can be easily recalibrated to match $\mathbb{P}_{\text{target}}(Y = 1 \mid X)$ under SLS (Saerens et al., 2002; Lipton et al., 2018). However, this is not possible under GLS since ν is unknown at inference time. Moreover, our approach does not assume such a known prior. The construction of *set-valued classifiers* of Section 5.3.4 is inspired by Sadinle et al. (2019); Dalmasso* et al. (2024); Masserano et al. (2023). There are also connections to *conformal prediction*: Conformal methods are widely used because they ensure prediction sets with marginal coverage when data are exchangeable (Papadopoulos et al., 2002; Vovk et al., 2005b; Lei et al., 2018). However, conformal methods need adjustments under distributional shift when data are no longer exchangeable. Such adjustments need to be tailored for the type of shift at hand (Tibshirani et al., 2019). For instance, label shift can be addressed through label-conditional conformal prediction (Vovk et al., 2014, 2016; Sadinle et al., 2019), which

guarantees coverage conditional on the label y (Podkopaev and Ramdas, 2021, Section 2.2) under SLS, but not under the presence of nuisance parameters and GLS. Finally, our work directly addresses the existing gap in methods for constructing *reliable simulator-based inference* algorithms with valid uncertainty quantification guarantees (Hermans et al., 2021). Our work is also inspired by the vast literature in high-energy physics on hypothesis testing and *nuisance-parameterized* machine-learning methods (Feldman and Cousins, 1998; Cousins, 2006; Sen et al., 2009; Chuang and Lai, 1998; Louppe et al., 2017; Cowan et al., 2011a), which also includes the so-called "mining gold" idea of leveraging hidden information on latent variables in an all-knowing simulator (Brehmer et al., 2020).

5.3 Methodology

For simplicity, we will restrict our discussion to $Y \in \{0, 1\}$.

5.3.1 Classification as Hypothesis Testing

We reformulate the binary classification problem as a composite-versus-composite hypothesis test:

$$H_{0,y} : \theta \in \Theta_0 \quad \text{versus} \quad H_{1,y} : \theta \in \Theta_1, \tag{5.2}$$

where $\Theta_0 = \{y\} \times \mathcal{N}$, $\Theta_1 = \{y\}^c \times \mathcal{N}$. We define

$$\tau_y(x) = \frac{\mathbb{P}_{\text{train}}(Y = y \mid x)\, \mathbb{P}_{\text{train}}(Y \neq y)}{\mathbb{P}_{\text{train}}(Y \neq y \mid x)\, \mathbb{P}_{\text{train}}(Y = y)} \tag{5.3}$$

as our test statistic, which is equivalent to the Bayes factor for the test in Equation (5.2); see Appendix D.1 for a derivation. Alternatively, one can define the test statistic as the probabilistic classifier $\mathbb{P}_{\text{train}}(Y = y \mid x)$ itself. Both quantities (which are related via a monotonic transformation) can be estimated directly from a *pre-trained* classifier based on $\mathcal{T}_B$. That is, there is no need for an extra step to, e.g., learn the likelihood function $\mathcal{L}(x; Y, \nu)$ or the associated likelihood ratio statistic from simulated data as done in Cranmer et al. (2020), Rizvi et al. (2023), and references therein.

We denote the estimate of τ_y by $\widehat{\tau}_y$ and reject the null $H_{0,y}$ for small values of $\widehat{\tau}_y$. For example, if the null represents $y = 0$, then a "positive" case ($y = 1$) in binary classification would correspond to small values of $\widehat{\tau}_0$, or equivalently, large values of the probabilistic classifier $\widehat{\mathbb{P}}_{\text{train}}(Y = 1 \mid x) = 1 - \widehat{\mathbb{P}}_{\text{train}}(Y = 0 \mid x)$. In this work, we define cutoffs for $\widehat{\tau}_y$ so that prediction sets are approximately valid under nuisance parameters and GLS.

5.3.2 The Rejection Probability Across the Entire Parameter Space

To choose the optimal cutoff to reject $H_{0,y}$ and construct valid prediction sets, we need to know how the classifier performs for different values of the nuisance parameters ν. The first step is to compute the following quantity:

Definition 5.1 (Rejection probability)**.** *Let λ be any test statistic, e.g, the estimated Bayes factor, $\lambda = \widehat{\tau}_y$. The rejection probability of λ is defined as*

$$W_\lambda(C; y, \nu) \coloneqq \mathbb{P}_{target}\left(\lambda(X) \leqslant C \mid y, \nu\right), \tag{5.4}$$

where $y \in \{0, 1\}$, $\nu \in \mathcal{N}$, and $C \in \mathbb{R}$.

For fixed ν and null $H_{0,0} : Y = 0$, the receiver operating characteristic (ROC) relates the true positive rate

$$\texttt{TPR}(C; \nu) \coloneqq W_{\widehat{\tau}_0}(C; 1, \nu)$$

to the false positive rate

$$\texttt{FPR}(C; \nu) \coloneqq W_{\widehat{\tau}_0}(C; 0, \nu),$$

while varying the cutoff C. Figure 5.3 shows examples of some ROC curves at different values of ν when the null represents the negative class $y = 0$, for the setting of Section 5.5.3.

A key insight behind our method is that the rejection probability (Equation (5.4)) is invariant under GLS even if estimated from p_{train}; in other words, it is always the same for train and target data (Lemma 5.3). As a result, our ROC curves reliably measure the performance of the classifier under nuisance parameters. In practice, we can estimate $W_\lambda(C; y, \nu)$ for all y and ν simultaneously using regression with a monotonic constraint in C. The whole procedure is amortized with respect to the target data, meaning that both the base classifier and the rejection probability are estimated only once, after which they can be evaluated on an arbitrary number of observations.

5.3.3 Selecting the Optimal Cutoff under GLS

Once we know the classifier's rejection probability function, we can apply it in various ways. All our choices are robust against GLS.

Controlling FPR or TPR. Based on $W_\lambda(C; y, \nu)$, we can find the cutoff C for a new test point that either controls type-I error (FPR), or guarantees a minimum recall (TPR), or maximizes some other metric of choice that depends on both FPR and TPR. For example, FPR control at some pre-specified level $\alpha \in [0, 1]$ and $\nu_0 \in \mathcal{N}$ implies $C_\alpha = \texttt{FPR}^{-1}(\alpha; \nu_0)$, and TPR control at some minimum recall α implies $\widetilde{C}_\alpha = \texttt{TPR}^{-1}(\alpha; \nu_0)$. To control FPR or TPR *uniformly* over ν, one can instead choose $C_\alpha = \inf_{\nu \in \mathcal{N}} \texttt{FPR}^{-1}(\alpha; \nu)$, and $\widetilde{C}_\alpha = \sup_{\nu \in \mathcal{N}} \texttt{TPR}^{-1}(\alpha; \nu)$, respectively. Although robust under GLS, such cutoffs can be overly conservative.

Controlling FPR or TPR, but with more power. An alternative approach, which is still valid for any ν and can increase power, is to restrict the search over nuisance parameters to a smaller region of $\mathcal{N}$. For this approach, we first construct a confidence set $S(x; \gamma)$ for ν and fixed $y \in \{0, 1\}$ at a pre-specified $(1 - \gamma)$ level (Definition 5.4). This allows to choose a data-dependent cutoff such that

$$C^*_\alpha(x) = \inf_{\nu \in S(x;\gamma)} \{\texttt{FPR}^{-1}(\beta; \nu)\},$$

Algorithm 5.1 `Nuisance-aware prediction sets`

Input: training set $\mathcal{T} = \{(Y_i, X_i)\}_{i=1}^{B}$; calibration set $\mathcal{T}' = \{(Y_i', \nu_i', X_i')\}_{i=1}^{B'}$; observation x; test statistic $\lambda = \tau_y$; mis-coverage levels $\alpha \in [0, 1]$ and $\gamma \in [0, \alpha]$.
Output: Prediction set $\mathbf{H}_\alpha(x)$ such that Equation (5.1) holds.

// **Training**
Estimate $\mathbb{P}_{\text{train}}(Y = y \mid X)$ via a probabilistic classifier
// **Calibration**
Estimate $W_{\tau_y}(C; y, \nu) \coloneqq \mathbb{P}_{\text{target}}\left(\tau_y(X) \leqslant C \mid y, \nu\right)$ as detailed in Algorithm D.1 by

i. Computing the test statistic $\widehat{\tau}_y(x)$ as in Equation (5.3) for all $X \in \mathcal{T}'$;

ii. Constructing the augmented calibration set $\mathcal{T}''$;

iii. Estimating the rejection probability function $W_{\widehat{\tau}_y}(C; y, \nu)$ from $\mathcal{T}''$ via monotone regression.

// **Inference**
for $y \in \{0, 1\}$ **do**
    Compute $\widehat{\tau}_y(x)$ as in Equation (5.3)
    if $\gamma = 0$ **then**
        $C^*_{\alpha,y}(x) \leftarrow \inf_{\nu \in \mathcal{N}} \{\widehat{W}^{-1}_{\widehat{\tau}_y}(\alpha; y, \nu)\}$
    else
        Constrain nuisances by constructing a level-γ confidence set $S_y(x; \gamma)$ for ν
        $C^*_{\alpha,y}(x) \leftarrow \inf_{\nu \in S_y(x;\gamma)} \{\widehat{W}^{-1}_{\widehat{\tau}_y}(\alpha - \gamma; y, \nu)\}$
$\mathbf{H}(x; \alpha) \leftarrow \left\{y \in \{0, 1\} \mid \widehat{\tau}_y(x) > C^*_{\alpha,y}(x)\right\}$
return Prediction set $\mathbf{H}(x; \alpha)$ for Y

where $\beta = \alpha - \gamma$, where the minimization is over the restricted set $S(x; \gamma) \subseteq \mathcal{N}$. In practice, $S(x; \gamma)$ can be either obtained from auxiliary measurements that are available at inference time, or from a separate pre-trained model that returns valid confidence sets on ν from data x. Lemma 5.5 demonstrates that this cutoff guarantees a maximum type-I error equal to α (FPR control) for any $\nu \in \mathcal{N}$. Similarly, for TPR control, choosing $\widetilde{C}^*_\alpha(x) = \sup_{\nu \in S(x;\gamma)} \texttt{TPR}^{-1}(\beta; \nu)$ with $\beta = \alpha + \gamma$ guarantees a minimum recall of at least α. The special case of $\gamma = 0$ (and $\beta = \alpha$) corresponds to $S(x; \gamma) = \mathcal{N}$; that is, no constraints on the nuisance parameters. Finally, note that hybrid cut-offs $\texttt{FPR}^{-1}(\beta; \widehat{\nu})$ and $\texttt{TPR}^{-1}(\beta; \widehat{\nu})$ based on a *point prediction* $\widehat{\nu}(x)$ of the nuisance parameters (such as the posterior mean) would not lead to valid uncertainty quantification under GLS (see Figure D.12 in Appendix).

5.3.4 Constructing Robust Set-Valued Classifiers

Rather than just returning a single label 0/1 for each observation x like the standard Bayes classifier (Appendix D.5), our method yields prediction sets from a set-valued classifier.

Definition 5.2 (Nuisance-aware prediction set)**.** *A nuisance-aware prediction set (NAPS)*

is the set returned from a set-valued classifier $\mathbf{H} : x \mapsto \{\varnothing, 0, 1, \{0, 1\}\}$ *with*

$$\mathbf{H}(x; \alpha) = \left\{ y \in \{0, 1\} \mid \widehat{\tau}_y(x) > C^*_{\alpha,y}(x) \right\}, \tag{5.5}$$

where

$$C^*_{\alpha,y}(x) = \inf_{\nu \in S_y(x;\gamma)} \{W^{-1}_{\widehat{\tau}_y}(\beta; y, \nu)\}, \tag{5.6}$$

is the rejection cutoff, $\beta = \alpha - \gamma$ *and* $S_y(x; \gamma)$ *is a* $(1 - \gamma)$ *confidence set for* ν *defined by Equation* (5.7)*.*

This classifier guarantees user-defined levels of coverage $1 - \alpha$ (the probability that the true label is included in the set), no matter what the true class y and the nuisance parameters ν are (Theorem 5.6). The resulting prediction sets contain all labels that were not rejected by the corresponding hypothesis test. Ambiguous sets can arise in two cases: *i)* When both null hypotheses are rejected, we obtain an empty set. However, empty sets only arise at very low confidence levels (high values of α), which is typically not considered an interesting regime; *ii)* When both null hypotheses are accepted, we obtain a prediction set that includes both 0 and 1. This latter type of ambiguity reflects the uncertainty of the classifier, which typically grows at higher confidence levels (low values of α). A low-quality classifier will often report an "I-don't-know answer" for ambiguous instances if forced to guarantee a certain confidence level, rather than returning a 0/1 answer that has a high chance of being incorrect.

While $\gamma = 0$ can be the default choice for NAPS, choosing a small $\gamma > 0$ often leads to higher power (see Section 5.5). Finally, note that while our set-valued classifier targets conditional coverage under GLS according to Equation (5.1), as a by-product we also achieve prediction sets with marginal coverage under GLS (see Theorem 5.6).

Algorithm 5.1 includes a step-by-step description of the entire procedure for constructing nuisance-aware prediction sets.

5.4 Theoretical Results

Proofs for this section can be found in Appendix D.2.

5.4.1 Validity and Robustness to GLS

Lemma 5.3 (Invariance of the Rejection Probability to GLS)**.** *Under GLS, the rejection probability (Definition 5.1) of any test statistic* λ *is invariant to GLS, that is*

$$\begin{aligned} W_\lambda(C; y, \nu) &= \mathbb{P}_{target}\left(\lambda(X) \leqslant C \mid y, \nu\right) \\ &= \mathbb{P}_{train}\left(\lambda(X) \leqslant C \mid y, \nu\right). \end{aligned}$$

Nuisance-Aware Cutoffs

Definition 5.4 (Confidence set for nuisance parameters)**.** *The random set* $S_y(x; \gamma)$ *is a valid* $(1 - \gamma)$ *level confidence set for* ν *at fixed* $y \in \{0, 1\}$*, if*

$$\mathbb{P}_{target}\left(\nu \in S_y(X; \gamma) \mid y, \nu\right) \geqslant 1 - \gamma, \quad \forall \nu \in \mathcal{N}, \tag{5.7}$$

for some pre-specified value $\gamma \in [0, 1]$.

The following theorem shows that nuisance-aware cutoffs control FPR and TPR at the specified level.

Theorem 5.5 (Nuisance-aware cutoffs for FPR/TPR control)**.** *Choose a threshold* $\alpha \in [0, 1]$ *and* $\gamma \in [0, \alpha]$*. Let* $S_y(x; \gamma)$ *be a valid* $(1-\gamma)$ *confidence set for* ν *at fixed* $y \in \{0, 1\}$ *according to Definition 5.4. Let* $\lambda(X)$ *be any test statistic that measures how plausible it is that* X *was generated from* $H_{0,y}$*. Define the nuisance-aware rejection cutoff to be*

$$C^*_{\alpha,y}(x) = \inf_{\nu \in S_y(x;\gamma)} \{W^{-1}_\lambda(\beta; y, \nu)\}, \tag{5.8}$$

where $\beta = \alpha - \gamma$*, and* W *is the rejection probability in Definition 5.1. Then, for all* $\nu \in \mathcal{N}$*, we have FPR control (maximum type-I error probability for* $H_{0,y}$*):*

$$\mathbb{P}_{target}\left(\lambda(X) \leqslant C^*_{\alpha,y}(X) \mid y, \nu\right) \leqslant \alpha \tag{5.9}$$

Similarly, if

$$\widetilde{C}^*_{\alpha,y}(x) = \sup_{\nu \in S_{1-y}(x;\gamma)} \{W^{-1}_\lambda(\beta; 1-y, \nu)\},$$

with $\beta = \alpha + \gamma$*, then for all* $\nu \in \mathcal{N}$*, we have TPR control (minimum recall for* $H_{0,y}$*):*

$$\mathbb{P}_{target}\left(\lambda(X) \leqslant \widetilde{C}^*_{\alpha,y}(X) \mid 1-y, \nu\right) \geqslant \alpha.$$

Properties of the Nuisance-Aware Prediction Set

The nuisance-aware prediction set (Definition 5.2) is both *conditionally* and *marginally* valid with respect to both y and ν under GLS.

Theorem 5.6. *Let* $\mathbf{H}(x; \alpha)$ *be the nuisance-aware prediction set of Definition 5.2. Under GLS, for every* $y \in \{0, 1\}$ *and* $\nu \in \mathcal{N}$

$$\mathbb{P}_{target}(Y \in \mathbf{H}(X; \alpha) \mid y, \nu) \geqslant 1 - \alpha.$$

Moreover,

$$\mathbb{P}_{target}(Y \in \mathbf{H}(X; \alpha)) \geqslant 1 - \alpha.$$

5.5 Experiments

5.5.1 Synthetic Example

Consider a simplified setting where we are certain about the data-generating process of $Y = 1$ cases of interest, but not about that of $Y = 0$ cases. We assume

$$p(x_i \mid Y_i = 1) = \frac{e^{x_i}}{e - 1}$$
$$p(x_i \mid Y_i = 0, \nu_i) = \frac{\nu_i e^{-\nu_i x_i}}{1 - e^{-\nu_i}},$$

where $\nu \in [1, 10]$ is a nuisance parameter, which enlarges the model for $Y = 0$ to reflect our uncertainty of how cases of no direct interest might manifest themselves.

Before Data Collection. Before having specific knowledge about target data and experimental conditions, we decide to draw ν from a uniform reference distribution $p_{\text{train}}(\nu) = \text{Uniform}(1, 10)$ (here $\mathbb{P}_{\text{train}}(Y = 1) = \mathbb{P}_{\text{target}}(Y = 1) = 0.5$ is fixed). We then pre-train a classifier[1] and compute the class posterior $\mathbb{P}_{\text{train}}(Y = 1 \mid x)$, and construct $(1 - \alpha)$ prediction sets

$$R_\alpha(x) \coloneqq \{y : \mathbb{P}_{\text{train}}(Y = y \mid x) > C^*_\alpha\} \tag{5.10}$$

with cutoffs

$$C^*_\alpha \text{ s.t. } \mathbb{P}_{\text{train}}\left(\mathbb{P}_{\text{train}}(Y = y \mid X) \leqslant C^*_\alpha\right) = \alpha,$$

for a pre-specified miscoverage level α. These are the oracle prediction sets that minimize ambiguity (i.e., average size) subject to having the correct total coverage according the Theorem 1 from Sadinle et al. (2019). We will henceforth refer to them as "standard prediction sets" to distinguish them from the oracle class-conditional prediction sets from Sadinle et al. (2019) and NAPS.

Setting 1: No GLS. When train and target data have the same distributions, the prediction sets $R_\alpha(X)$ have guaranteed marginal coverage

$$\mathbb{P}_{\text{train}}(Y \in R_\alpha(X)) = 1 - \alpha$$

at the nominal $(1 - \alpha)$ level by construction (red curve overlapping black bisector in Figure 5.1, top left), although they might still undercover in specific regions of the nuisance parameter space (see Figure D.12 in Appendix D.9). NAPS with $\gamma = 0$ are instead both marginally valid (blue curve, top left) and conditionally valid (Theorem 5.6). The latter "universality" can cause overly conservative prediction sets and a loss of power (defined as the probability of rejecting $H_{0,y} : Y = y$ when $Y \neq y$); see bottom left panel.

Setting 2: With GLS. Suppose now that we apply the pre-trained classifier to a target distribution with a *different* distribution over the nuisance parameters, namely $p_{\text{target}}(\nu) = \mathcal{N}(4, 0.1) \neq p_{\text{train}}(\nu)$. The top right panel of Figure 5.1 shows that the prediction sets $R_\alpha(X)$ are no longer valid even marginally (red curve below bisector), whereas NAPS are still valid. Moreover, we can achieve higher power by constraining the optimization to a high-confidence set of the nuisance parameter (compare green with blue NAPS curves). In summary: our proposed method can leverage the original $\mathbb{P}_{\text{train}}(Y = 1 \mid x)$ classifier to provide prediction sets that are both valid and precise for any distribution $p(y, \nu)$ as long as $x \mid y, \nu$ stays the same. Additional results for other prediction set methods and NAPS with $\gamma > 0$ are available in Appendix D.9.

5.5.2 Single-Cell RNA Sequencing

RNA sequencing, or RNA-Seq, is a vital technique in genetics and genomics research that has revolutionized our understanding of gene expression. Many RNA-seq experiments involve extracting RNA from target cells and examining counts of specific genes. While the natural

[1] In this simplified example we can actually compute everything semi-analytically.

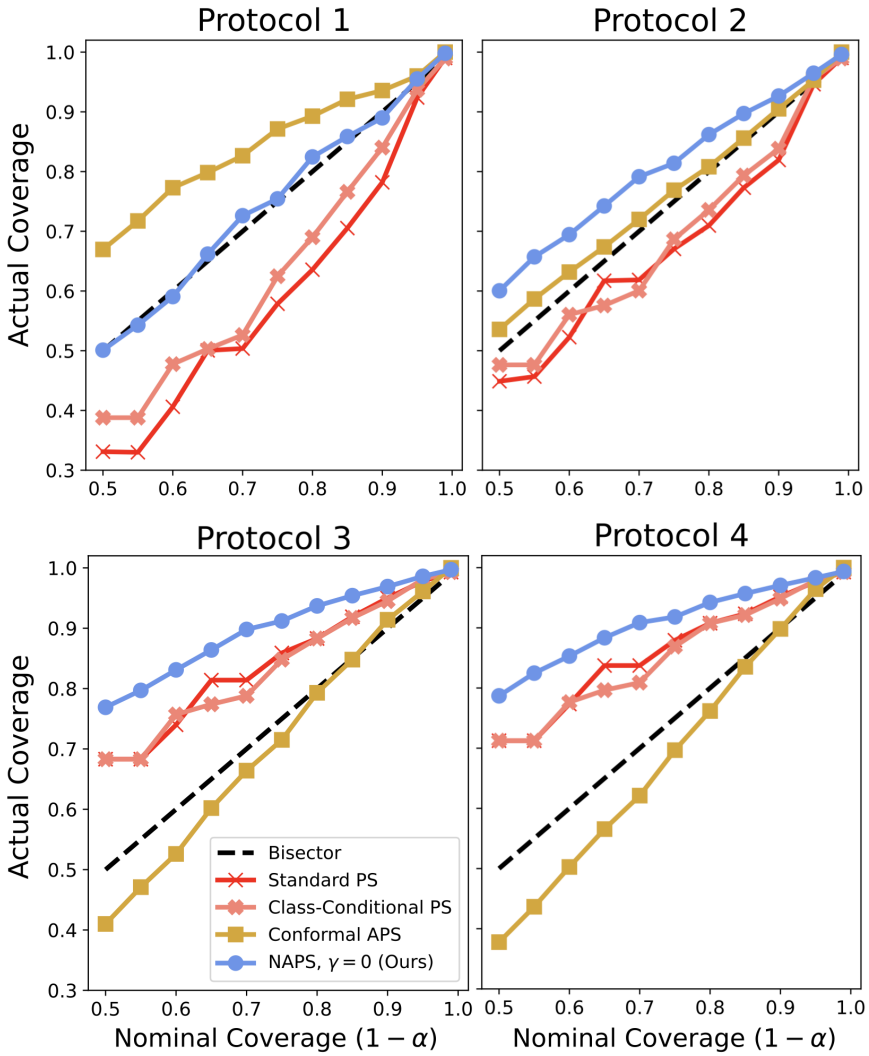


Figure 5.2: **Coverage under different batch protocols ν for the RNA-Seq example.** Each marker represents the proportion of samples in the test set whose true label was included in the constructed prediction sets. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) are valid regardless of the protocol, which is unknown at inference time. All other methods for prediction sets with marginal coverage (red), class-conditional coverage (pink), and conformal adaptive prediction sets (gold) undercover for at least two batch protocols.

variation in gene counts between different types of cells is interesting to researchers, the observed gene counts depend also on the precise steps of the sequencing process. For example, the exact chemicals, equipment, room temperature and lab technician can greatly influence the final measurements, in addition to the cell type. In practice, these so-called "batch effects" are often unmeasured confounders whose exact value is unknown at the inference stage. Thus, analysis of experimental gene counts must take them into account in order to conduct reliable scientific analysis. In what follows, we define a "batch protocol" to be a

particular set of these conditions common to a batch of cells.

We use data from the recently proposed `scDesign3` simulator (Song et al., 2023), with reference data taken from the PBMC Systematic Comparative Analysis (Ding et al., 2019). We consider two cell types (`CD4`$^+$ T-cells and `Cytotoxic` T-cells) and a subset of 100 random genes. The reference data contains counts from two separate experiments, which will serve as the basis of our simulated batch protocols. We use the two original experimental conditions as well as two artificial perturbations derived from them to generate four possible batch protocols. Following our terminology, this corresponds to a discrete nuisance parameter with four groups. We consider the setup of a classifier trained on data from all four possible protocols and tested on different x_{target} whose true protocol value is unknown (in addition to the cell type). In total, we have available 80,000 samples which we divide into train (60%), calibration (35%) and test (5%) sets. Our goal is to infer the cell's type from the observed gene count under the presence of the unknown nuisance parameter.

We compare our method with three baselines: (i) standard prediction sets for which cutoffs are computed from $\mathbb{P}(Y \mid X)$ (Sadinle et al., 2019, Theorem 1); (ii) class-conditional prediction sets with cutoffs derived separately from each $\mathbb{P}(Y = i \mid X)$, $i \in \{0, 1\}$ (Sadinle et al., 2019); and (iii) conformal adaptive prediction sets (APS; Romano et al. (2020)). Figure 5.2 shows that nuisance-aware prediction sets (NAPS) are valid regardless of the protocol, which is unknown at inference time. On the other hand, all of the other prediction sets from the analyzed baselines undercover for at least two protocols. Nuisance-aware cutoffs need to control type-I error for every single value of the nuisance parameter, including the hardest case. Here, Protocol 1 (top left) appears to be the most difficult to classify correctly. Finally, we note that while conformal APS approximately achieves coverage for $(1 - \alpha) \approx 1$, this comes at the expense of uninformative prediction sets that contain both labels for all x_{target}. NAPS, on the other hand, is able to maintain high power (see Figure D.5 in Appendix D.7). Additional results and details on the base classifier, the model used to estimate the rejection probability function, and the baselines adopted for comparison can be found in Appendix D.7.

5.5.3 Atmospheric Cosmic-Ray Showers

High-energy cosmic rays, both charged and neutral, are extremely informative probes of astrophysical sources in our galaxy and beyond. Gamma rays (which constitute the vast majority of neutral cosmics) reach the Earth atmosphere from specific directions that coincide with the location of the originating source in the sky. On the other hand, charged cosmic rays (hadrons) arrive from non-informative directions as they get deflected by galactic magnetic fields while travelling. An important step in analyzing gamma-ray sources is to separate gamma-induced showers (G) from the very large background ($> 1000 : 1$) of hadron-induced showers (H) using ground-based detector arrays that collect particles x from secondary showers (Dorigo et al. (2023, 2025); see top left of Figure 5.4 for an illustration). G/H separation is a challenging rare-event detection problem, where the true distribution of both the shower type Y and the shower parameters ν might be misspecified in simulated data. Our goal is to infer the cosmic ray identity Y from ground measurements X while

accounting for additional shower parameters: energy E, azimuth angle A and zenith angle Z. Together, these form a nuisance parameter vector $\nu = (E, A, Z)$. We construct a data set of 99,850 samples simulated from `CORSIKA` (Heck et al., 1998) divided into train (45%), calibration (45%) and test (10%) sets. Figure 5.3 (left) shows several ROC curves as a function of different energy values, demonstrating a clear dependency of the classification problem on this shower parameter.

Figure 5.5 summarizes our results as a function of the confidence level $(1-\alpha)$ for different classification metrics. These are computed within true and within predicted gamma rays for two different bins whose border is the median energy level. Nuisance-aware prediction sets (NAPS with $\gamma = 0$) achieve high precision and low false discovery rates but slightly under-perform relative to the standard Bayes classifier (Appendix D.5) for lower energy values (left column in Figure 5.5), specifically at low confidence levels. This behaviour originates from the complexity of the data: at lower energies it is indeed much harder to distinguish gamma rays from hadrons (see bottom left panel of Figure 5.4).

By constructing $(1-\gamma)$ confidence sets for ν (see the right panel of Figure 5.4 for an example), we are able to outperform the standard Bayes classifier at all confidence levels (NAPS with $\gamma > 0$). This result is explained by the bottom panel in Figure 5.5: NAPS predicts a single label only when it is relatively certain about it, and otherwise outputs an

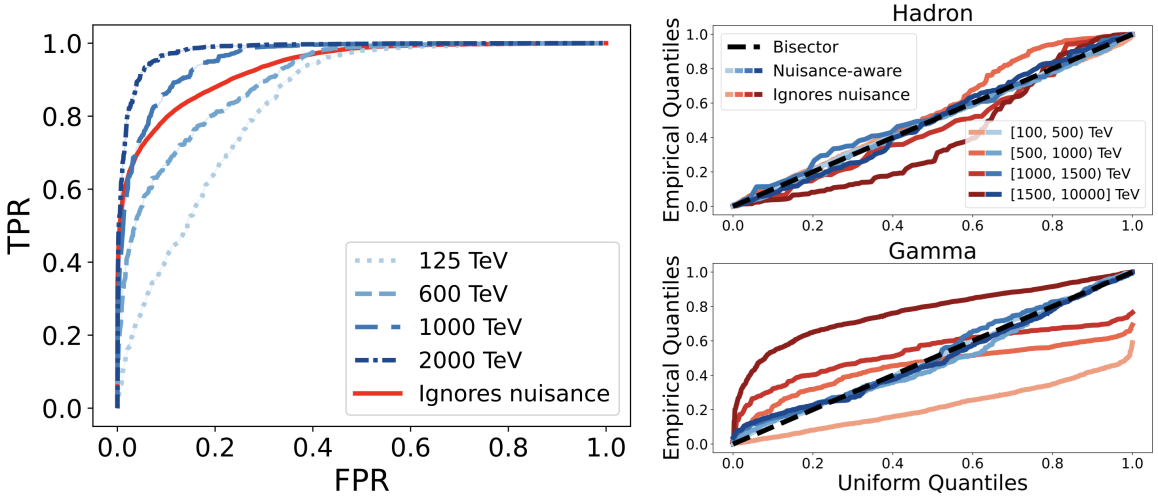


Figure 5.3: **Dependence of the ROC on the energy of the cosmic-ray shower. Left:** Receiver operating characteristic evaluated according to our method at different energy values (shades of blue). By estimating the entire ROC, we can control FPR or TPR at specified confidence levels for all $\nu \in \mathcal{N}$, which is not possible with the "marginal" ROC curve (red). **Right:** Diagnostic P-P plot evaluated at four bins over energy for nuisance-aware ROC (shades of blue) and ROC that ignores nuisances (shades of red). To check if $\mathbb{P}_{\text{target}}(\lambda(X) \leqslant C \mid y, \nu)$ is well estimated, we plot PIT values against a Uniform$(0, 1)$ distribution (dashed bisector; see Appendix D.4 for details). This is clearly not the case if one ignores nuisance parameters.

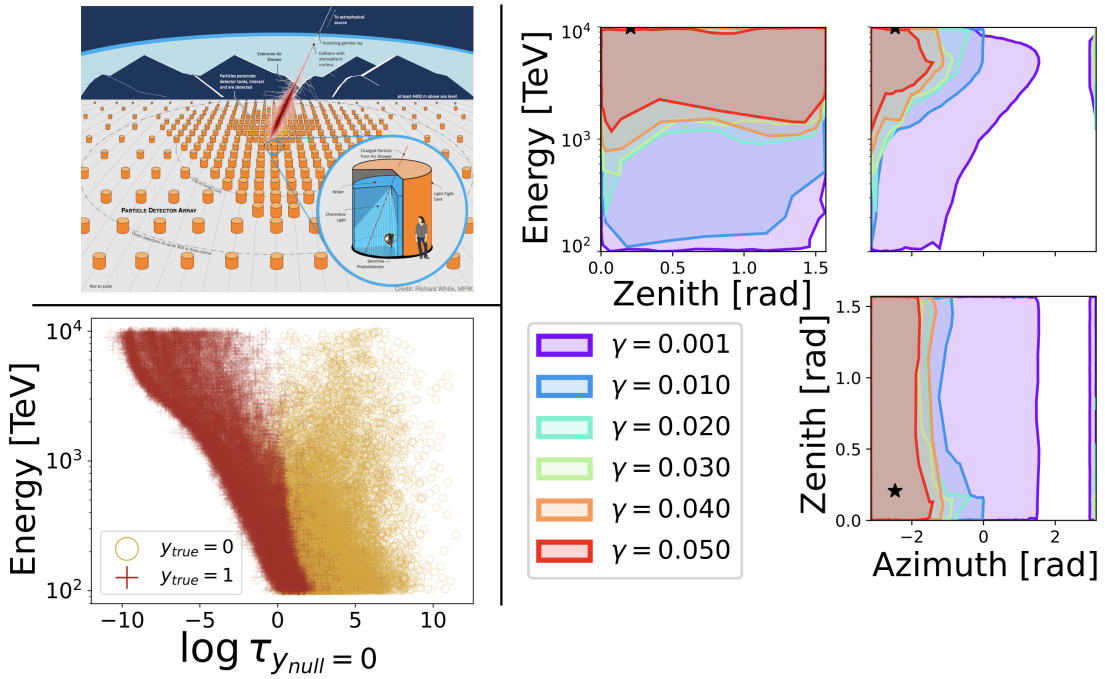


Figure 5.4: **Constraining the cosmic ray shower parameters. Top left**: Illustration of the Southern Wide-field Gamma-ray Observatory (SWGO; Abreu et al. (2019); image credit: Richard White) array of detectors with an incoming gamma ray (red). **Bottom Left**: Test statistic under $y_0 = 0$ (hadron) as a function of energy. At high energies, the class-conditional test statistics are well separated, implying that it is easier to distinguish gamma showers (red) from hadron showers (gold). **Right**: Confidence set for ν at different $(1 - \gamma)$ confidence levels obtained via the framework of Masserano et al. (2023). The true value of ν is the black star.

ambiguous prediction set that contains both labels. Nonetheless, for this example, NAPS with $\gamma = 0$ is able to achieve a higher number of true positives and lower number of false negatives relative to the Bayes classifier. Additional results and details on the models used can be found in Appendix D.6.

5.6 Conclusion and Discussion

The introduction of nuisance parameters complicates the effectiveness and reliability of machine learning models in tasks such as classification. This paper introduces a new method for handling prior probability shift of both label and nuisance parameters in likelihood-free inference when a high-fidelity mechanistic model is available. We demonstrate a new technique for estimating the ROC across the entire parameter space for binary classification problems. We also show how to construct set-valued classifiers that have a guaranteed user-specified probability $(1 - \alpha)$ of including the true label (parameter of interest), for all levels $\alpha \in [0, 1]$ simultaneously, without having to retrain the model for every α. These set-valued classifiers are valid, no matter what the true label and unknown nuisance parameters are.

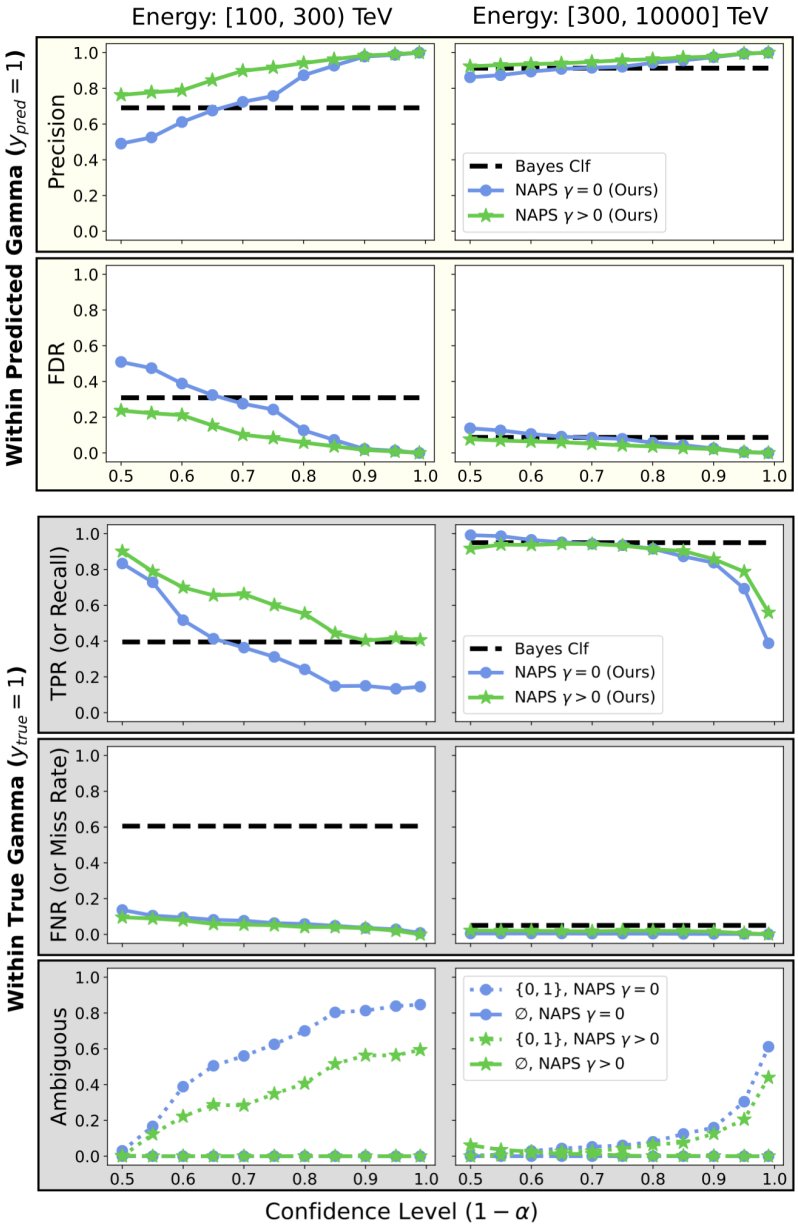


Figure 5.5: **Classification metrics within true and within predicted Gamma rays ($y = 1$).** Results are binned according to whether the shower energy is below (left) or above (right) the median value. **Top panel:** Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) achieve high precision and low false discovery rates (FDR), especially at high confidence levels. In addition, by constraining the nuisance parameters $\nu = (E, A, Z)$, we can increase performance (NAPS $\gamma > 0$; green) with uniformly better results relative to the standard Bayes classifier (black dashed line). **Bottom panel:** Our set-valued classifier makes explicit its level of uncertainty on the label y by returning ambiguous prediction sets (bottom row) for hard-to-classify x_{target}. Even so, NAPS with $\gamma > 0$ is able to achieve a higher number of true positives and lower number of false negatives relative to the Bayes classifier. Here $\gamma = \alpha \times 0.3$.

Finally, we demonstrate how to increase power while maintaining validity by constraining nuisance parameters.

Extensions and Limitations. Our approach can be extended to standard classification problems where the training data does not come from a simulator, as long as (i) the nuisance parameters ν in the data-generating process have been identified and are available at training time, and (ii) we can reliably estimate the rejection probability function across the entire parameter space as in Section 5.3.2. We recommend checking the latter with diagnostic P-P plots (see Appendix D.4, and Figure 5.3 (right) for an example).

NAPS directly extends to multiclass as one-vs-one problems, since we can estimate one-vs-one ROC curves for each $\nu \in \mathcal{N}$. The computational cost for K classes would increase by a factor of $\binom{K}{2}$. However, an extension to multiclass as one-vs-rest problems is non-trivial, because estimating ROC curves requires knowledge of the distribution of labels Y on the target set for every nuisance parameter ν. Without such knowledge, the ROC curves would not be invariant to GLS.

NAPS achieves validity under GLS. However, in the absence of a shift, this results in reduced power compared to standard prediction sets (Equation (5.10)). Although we can recover some of this power by constraining nuisance parameters (i.e. setting $\gamma > 0$), the cutoffs need to be computed for *each* test point, which can be computationally expensive, especially for high-dimensional ν. Furthermore, setting $\gamma > 0$ is not guaranteed to increase power relative to $\gamma = 0$: Since rejection probability inversion is performed at level $\alpha - \gamma$, power might *decrease* when optimizing the NAPS cutoff over the $(1 - \gamma)$ confidence set for ν (see Equation (5.8)). This can occur if the $(1 - \gamma)$ confidence sets are too large, or when the distribution of ν is skewed toward certain regions (Figure D.13). For further discussion, refer to Appendix D.9.4.

Finally, we note that NAPS may sometimes result in empty prediction sets, though this is uncommon when $(1 - \alpha)$ is large. Future adaptations could incorporate strategies from Sadinle et al. (2019) to mitigate this issue.

6

The `lf2i` package

A central goal of this thesis is to provide methods that are not only methodologically or theoretically appealing, but that are also easy to use in practice, so that domain scientists can benefit from them during their investigations. As such, we devoted a crucial effort into developing and maintaining a friendly Python package that provides scalable implementations of all the methods presented in this thesis. In this Chapter, we briefly review the main structure and contributions of this package[1].

6.1 Description of the Main Components

The central objective of the `lf2i` package is to provide an easy-to-use Python implementation of the methods and algorithms we developed in this thesis. In addition, as a long term trajectory of this effort, we would like `lf2i` to become the standard software reference for likelihood-free inference methods that bridge modern machine learning with sound frequentist guarantees, which we believe to be highly desirable in scientific inference settings.

A significant challenge in designing `lf2i` lied in accommodating different data, simulators, machine learning algorithms, test statistics, calibration methods, and in general varying degrees of flexibility within our inferential framework. As such, we tried to strike a balance between customization and ease of use, while also building an infrastructure that could potentially leverage the continuously evolving space of software packages implementing recent advancements at the intersection of machine learning and SBI/LFI. Below we describe each component that is summarized in Figure 6.1.

Data. The space of simulators across different domains of science is vast and complex, therefore we limited ourselves to providing a minimal infrastructure for practitioners to adapt their simulators so that they abide the data structures required by our inferential methods with the least possible effort. In addition, we made sure that one can directly

[1]These efforts started from an initial code-base provided by Dalmasso et al. (2020). From this starting point, we derived the implementation of `ACORE` and of critical values via quantile regression. The structure, utilities and other components of the `lf2i` package are otherwise entirely novel.

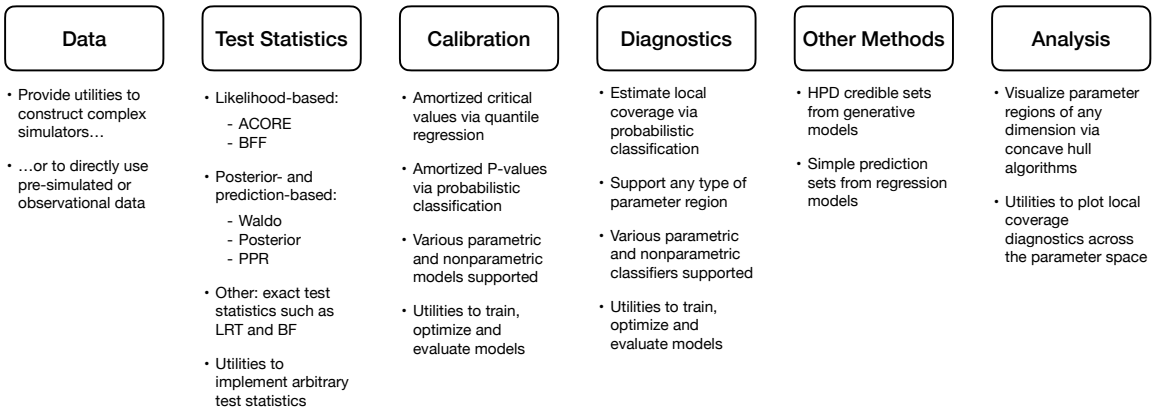


Figure 6.1: **Main components of the `lf2i` package.**

feed pre-simulated or observational data without necessarily requiring the specification of a simulator.

Test statistics. All test statistics share a common object-oriented structure. One can decide whether to train the underlying estimator (e.g., a regression, classification or generative model) from within each test statistics, or do it separately and then pass the trained estimator to the class constructor. We find the latter method fits better into the scope of `lf2i`, therefore we plan to deprecate the former (training the estimator within the class) soon. The central component is the evaluation of the test statistics, which proceeds separately for the construction of confidence sets with respect to calibration and diagnostics, so that we can fully exploit parallelization via `joblib` (Joblib Development Team, 2020) and vectorization over the appropriate dimensions. We support several test statistics that we developed over the years: some are likelihood-based, such as `ACORE` and `BFF` which leverage odds estimation; some are based on predictions and posteriors, such as `Waldo`; some are instead entirely based on posteriors (especially those obtained from generative models), such as `Posterior` and `PPR` (the prior-posterior ratio). Finally, since our framework is not necessarily based on test statistics estimated via machine learning tools, we plan to release "exact" test statistics like the standard likelihood-ratio test and Bayes Factor, since some practitioner might still find them useful in conjunction with our amortized calibration methods.

Calibration. This module mainly supports two calibration methods that are relevant to construct confidence sets via Neyman inversion: critical values via quantile regression and p-values via monotone probabilistic classification. In both cases, we limit ourselves to implement a dispatcher that trains and evaluates an appropriate calibration model, allowing the user to choose among different machine learning and statistical methods. For p-values estimation, we also implement a few ad-hoc utilities to augment the calibration dataset to re-sample the test statistics, so that the resulting estimates are also amortized with respect to the confidence level $1 - \alpha$.

While not present in Figure 6.1, the natural step after both the test statistic and the calibration method have been learned is to apply the Neyman inversion of hypothesis tests across the entire parameter space. This entails checking for which values of θ the test statistic is in the acceptance region, and retaining those values to form the confidence set. An important portion of this process is the definition of an evaluation grid over the parameter space that indexes the tests to be inverted. We foresee this to be a crucial components that should ideally be automatized and released soon.

Diagnostics. This component implements the independent diagnostic procedure described in Section 2.3.4 by i) computing indicators that signal whether a specific true parameter value is included or not in a parameter region, and ii) training a probabilistic classifier to estimate local coverage across the parameter space. This tools has proven to be very useful in practice because it allows to identify regions where either the calibration procedure is failing or the domain scientist needs to collect more data to provide more reliable results. Note that here we used the general term "parameter regions" because this module explicitly allows to check the local empirical coverage for arbitrary sets, whether they are confidence sets constructed with our methods, credible regions from posterior distributions, prediction sets, and more.

Other methods. This module aims at being a collection of alternative methods to construct parameter regions and to provide comparisons. It currently implements an efficient method to compute high-posterior density credible regions of any dimension from neural density estimators, and it also provide a simple function to compute gaussian-like prediction sets. In the future, we foresee this module to potentially accommodate additional methods such as conformal inference, when their implementation is not sufficiently easy to use from other packages.

Analysis. Finally, this module implements several utilities to analyze and visualize results. Providing an exhaustive visualization module is beyond the scope of the package, especially given the numerous customizations that are often needed in practice to obtain figures of publishable quality. This said, we tried to at least provide a minimal amount of structure to plot local coverage diagnostics across the parameter space, and to visualize parameter regions of any dimension. The latter task required the use of specialized libraries that implement concave hull algorithms (e.g., `alphashape` (Bellock, 2021)) to plot contours of regions of arbitrary shape.

6.2 Related Software

Recently, most of the advancements in LFI have been driven by the development of new machine learning methods, most of which have also been influenced by innovations in deep learning and generative models. In parallel, several packages that implement neural network-based SBI algorithms have emerged, such as `SBI` (Tejero-Cantero et al., 2020), `BayesFlow` (Radev et al., 2023b), `sbijax` (Dirmeier et al., 2024) and `Swyft` (undark lab,

2023). Our package — `lf2i` — is complementary to these efforts. We do not aim to provide the same functionalities, but rather to leverage those packages and their implementations of modern deep learning methods for SBI to estimate better and more flexible test statistics, with which we can then construct confidence sets with sounds statistical guarantees.

7

Extensions and Future Work

We conclude by discussing a few methodological extensions and novel applications that we have been working on and that will set the ground for future explorations.

7.1 Anytime-Valid Sequential Likelihood-Free Inference

Our discussion so far has revolved around the problem of constructing confidence sets in LFI settings leveraging machine learning models trained on a pre-determined number of simulations. Consider now a setting where, given a fixed x_{obs} and an implicit likelihood model F_θ from which we can simulate pairs (θ, X), we would like to learn a neural density estimator $s_\phi : X \mapsto q_\psi$ such that $q_\psi \approx \pi(\theta \mid x_{\text{obs}})$. An approach is to proceed sequentially over several rounds, using the last posterior estimate to adaptively choose where to query F_θ and obtain more informative simulations. See Algorithm 7.1 for a sketch of the steps involved. Several methods have been proposed for this setting (e.g., Lueckmann et al. (2017); Papamakarios and Murray (2016); Greenberg et al. (2019)), but none of them provides theoretical guarantees on the coverage and size of the credible regions obtained from the resulting estimated posterior distributions, nor on the optimality (if any) of the updates on the proposal used to sample θ at each round. In addition, it is usually nearly impossible to apply modern diagnostics to these algorithms, as it would require to re-train them for each different x_{obs}.

To fill this gap, an interesting are for future work would entail the development of a method that ensures anytime validity (similarly to Equation (2.1) but for sequential settings) while providing meaningful (i.e., tight) constraints on θ. Part of the motivation behind this direction is the need to improve sample efficiency in settings where obtaining more data can be expensive. More specifically, taking inspiration from the literature on sequential testing and game-theoretic statistics (e.g., Grünwald et al. (2020); Neiswanger and Ramdas (2021); Waudby-Smith and Ramdas (2020, 2024)), we started by considering the following: let $p_0(\theta) = \pi(\theta)$ be the prior distribution and $\hat{p}_r(\theta \mid x_{\text{obs}})$ be the estimated posterior after having simulated $(X_1, \dots, X_r)$ in r sequential rounds, and define the prior-posterior ratio to

Algorithm 7.1 `Sequential Neural Posterior Estimation`

Input: observation x_{obs}, simulator F_θ, prior $\pi(\theta)$, simulations per round N, rounds R, neural network s_ϕ
Output: $\hat{q}_{\psi=s_\phi}$

Set $\tilde{p}_1 \leftarrow \pi(\theta)$
for r in $\{1, \dots, R\}$ **do**
    for j in $\{1, \dots, N\}$ **do**
        Sample $\theta_{r,j} \sim \tilde{p}_r(\theta)$
        Simulate $X_{r,j} \sim F_{\theta_{r,j}}$
    $\phi_{r+1} \leftarrow \text{argmin}_\phi$ L [some loss ensuring that the estimate is close to $p(\theta \mid x_{\text{obs}})$ and not to $\tilde{p}(\theta \mid x_{\text{obs}})$]
    $\psi_{r+1} \leftarrow s_{\phi_{r+1}}(x_{\text{obs}})$
    $\tilde{p}_{r+1} \leftarrow q_{\psi_{r+1}}(\theta)$
return $q_{\psi_R}(\theta)$

be

$$\tau_r(x_{\text{obs}}; \theta) \coloneqq \frac{p_0(\theta)}{\hat{p}_r(\theta \mid x_{\text{obs}})}. \tag{7.1}$$

Note that Equation (7.1) is equivalent to the the inverse of the Bayes factor, and is in fact an alternative way of computing the test statistic proposed earlier in Section 2.3. Following Waudby-Smith and Ramdas (2020), it should be the case that the sequence $(\tau_r(x_{\text{obs}}; \theta^\star))_{r=1}^R$ is a non-negative martingale with respect to $(\mathcal{F}_r)_{r=1}^R$, i.e. the filtration induced by the sequence of simulations $(X_1, \dots, X_R)$. Furthermore, from Ville's inequality (Ville, 1939)

$$\mathcal{R}_r(x_{\text{obs}}) = \{\theta \in \Theta : \tau_r(X; \theta) < 1/\alpha\} \tag{7.2}$$

forms a $(1 - \alpha)$ confidence sequence for $\theta^\star$, i.e. $\mathbb{P}\left(\exists r : \theta^\star \notin \mathcal{R}_r(x_{\text{obs}})\right) \leqslant \alpha$.
To quickly test this idea, we setup a simple toy experiment where we train a sequential neural posterior estimator on data simulated in rounds from $X \sim Beta(\frac{30 \cdot \theta}{1-\theta}, 30)$ with a $\theta \sim Unif(0, 1)$ prior at $r = 0$. The goal is to estimate the mean of the Beta distribution, for which we generate a fixed observation $x_{\text{obs}} \sim Beta(10, 30)$, implying $\theta^\star = 1/4$. Figure 7.1 compares HPD regions from the estimated posterior and confidence sets computed according to Equation (7.2), at each round. While HPD regions roughly concentrate around $\theta^\star$ as more simulations are sampled, confidence sets seem extremely conservative (only the lower bound is marginally improving over rounds).

One important difference of this approach relative to the previous sections is that here the confidence sets are constructed using a single common cutoff (i.e., $1/\alpha$) for all the null hypotheses tested for Neyman inversion. While this guarantees validity, it is known to lead to conservative estimates in several settings. A possible promising direction is to look at the literature that re-frames testing as the game of a gambler that bets sequentially to increase its capital. Loosely speaking, one can construct a capital process for each null hypothesis and reject the null when the capital is greater than some threshold. This framework allows to obtain parameter constraints that are both anytime-valid and powerful (see, e.g., Waudby-Smith and Ramdas (2024); Ramdas and Wang (2024)). The feasibility

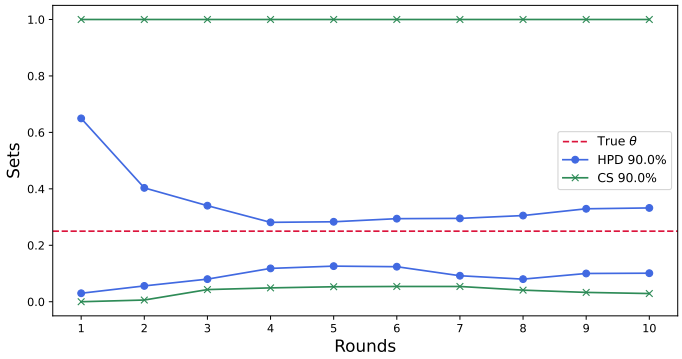


Figure 7.1: **Confidence sequences and credible regions for the mean of a $Beta(10, 30)$ distribution.** HPD sets correctly concentrate around the true parameter as the posterior estimates improve from additional training data. Confidence sequences achieve validity, but remain very conservative.

of leveraging these tools relies on their transferability to LFI settings, where little-to-no assumptions can usually be made on the distribution that generates the data, since the likelihood is by definition intractable. Another interesting area to attack this problem is the literature on the well-known best-arm identification problem for multi-armed bandits (see, e.g., Jamieson and Nowak (2014); Kuchibhotla and Zheng (2020)): re-framing the task of actively choosing where to sample in the parameter space to increase sample efficiency as that of selecting the arm with the highest reward might yield a procedure with important optimality properties. Finally, another promising approach to actively decide where to sample in the parameter space is to leverage influence functions, specifically the practical results derived by Koh and Liang (2017). In this work, the authors show how to compute influence functions for black-box predictors to analyze the effect that perturbations on the input data have on the loss and on the model predictions at a certain test point. In our context, this could provide a powerful tool to design "friendly"[1] perturbations of the inputs (for us, the parameters) that would lead to a maximal decrease of the loss, thereby guiding sampling across the parameter space to a prior distribution that is "optimal" to infer the unknown $\theta^\star$.

7.2 LF2I for Data Assimilation

In terms of sample efficiency, the calibration methods introduced in Chapters 2 and 4 scale exponentially better with the dimensionality of the parameter space relative to Monte Carlo methods, when applied in the context of the Neyman inversion. Nonetheless, with very high-dimensional parameter spaces the curse of dimensionality is inevitable. One setting that opens the way to new areas of investigation is the application of LF2I-related methods to data assimilation and inference on the latent states of a dynamical system. By construction, these latent states have an inherent structure determined by the stochastic process that defines the system itself. How do we actively exploit this structure to achieve type-I error

[1] As opposed to adversarial.

control in very high-dimensional parameter spaces and construct confidence sets for latent states? A possible solution is to use nonlinear dimension reduction methods and apply quantile regression and Neyman inversion in this space. For example, this could be achieved by learning a geometric graph neural network over the latent space. A precise formalization and solution of this problem is left to future work.

7.3 Improving on the LRT by Leveraging Prior Distributions in Particle Physics

This section presents an extension and application of some of the results in Chapters 2-4. In particle physics, experiments rely on sophisticated statistical methods to extract physics information from data. Searches for new physics phenomena and measurement of standard model parameters are typically performed as composite hypothesis tests in a frequentist framework using the generalized likelihood ratio test (LRT). Across experiments, the reach and precision of these studies are often limited by the amount of data that can be collected. A procedure that could improve the power of the statistical technique itself around the parameter space that is being tested would benefit experiments spanning collider physics, neutrino physics, dark matter searches and beyond.

The Neyman-Person lemma guarantees that the LRT is the most powerful test statistic for simple hypothesis tests; however, this does not extend to composite hypothesis tests. Here we leverage the Bayes Factor as a frequentist test statistic to provide particle physicists with the ability to assign more statistical power in relevant regions of the hypothesis space, at the cost of giving up statistical power in less interesting parts of the parameter space.

One of the main results is summarized in Figure 7.2. When scientific knowledge is well aligned with the truth via a prior distribution, a significant improvement in constraining power can be achieved. In this work, we also plan to demonstrate that a significant improvement in sensitivity can be achieved under various scenarios using the HiggsML collider physics dataset simulated by the ATLAS experiment (Adam-Bourdarios et al., 2014).

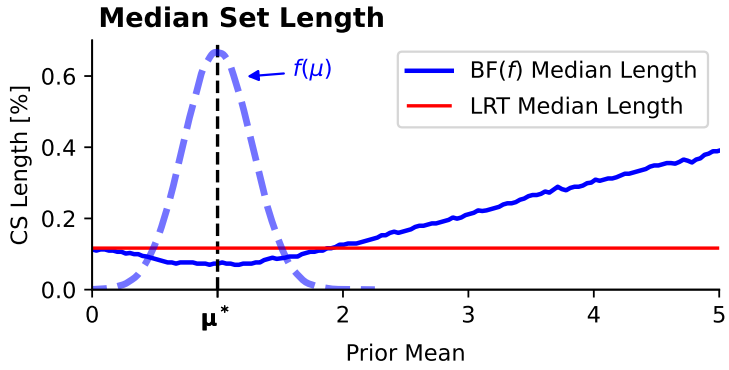


Figure 7.2: **Median 68.3% confidence interval length for LRT (red) and for BF (solid blue curve) with a truncated normal prior distribution (dashed blue curve).** Observations are sampled from different values of μ to show the gain of power around the prior and the loss of power far from it.

Bibliography

G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A.A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, Sep 2012a. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.020.

Georges Aad, Tatevik Abajyan, B Abbott, J Abdallah, S Abdel Khalek, Ahmed Ali Abdelalim, R Aben, B Abi, M Abolins, OS AbouZeid, et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, 2012b.

A. A. Abdo and Others. Fermi large area telescope observations of markarian 421: The missing piece of its spectral energy distribution. *The Astrophysical Journal*, 736(2):131, jul 2011. doi: 10.1088/0004-637X/736/2/131. URL `https://dx.doi.org/10.1088/0004-637X/736/2/131`.

A.U. Abeysekara et al. The high-altitude water cherenkov (hawc) observatory in méxico: The primary detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1052:168253, July 2023. ISSN 0168-9002. doi: 10.1016/j.nima.2023.168253. URL `http://dx.doi.org/10.1016/j.nima.2023.168253`.

P Abreu, A Albert, R Alfaro, C Alvarez, R Arceo, P Assis, F Barao, J Bazo, JF Beacom, J Bellido, et al. The southern wide-field gamma-ray observatory (swgo): A next-generation ground-based survey instrument for vhe gamma-ray astronomy. *arXiv preprint arXiv:1907.07737*, 2019.

Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge. *URL http://higgsml. lal. in2p3. fr/documentation*, 9, 2014.

Sea Agostinelli, John Allison, K al Amako, John Apostolakis, H Araujo, Pedro Arce, Makoto Asai, D Axen, Swagato Banerjee, GJNI Barrand, et al. Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

J. Aleksić et al. Measurement of the crab nebula spectrum over three decades in energy with the magic telescopes. *Journal of High Energy Astrophysics*, 5-6:30–38, 2015. ISSN 2214-4048. doi: https://doi.org/10.1016/j.jheap.2015.01.002. URL https://www.sciencedirect.com/science/article/pii/S2214404815000038.

R. Alfaro et al. Gamma/hadron separation with the hawc observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 1039:166984, 2022. ISSN 0168-9002. doi: https://doi.org/10.1016/j.nima.2022.166984. URL https://www.sciencedirect.com/science/article/pii/S0168900222004247.

Sara Algeri, Jelle Aalbers, Knut Dundas Morå, and Jan Conrad. Searching for new physics with profile likelihoods: Wilks and beyond. *arXiv preprint arXiv:1911.10237*, 2019.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023b.

Astropy Collaboration and Astropy Project Contributors. The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package. *The Astrophysical Journal*, 935(2):167, 2022. doi: 10.3847/1538-4357/ac7c74.

ATLAS Collaboration. The quest to discover supersymmetry at the atlas experiment, 2024. URL https://arxiv.org/abs/2403.02455.

J-E Augustin, Adam M Boyarski, Martin Breidenbach, F Bulos, JT Dakin, GJ Feldman, GE Fischer, D Fryberger, G Hanson, B Jean-Marie, et al. Discovery of a narrow resonance in e+ e- annihilation. *Physical Review Letters*, 33(23):1406, 1974.

Meili Baragatti, Casenave Céline, Bertrand Cloez, David Métivier, and Isabelle Sanchez. Approximate bayesian computation with deep learning and conformal prediction. *arXiv preprint arXiv:2406.04874*, 2024.

M. J. Bayarri and J. O. Berger. The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004. doi: 10.1214/088342304000000116.

Mark Beaumont and Bruce Rannala. The Bayesian revolution in genetics. *Nature reviews. Genetics*, 5:251–61, 05 2004. doi: 10.1038/nrg1318.

Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, 41(1):379–406, 2010.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Kenneth E. Bellock. Alpha shape toolbox, 2021. URL https://pypi.org/project/alphashape/.

James Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–402, 2006. doi: 10.1214/06-BA115.

Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

Herman J Bierens. Uniform consistency of kernel estimators of a regression function under generalized conditions. *Journal of the American Statistical Association*, 78(383):699–707, 1983.

Michael GB Blum and Olivier François. Non-linear regression models for approximate bayesian computation. *Statistics and computing*, 20:63–73, 2010.

Mark Bolden and Paul Kervin. Panoramic-survey telescope and rapid response system: Leveraging astronomical technology for satellite situational awareness. *38th COSPAR Scientific Assembly*, 38:3, 2010.

Rongmon Bordoloi, Simon J Lilly, and Adam Amara. Photo-z performance for precision cosmology. *Monthly Notices of the Royal Astronomical Society*, 406(2):881–895, 2010.

Denis Boyda, Gurtej Kanwar, Sébastien Racanière, Danilo Jimenez Rezende, Michael S Albergo, Kyle Cranmer, Daniel C Hackett, and Phiala E Shanahan. Sampling using su (n) gauge equivariant flows. *Physical Review D*, 103(7):074504, 2021.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1915980117.

Richard P Brent. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Paul-Christian Bürkner, Marvin Schmitt, and Stefan T Radev. Simulations in statistical workflows. *arXiv preprint arXiv:2503.24011*, 2025.

Nicola Cabibbo. Unitary Symmetry and Leptonic Decays. *Phys. Rev. Lett.*, 10:531–533, 1963. doi: 10.1103/PhysRevLett.10.531.

Kenneth G. Carpenter, Richard D. Robinson, Graham M. Harper, Philip D. Bennett, Alexander Brown, and Dermott J. Mullan. GHRS Observations of Cool, Low-Gravity Stars. V. The Outer Atmosphere and Wind of the Nearby K Supergiant λ Velorum*. *The Astrophysical Journal*, 521(1):382, 1999. doi: 10.1086/307520.

James Carzon, Luca Masserano, Aishik Ghosh, Daniel Whiteson, Rafael Izbicki, and Ann Lee. On focusing statistical power for searches and measurements in particle physics. *In Submission*, 2025.

Paula Chadwick. 35 years of ground-based gamma-ray astronomy. *Universe*, 7(11):432, 2021.

MT Chao. The asymptotic behavior of Bayes' estimators. *The Annals of Mathematical Statistics*, 41(2):601–608, 1970.

Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, Thomas Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 gev with the cms experiment at the lhc. *Physics Letters B*, 716(1):30–61, 2012.

Jiahua Chen and Pengfei Li. Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542, 2009.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Yanzhi Chen and Michael U. Gutmann. Adaptive Gaussian copula ABC. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1584–1592. PMLR, 16–18 Apr 2019.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Chin-San Chuang and Tze Leung Lai. Resampling methods for confidence intervals in group sequential trials. *Biometrika*, 85(2):317–332, 06 1998. ISSN 0006-3444. doi: 10.1093/biomet/85.2.317. URL `https://doi.org/10.1093/biomet/85.2.317`.

Chin-Shan Chuang and Tze Leung Lai. Hybrid resampling methods for confidence intervals. *Statistica Sinica*, 10(1):1–33, 2000. ISSN 10170405, 19968507.

Marco Cirelli, Alessandro Strumia, and Jure Zupan. Dark matter. *arXiv preprint arXiv:2406.01705*, 2024.

Grégoire Clarté, Christian P Robert, Robin J Ryder, and Julien Stoehr. Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607, 2021.

Cdf Collaboration et al. Observation of top quark production in pbar-p collisions. *arXiv preprint hep-ex/9503002*, 1995.

Gaia Collaboration et al. The gaia mission. *arXiv preprint arXiv:1609.04153*, 2016.

Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006a.

Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692, 2006b. doi: 10.1198/106186006X136976.

Gabriele Corso, Bowen Jing, Regina Barzilay, Tommi Jaakkola, et al. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR 2023)*, 2023.

Robert D Cousins. Treatment of nuisance parameters in high energy physics, and possible justifications and improvements in the statistics literature. In *Statistical Problems In Particle Physics, Astrophysics And Cosmology*, pages 75–85. World Scientific, 2006.

Robert D. Cousins. Lectures on statistics in theory: Prelude to statistics in practice, 2018.

Robert D Cousins, James T Linnemann, and Jordan Tucker. Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a poisson process. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 595(2):480–501, 2008.

Glen Cowan. Discovery sensitivity for a counting experiment with back- ground uncertainty. *Technical Report*, 2012.

Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71:1–19, 2011a.

Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71(2), Feb 2011b. ISSN 1434-6052. doi: 10.1140/epjc/s10052-011-1554-0.

Kyle Cranmer. Practical Statistics for the LHC. *arXiv e-prints*, art. arXiv:1503.07622, Mar 2015.

Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1912789117.

Yifan Cui and Min-ge Xie. Confidence distribution and distribution estimation for modern statistical inference. In *Springer Handbook of Engineering Statistics*, pages 575–592. Springer, 2023.

Didier Dacunha-Castelle and Elisabeth Gassiat. Testing in locally conic models, and application to mixture models. *ESAIM: Probability and Statistics*, 1:285–317, 1997.

Niccolo Dalmasso, Rafael Izbicki, and Ann Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2323–2334, Virtual, 13–18 Jul 2020. PMLR.

Niccolò Dalmasso*, Luca Masserano*, David Zhao, Rafael Izbicki, and Ann B Lee. Likelihood-free frequentist inference: Bridging classical statistics and machine learning for reliable simulator-based inference. *Electronic Journal of Statistics*, 18(2):5045–5090, 2024.

Gauri Sankar Datta and Trevor J. Sweeting. Probability matching priors. In D.K. Dey and C.R. Rao, editors, *Bayesian Thinking*, volume 25 of *Handbook of Statistics*, pages 91–114. Elsevier, 2005. doi: https://doi.org/10.1016/S0169-7161(05)25003-4.

Anthony C Davison, Simone A Padoan, and Mathieu Ribatet. Statistical modeling of spatial extremes. *Statistical Science*, 27(2):161–186, 2012.

Arnaud Delaunoy, Joeri Hermans, François Rozet, Antoine Wehenkel, and Gilles Louppe. Towards reliable simulation-based inference with balanced neural ratio estimation. *Advances in Neural Information Processing Systems*, 35:20025–20037, 2022.

Biprateep Dey, Jeffrey A Newman, Brett H Andrews, Rafael Izbicki, Ann B Lee, David Zhao, Markus Michael Rau, and Alex I Malz. Re-calibrating photometric redshift probability distributions using feature-space regression. *arXiv preprint arXiv:2110.15209*, 2021.

Biprateep Dey, David Zhao, Jeffrey A Newman, Brett H Andrews, Rafael Izbicki, and Ann B Lee. Calibrated predictive distributions via diagnostics for conditional coverage. *arXiv preprint arXiv:2205.14568*, 2022.

Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. Systematic comparative analysis of single cell rna-sequencing methods. *BioRxiv*, page 632216, 2019.

Simon Dirmeier, Simone Ulzega, Antonietta Mira, and Carlo Albert. Simulation-based inference with the python package sbijax. *arXiv preprint arXiv:2409.19435*, 2024.

Axel Donath and Others. Gammapy: A python package for gamma-ray astronomy. *Astronomy and Astrophysics*, 678:A157, 2023. doi: 10.1051/0004-6361/202346488. URL `https://doi.org/10.1051/0004-6361/202346488`.

David L Donoho. Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probability Theory and Related Fields*, 99(2):145–170, 1994.

Tommaso Dorigo and Pablo de Castro. Dealing with nuisance parameters using machine learning in high energy physics: a review. *arXiv preprint arXiv:2007.09121*, 2020.

Tommaso Dorigo, Sofia Guglielmini, Jan Kieseler, Lukas Layer, and Giles C Strong. Deep regression of muon energy with a k-nearest neighbor algorithm. *arXiv preprint arXiv:2203.02841*, 2022.

Tommaso Dorigo, Max Aehle, Julien Donini, Michele Doro, Nicolas R Gauger, Rafael Izbicki, Ann Lee, Luca Masserano, Federico Nardi, Alexander Shen, et al. End-to-end optimization of the layout of a gamma ray observatory. *arXiv preprint arXiv:2310.01857*, 2023.

Tommaso Dorigo, Michele Doro, Max Aehle, Muhammad Awais, Nicolas R Gauger, Rafael Izbicki, Jan Kieseler, Ann B Lee, Luca Masserano, Federico Nardi, et al. On the utility function of experiments in fundamental science. *Physics Open*, page 100270, 2025.

Michele Doro, Miguel Angel Sánchez-Conde, Moritz Hütten, et al. Advances in very high energy astrophysics. In *Advances in Very High Energy Astrophysics*. World Scientific, 2024.

Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012, Apr 2009. ISSN 0090-5364. doi: 10.1214/07-aos571.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. `nflows`: Normalizing flows in PyTorch, November 2020a. URL `https://doi.org/10.5281/zenodo.4296287`.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2771–2781. PMLR, 13–18 Jul 2020b.

Antonio D'Isanto and Kai Lars Polsterer. Photometric redshift estimation via deep learning-generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy & Astrophysics*, 609:A111, 2018.

Matteo Fasiolo, Simon N. Wood, Florian Hartig, and Mark V. Bravington. An extended empirical saddlepoint approximation for intractable likelihoods. *Electron. J. Statist.*, 12 (1):1544–1578, 2018. doi: 10.1214/18-EJS1433.

Tom Fawcett and Peter A Flach. A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58:33–38, 2005.

G. Feldman. Multiple measurements and parameters in the unified approach. Technical report, Technical Report, Talk at the FermiLab Workshop on Confidence Limits, 2000.

Gary J. Feldman and Robert D. Cousins. Unified approach to the classical statistical analysis of small signals. *Physical Review D*, 57(7):3873–3889, Apr 1998. ISSN 1089-4918. doi: 10.1103/physrevd.57.3873.

R.A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 11th ed. rev. edition, 1925.

Edwin Fong and Chris C Holmes. Conformal bayesian computation. *Advances in Neural Information Processing Systems*, 34:18268–18279, 2021.

Peter E Freeman, Rafael Izbicki, and Ann B Lee. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565, 2017.

Gaia Collaboration et al. Gaia Data Release 3. Summary of the content and survey properties. *Astronomy and Astrophysics*, 674:A1, 2023. ISSN 0004-6361. doi: 10.1051/0004-6361/202243940. URL https://ui.adsabs.harvard.edu/abs/2023A&A...674A...1G.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Score modeling for simulation-based inference. In *NeurIPS 2022 workshop on score-based methods*, 2022.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis*. CRC Press, 2013.

Florian Gerber and Douglas Nychka. Fast covariance parameter estimation of spatial gaussian process models using neural networks. *Stat*, 10(1):e382, 2021.

JK Ghosh. Higher order asymptotics for the likelihood ratio, rao's and wald's tests. *Statistics & probability letters*, 12(6):505–509, 1991.

JK Ghosh and RV Ramamoorthi. Preliminaries and the finite dimensional case. *Bayesian Nonparametrics*, pages 9–55, 2003.

JK Ghosh, BK Sinha, and SN Joshi. Expansions for posterior probability and integrated Bayes risk. *Statistical Decision Theory and Related Topics III*, 1:403–456, 1982.

Stéphane Girard, Armelle Guillou, and Gilles Stupfler. Uniform strong consistency of a frontier estimator using kernel regression on high order moments. *ESAIM: Probability and Statistics*, 18:642–666, 2014.

Sheldon L. Glashow. The renormalizability of vector meson interactions. *Nuclear Physics*, 10:107–117, 1959. ISSN 0029-5582. doi: https://doi.org/10.1016/0029-5582(59)90196-8. URL https://www.sciencedirect.com/science/article/pii/0029558259901968.

Irving John Good. The bayes/non-bayes compromise: A brief review. *Journal of the American Statistical Association*, 87(419):597–606, 1992.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

Michael U. Gutmann and Jukka Corander. Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, 17(125): 1–47, 2016.

W Hardle, Stephan Luckhaus, et al. Uniform consistency of a class of regression function estimators. *The Annals of Statistics*, 12(2):612–623, 1984.

Dieter Heck, Johannes Knapp, JN Capdevielle, G Schatz, T Thouw, et al. Corsika: A monte carlo code to simulate extensive air showers. *Report fzka*, 6019(11), 1998.

Lukas Heinrich. Learning optimal test statistics in the presence of nuisance parameters. *arXiv preprint arXiv:2203.13079*, 2022.

HEP ML Community. A Living Review of Machine Learning for Particle Physics, 2025. URL `https://iml-wg.github.io/HEPML-LivingReview/`.

SW Herb, DC Hom, LM Lederman, JC Sens, HD Snyder, JK Yoh, JA Appel, BC Brown, CN Brown, WR Innes, et al. Observation of a dimuon resonance at 9.5 gev in 400-gev proton-nucleus collisions. *Physical Review Letters*, 39(5):252, 1977.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. *arXiv preprint arXiv:1903.04057*, 2020.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference. *arXiv preprint arXiv:2110.06581*, 2021.

Matthew Ho, Markus Michael Rau, Michelle Ntampaka, Arya Farahi, Hy Trac, and Barnabás Póczos. A robust and efficient deep learning method for dynamical mass measurements of galaxy clusters. *The Astrophysical Journal*, 887(1):25, 2019.

Matthew Ho, Arya Farahi, Markus Michael Rau, and Hy Trac. Approximate bayesian uncertainties on deep learning dynamical mass estimates of galaxy clusters. *The Astrophysical Journal*, 908(2):204, 2021.

Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2): 901–928, 2023.

Benjamin Holzschuh and Nils Thuerey. Flow matching for posterior inference with simulator feedback. *arXiv preprint arXiv:2410.22573*, 2024.

EA Huerta, Asad Khan, Xiaobo Huang, Minyang Tian, Maksim Levental, Ryan Chard, Wei Wei, Maeve Heflin, Daniel S Katz, Volodymyr Kindratenko, et al. Accelerated, scalable and reproducible ai-driven gravitational wave detection. *Nature Astronomy*, 5 (10):1062–1068, 2021.

Željko Ivezić, Steven M Kahn, J Anthony Tyson, Bob Abel, Emily Acosta, Robyn Allsman, David Alonso, Yusra AlSayyad, Scott F Anderson, John Andrew, et al. Lsst: from science drivers to reference design and anticipated data products. *The Astrophysical Journal*, 873 (2):111, 2019.

Rafael Izbicki, Ann Lee, and Chad Schafer. High-Dimensional Density Ratio Estimation with Extensions to Approximate Likelihood Computation. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 420–429, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.

Rafael Izbicki, Ann B Lee, and Peter E Freeman. Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *The Annals of Applied Statistics*, 2017.

Rafael Izbicki, Ann B Lee, and Taylor Pospisil. ABC–CDE: Toward Approximate Bayesian Computation with complex high-dimensional data and limited simulations. *Journal of Computational and Graphical Statistics*, pages 1–20, 2019. doi: 10.1080/10618600.2018.1546594.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

Harold Jeffreys. Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222, 1935. doi: 10.1017/S030500410001330X.

Harold Jeffreys. *Theory of probability*. Clarendon Press Oxford, 3rd ed. edition, 1961.

Joblib Development Team. Joblib: running python functions as pipeline jobs, 2020. URL `https://joblib.readthedocs.io/`.

Adil Jueid, Simone Amoroso, Sascha Caron, Peter Skands, and Roberto Ruiz de austri. Particle spectra from dark matter annihilation: physics modelling and QCD uncertainties. *PoS*, TOOLS2020:028, 2021. doi: 10.22323/1.392.0028.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873): 583–589, 2021.

Marko Järvenpää, Michael U. Gutmann, Aki Vehtari, and Pekka Marttinen. Parallel Gaussian process surrogate Bayesian inference with noisy likelihood evaluations. *Bayesian Anal.*, 16(1):147–178, 2021. doi: 10.1214/20-BA1200.

Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996. doi: 10.1080/01621459.1996.10477003.

Matthäus Kiel, Christopher W O'Dell, Brendan Fisher, Annmarie Eldering, Ray Nassar, Cameron G MacDonald, and Paul O Wennberg. How bias correction goes wrong: Measurement of X_{CO_2} affected by erroneous surface pressure estimates. *Atmospheric Measurement Techniques*, 12(4):2241–2259, 2019.

Jan Kieseler, Giles Chatham Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Preprocessed dataset for "calorimetric measurement of multi-tev muons via deep regression". *URL https://doi. org/10.5281/zenodo*, 5163817, 2021.

Jan Kieseler, Giles C Strong, Filippo Chiandotto, Tommaso Dorigo, and Lukas Layer. Calorimetric measurement of multi-tev muons via deep regression. *The European Physical Journal C*, 82(1):1–26, 2022.

TD Kitching, A Amara, FB Abdalla, B Joachimi, and Alexandre Refregier. Cosmological systematics beyond nuisance parameters: form-filling functions. *Monthly Notices of the Royal Astronomical Society*, 399(4):2107–2128, 2009.

Roger Koenker, Victor Chernozhukov, Xuming He, and Limin Peng. *Handbook of quantile regression*. CRC press, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Sergey E Koposov, C Allende Prieto, A P Cooper, T S Li, L Beraldo e Silva, B Kim, A Carrillo, A Dey, C J Manser, F Nikakhtar, A H Riley, C Rockosi, M Valluri, J Aguilar, S Ahlen, S Bailey, R Blum, D Brooks, T Claybaugh, S Cole, A de la Macorra, B Dey, J E Forero-Romero, E Gaztañaga, J Guy, A Kremin, L Le Guillou, M E Levi, M Manera, A Meisner, R Miquel, J Moustakas, J Nie, N Palanque-Delabrouille, W J Percival, M Rezaie, G Rossi, E Sanchez, E F Schlafly, M Schubnell, G Tarlé, B A Weaver, and Z Zhou. Desi early data release milky way survey value-added catalogue. *Monthly Notices of the Royal Astronomical Society*, 533(1):1012–1031, 07 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae1842. URL `https://doi.org/10.1093/mnras/stae1842`.

Arun Kumar Kuchibhotla and Qinqing Zheng. Near-optimal confidence sequences for bounded random variables. *arXiv preprint arXiv:2006.05022*, 2020.

N. Lagarde, C. Reylé, C. Chiappini, R. Mor, F. Anders, F. Figueras, A. Miglio, M. Romero-Gómez, T. Antoja, N. Cabral, J.-B. Salomon, A. C. Robin, O. Bienaymé, C. Soubiran, D. Cornu, and J. Montillaud. Deciphering the evolution of the Milky Way discs: Gaia APOGEE Kepler giant stars and the Besançon Galaxy Model. *Astronomy and Astrophysics*, 654:A13, October 2021. ISSN 0004-6361. doi: 10.1051/0004-6361/202039982. URL `https://ui.adsabs.harvard.edu/abs/2021A&A...654A..13L/abstract`.

Alexander Laroche and Joshua S Speagle. Closing the stellar labels gap: Stellar label independent evidence for $[\alpha/m]$ information in *Gaia BP/RP* spectra. *arXiv preprint arXiv:2404.07316*, 2024.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. *arXiv preprint arXiv:2302.03026*, 2023.

Amanda Lenzi, Julie Bessac, Johann Rudi, and Michael L Stein. Neural networks for parameter estimation in intractable models. *arXiv preprint arXiv:2107.14346*, 2021.

Yong Li, Jun Yu, and Tao Zeng. Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics*, 216(2):450–493, 2020.

Hannelore Liero. Strong uniform consistency of nonparametric regression function estimates. *Probability theory and related fields*, 82(4):587–614, 1989.

Julia Linhart, Alexandre Gramfort, and Pedro LC Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference. *arXiv preprint arXiv:2306.03580*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Advances in neural information processing systems*, 30, 2017.

Jan-Matthis Lueckmann, Pedro J Goncalves, Giacomo Bassetto, Kaan Öcal, Marcel Nonnenmacher, and Jakob H Macke. Flexible statistical inference for mechanistic models of neural dynamics. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1289–1299. Curran Associates, Inc., 2017.

Jan-Matthis Lueckmann, Giacomo Bassetto, Theofanis Karaletsos, and Jakob H Macke. Likelihood-free inference with emulator networks. In *Symposium on Advances in Approximate Bayesian Inference*, pages 32–53, 2019.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

Louis Lyons. Open statistical issues in Particle Physics. *The Annals of Applied Statistics*, 2 (3):887 – 915, 2008. doi: 10.1214/08-AOAS163.

James G MacKinnon. Bootstrap hypothesis testing. *Handbook of computational econometrics*, 183:213, 2009.

Steven R. Majewski et al. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *The Astronomical Journal*, 154, 2017. ISSN 0004-6256. doi: 10.3847/1538-3881/aa784d. URL `https://ui.adsabs.harvard.edu/abs/2017AJ....154...94M`.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180, 2012.

Jean-Michel Marin, Louis Raynal, Pierre Pudlo, Mathieu Ribatet, and Christian Robert. ABC random forests for Bayesian parameter inference. *Bioinformatics (Oxford, England)*, 35, 05 2016. doi: 10.1093/bioinformatics/bty867.

Luca Masserano. `lf2i`: Likelihood-free frequentist inference. `https://github.com/lee-group-cmu/lf2i`, 2023. URL `https://github.com/lee-group-cmu/lf2i`. GitHub repository.

Luca Masserano, Tommaso Dorigo, Rafael Izbicki, Mikael Kuusela, and Ann Lee. Simulator-based inference with waldo: Confidence regions by leveraging prediction algorithms and posterior estimators for inverse problems. In *International Conference on Artificial Intelligence and Statistics*, pages 2960–2974. PMLR, 2023.

Luca Masserano*, Alexander Shen*, Michele Doro, Tommaso Dorigo, Rafael Izbicki, and Ann Lee. Classification under nuisance parameters and generalized label shift in likelihood-free inference. In *International Conference on Machine Learning*, pages 34987–35012. PMLR, 2024.

Luca Masserano*, James Carzon*, Alexander Shen*, Antonio Herling Ribeiro*, Tommaso Dorigo, Michele Doro, Joshua Speagle, Rafael Izbicki, and Ann Lee. Valid scientific inference with neural density estimators and generative models. *In Submission*, 2025.

Geoffrey J. McLachlan. On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3):318–324, 1987.

Edward Meeds and Max Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint arXiv:1401.2838*, 2014.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (35):983–999, 2006.

Benjamin K Miller, Alex Cole, Patrick Forré, Gilles Louppe, and Christoph Weniger. Truncated marginal neural ratio estimation. *Advances in Neural Information Processing Systems*, 34:129–143, 2021.

I Minchev, F Anders, A Recio-Blanco, C Chiappini, P de Laverny, A Queiroz, M Steinmetz, V Adibekyan, I Carrillo, G Cescutti, G Guiglion, M Hayden, R S de Jong, G Kordopatis, S R Majewski, M Martig, and B X Santiago. Estimating stellar birth radii and the

time evolution of Milky Way's ISM metallicity gradient. *Monthly Notices of the Royal Astronomical Society*, 481(2):1645–1657, 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty2033. URL https://doi.org/10.1093/mnras/sty2033.

Siddharth Mishra-Sharma and Kyle Cranmer. Neural simulation-based inference approach for characterizing the galactic center γ-ray excess. *Physical Review D*, 105(6):063017, 2022.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1): 521–530, 2012.

Saralees Nadarajah, Sergey Bityukov, and Nikolai Krasnikov. Confidence distributions: A review. *Statistical Methodology*, 22:23–46, 2015.

Willie Neiswanger and Aaditya Ramdas. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic learning theory*, pages 963–982. PMLR, 2021.

J. Neyman. On the problem of confidence intervals. *Ann. Math. Statist.*, 6(3):111–116, 09 1935a. doi: 10.1214/aoms/1177732585.

J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937a. ISSN 00804614.

J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. *Biometrika*, 20A(1/2):175–240, 1928. ISSN 00063444.

Jerzy Neyman. On the problem of confidence intervals. *The annals of mathematical statistics*, 6(3):111–116, 1935b.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236(767):333–380, 1937b.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with neural networks. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 388–395. IEEE, 2007.

George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with Bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1028–1036. Curran Associates, Inc., 2016.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Yash Patel, Declan McNamara, Jackson Loper, Jeffrey Regier, and Ambuj Tewari. Variational inference with coverage guarantees in simulation-based inference. *arXiv preprint arXiv:2305.14275*, 2023.

Pratik Patil, Mikael Kuusela, and Jonathan Hobbs. Objective frequentist uncertainty quantification for atmospheric CO_2 retrievals. *SIAM/ASA Journal on Uncertainty Quantification*, 10(3):827–859, 2022. doi: 10.1137/20M1356403.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

Umberto Picchini, Umberto Simola, and Jukka Corander. Adaptive MCMC for synthetic likelihoods and correlated synthetic likelihoods. *arXiv preprint arXiv:2004.04558*, 2020.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Felipe Maia Polo, Rafael Izbicki, Evanildo Gomes Lacerda Jr, Juan Pablo Ibieta-Jimenez, and Renato Vicente. A unified framework for dataset shift diagnostics. *Information Sciences*, page 119612, 2023.

Klaus M Pontoppidan, Jaclyn Barrientes, Claire Blome, Hannah Braun, Matthew Brown, Margaret Carruthers, Dan Coe, Joseph DePasquale, Néstor Espinoza, Macarena Garcia Marin, et al. The jwst early release observations. *The Astrophysical Journal Letters*, 936 (1):L14, 2022.

Alexandre Pouget, Jeffrey M Beck, Wei Ji Ma, and Peter E Latham. Probabilistic brains: knowns and unknowns. *Nature neuroscience*, 16(9):1170–1178, 2013.

John W Pratt. Length of confidence intervals. *Journal of the American Statistical Association*, 56(295):549–567, 1961.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

X. Qian, A. Tan, J.J. Ling, Y. Nakajima, and C. Zhang. The Gaussian CL_s method for searches of new physics. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 827(35):63–78, 2016.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Stefan T. Radev, Ulf K. Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2020. doi: 10.1109/TNNLS.2020.3042395.

Stefan T Radev, Marvin Schmitt, Valentin Pratz, Umberto Picchini, Ullrich Köthe, and Paul-Christian Bürkner. Jana: Jointly amortized neural approximation of complex bayesian models. In *Uncertainty in Artificial Intelligence*, pages 1695–1706. PMLR, 2023a.

Stefan T Radev, Marvin Schmitt, Lukas Schumacher, Lasse Elsemüller, Valentin Pratz, Yannik Schälte, Ullrich Köthe, and Paul-Christian Bürkner. Bayesflow: Amortized bayesian workflows with neural networks. *arXiv preprint arXiv:2306.16015*, 2023b.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv preprint arXiv:2410.23614*, 2024.

Richard Redner. Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics*, 9(1):225–228, 1981.

Shahzar Rizvi, Mariel Pettee, and Benjamin Nachman. Learning likelihood ratios with neural network classifiers. *arXiv preprint arXiv:2305.10500*, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1): 21–41, 2002.

Matthew Sainsbury-Dale, Andrew Zammit-Mangion, and Raphaël Huser. Likelihood-free parameter estimation with neural bayes estimators. *The American Statistician*, 78(1): 1–14, 2024.

Abdus Salam. Weak and electromagnetic interactions. *Il Nuovo Cimento (1955-1965)*, 11: 568–577, 1959. URL `https://api.semanticscholar.org/CorpusID:15889731`.

Chad M Schafer and Philip B Stark. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104(487):1080–1089, 2009.

S J Schmidt, A I Malz, J Y H Soo, I A Almosallam, M Brescia, S Cavuoti, J Cohen-Tanugi, A J Connolly, J DeRose, P E Freeman, M L Graham, K G Iyer, M J Jarvis, J B Kalmbach, E Kovacs, A B Lee, G Longo, C B Morrison, J A Newman, E Nourbakhsh, E Nuss, T Pospisil, H Tranin, R H Wechsler, R Zhou, R Izbicki, and (The LSST Dark Energy Science Collaboration). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society*, 499(2):1587–1606, 09 2020. ISSN 0035-8711. doi: 10.1093/mnras/staa2799.

Marvin Schmitt, Valentin Pratz, Ullrich Köthe, Paul-Christian Bürkner, and Stefan Radev. Consistency models for scalable and fast simulation-based inference. *Advances in Neural Information Processing Systems*, 37:126908–126945, 2024.

Tore Schweder and Nils Lid Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002.

Catia Scricciolo. Probability matching priors: A review. *Journal of the Italian Statistical Society*, 8:83–100, 1999. doi: 10.1007/BF03178943.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

Alistair A Sellar, Colin G Jones, Jane P Mulcahy, Yongming Tang, Andrew Yool, Andy Wiltshire, Fiona M O'connor, Marc Stringer, Richard Hill, Julien Palmieri, et al. Ukesm1: Description and evaluation of the uk earth system model. *Journal of Advances in Modeling Earth Systems*, 11(12):4513–4558, 2019.

Bodhisattva Sen, Matthew Walker, and Michael Woodroofe. On the unified method with nuisance parameters. *Statistica Sinica*, 19(1):301–314, 2009. ISSN 10170405, 19968507.

Louis Sharrock, Jack Simons, Song Liu, and Mark Beaumont. Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models. *arXiv preprint arXiv:2210.04872*, 2022.

Umberto Simola, Jessi Cisewski-Kehe, Michael U Gutmann, and Jukka Corander. Adaptive Approximate Bayesian Computation tolerance selection. *Bayesian analysis*, 16(2):397–423, 2021.

Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of Approximate Bayesian Computation*. Chapman and Hall/CRC, 2018.

Michael F Skrutskie, RM Cutri, R Stiening, MD Weinberg, S Schneider, JM Carpenter, Capps Beichman, R Capps, T Chester, J Elias, et al. The two micron all sky survey (2mass). *The Astronomical Journal*, 131(2):1163, 2006.

Dongyuan Song, Qingyang Wang, Guanao Yan, Tianyang Liu, Tianyi Sun, and Jingyi Jessica Li. scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology*, pages 1–6, 2023.

Joshua S Speagle, Catherine Zucker, Angus Beane, Phillip A Cargile, Aaron Dotter, Douglas P Finkbeiner, Gregory M Green, Benjamin D Johnson, Edward F Schlafly, Ana Bonaca, et al. Deriving stellar properties, distances, and reddenings using photometry and astrometry with brutus. *arXiv preprint arXiv:2503.02227*, 2025.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Amos Storkey et al. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, 30(3-28):6, 2009.

Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11:341–359, 1997.

Mikael Sunnåker, Alberto Giovanni Busetto, Elina Numminen, Jukka Corander, Matthieu Foll, and Christophe Dessimoz. Approximate bayesian computation. *PLoS computational biology*, 9(1):e1002803, 2013.

Hyungsuk Tak, Yang Chen, Vinay L Kashyap, Kaisey S Mandel, Xiao-Li Meng, Aneta Siemiginowska, and David A van Dyk. Six maxims of statistical acumen for astronomical data analysis. *The Astrophysical Journal Supplement Series*, 275(2):30, 2024.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Alvaro Tejero-Cantero, Jan Boelts, Michael Deistler, Jan-Matthis Lueckmann, Conor Durkan, Pedro J. Gonçalves, David S. Greenberg, and Jakob H. Macke. `sbi`: A toolkit for simulation-based inference. *Journal of Open Source Software*, 5(52):2505, 2020. doi: 10.21105/joss.02505. URL `https://doi.org/10.21105/joss.02505`.

Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Anal.*, 2021. doi: 10.1214/20-BA1238. Advance publication.

Suzanne Thornton and Min-ge Xie. Bridging bayesian, frequentist and fiducial inferences using confidence distributions. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 106–131. Chapman and Hall/CRC, 2024.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate Bayesian Computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

undark lab. Swyft: A system for scientific simulation-based inference at scale., 2023. URL `https://github.com/undark-lab/swyft`.

W van den Boom, G Reeves, and D B Dunson. Approximating posteriors with high-dimensional nuisance parameters via integrated rotated Gaussian approximation. *Biometrika*, Aug 2020. ISSN 1464-3510. doi: 10.1093/biomet/asaa068.

Afonso Fernandes Vaz, Rafael Izbicki, and Rafael Bassi Stern. Quantification under prior probability shift: The ratio estimator and its extensions. *The Journal of Machine Learning Research*, 20(1):2921–2953, 2019.

Valérie Ventura. Bootstrap tests of hypotheses. In *Analysis of parallel spike trains*, pages 383–398. Springer, 2010.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17 (3):261–272, 2020.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005a.

Vladimir Vovk, Ivan Petej, and Valentina Fedorova. From conformal to probabilistic prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 221–230. Springer, 2014.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pages 23–39. Springer, 2016.

Vladimir Vovk et al. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005b.

Julia Walchessen, Amanda Lenzi, and Mikael Kuusela. Neural likelihood surfaces for spatial processes with computationally intensive or intractable likelihoods. *arXiv preprint arXiv:2305.04634*, 2023.

Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54 (3):426–482, 1943.

Bingjie Wang, Joel Leja, V Ashley Villar, and Joshua S Speagle. Sbi++: Flexible, ultra-fast likelihood-free inference customized for astronomical applications. *The Astrophysical Journal Letters*, 952(1):L10, 2023.

David J Warne, Oliver J Maclaren, Elliot J Carr, Matthew J Simpson, and Christopher Drovandi. Generalised likelihood profiles for models with intractable likelihoods. *Statistics and Computing*, 34(1):50, 2024.

Larry Wasserman. Frasian inference. *Statistical Science*, 26(3):322–325, 2011.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33:20204–20214, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86 (1):1–27, 2024.

Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967. doi: 10.1103/PhysRevLett.19.1264.

Jonas Wildberger, Maximilian Dax, Simon Buchholz, Stephen Green, Jakob H Macke, and Bernhard Schölkopf. Flow matching for scalable simulation-based inference. *Advances in Neural Information Processing Systems*, 36, 2024.

Richard Wilkinson. Accelerating ABC methods using Gaussian processes. In *Artificial Intelligence and Statistics*, pages 1015–1023, 2014.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9(1):60–62, 03 1938. doi: 10.1214/aoms/1177732360.

Simon Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–4, 08 2010. doi: 10.1038/nature09319.

Simon Wood. Package 'mgcv'. *R package version*, 1(29):729, 2015.

Min-ge Xie and Kesar Singh. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39, 2013.

Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

Chaoyu Yu and Peter D Hoff. Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, 105(2):319–335, 2018.

David Zhao, Niccolò Dalmasso, Rafael Izbicki, and Ann B Lee. Diagnostics for conditional density models and bayesian inference algorithms. In *Uncertainty in Artificial Intelligence*, pages 1830–1840. PMLR, 2021.

Yunzhang Zhu, Xiaotong Shen, and Wei Pan. On high-dimensional constrained maximum likelihood inference. *Journal of the American Statistical Association*, 115(529):217–230, 2020.

A

Additional Results for Chapter 2

A.1 Estimating Odds

Algorithm A.1 shows how to create the training set $\mathcal{T}$ for estimating odds. Out of the total number of simulations B, a proportion p is generated by the stochastic forward simulator F_θ at different parameter values θ, while the rest is sampled from a reference distribution G. Note that G can be any distribution that dominates F_θ. If G is the marginal distribution F_x and $n = 1$, then computations for `BFF` are simplified because its denominator equals one. Algorithm A.2 shows how to sample from the marginal distribution F_x. In practice, if the data is pre-simulated, one can sample from the (empirical) marginal using permutations to break the relationship between θ and X for $X \sim G = F_x$.

Algorithm A.1 Generate a labeled sample of size B for estimating odds

Input: simulator F_θ; reference distribution G; proposal distribution π_Θ over parameter space; number of simulations B; parameter p of Bernoulli distribution
Output: labeled training sample $\mathcal{T}$

Set $\mathcal{T} \leftarrow \varnothing$
for i in $\{1, ..., B\}$ **do**
    Draw parameter value $\theta_i \sim \pi_\Theta$
    Draw $Y_i \sim \text{Bernoulli}(p)$
    if $Y_i == 1$ **then**
        Draw sample $X_i \sim F_{\theta_i}$
    else
        Draw sample $X_i \sim G$
    $\mathcal{T} \leftarrow \mathcal{T} \cup (\theta_i, X_i, Y_i)$
return $\mathcal{T} = \{\theta_i, X_i, Y_i\}_{i=1}^B$

A.2 Estimating p-values

Given observed data D and a test statistic λ, we can compute p-values $p(D; \theta_0) := \mathbb{P}_{\mathcal{D}|\theta_0}(\lambda(\mathcal{D}; \theta_0) < \lambda(D; \theta_0))$ for each hypothesis $H_{0,\theta_0} : \theta = \theta_0$. Algorithm A.3 describes

Algorithm A.2 Sample from the marginal distribution $G = F_X$

Input: simulator F_θ; proposal distribution π_Θ over parameter space
Output: sample X_i from the marginal distribution F_X

1: Draw parameter value $\theta_i \sim \pi_\Theta$
2: Draw sample $X_i \sim F_{\theta_i}$
3: **return** X_i

how to estimate such p-values for all $\theta_0 \in \Theta$ simultaneously.

A.3 Constructing Confidence Sets

Algorithm A.4 details the construction of LF2I confidence sets with `ACORE` and `BFF` as defined in Section 2.3 (the algorithm based on p-value estimation is analogous). Algorithm A.5 details the construction of the (hybrid) `ACORE` and `BFF` confidence sets defined in Section 2.5 for the general setting with nuisance parameters. Note that the first chunk on estimating the odds and the last chunk with Neyman inversion are the same for `ACORE` and `BFF`. Furthermore, the test statistics are the same whether or not there are nuisance parameters.

A.4 Theoretical Guarantees of Power for `ACORE` with Calibrated Critical Values

Next, we show for finite Θ that as long as the probabilistic classifier is consistent and the critical values are well estimated (which holds for large enough B' according to Theorem A.4), the power of the `ACORE` test converges to the power of the LRT as B grows.

Algorithm A.3 Estimate p-values $p(D;\theta_0)$ given observed data D for a level-α test of $H_{0,\theta_0} : \theta = \theta_0$ vs. $H_{1,\theta_0} : \theta \neq \theta_0$, for all $\theta_0 \in \Theta$ simultaneously.

Input: observed data D; simulator F_θ; number of simulations B'; π_Θ (fixed proposal distribution over the parameter space Θ); test statistic λ; probabilistic classifier
Output: estimated p-value $\hat{p}(D;\theta)$ for all $\theta = \theta_0 \in \Theta$

1: Set $\mathcal{T}' \leftarrow \varnothing$
2: **for** i in $\{1, \ldots, B'\}$ **do**
3: Draw parameter $\theta_i \sim \pi_\Theta$
4: Draw sample $X_{i,1}, \ldots, X_{i,n} \overset{\text{iid}}{\sim} F_{\theta_i}$
5: Compute test statistic $\lambda_i \leftarrow \lambda((X_{i,1}, \ldots, X_{i,n}); \theta_i)$
6: Compute indicator $Z_i \leftarrow \mathbb{1}(\lambda_i < \lambda(D;\theta_i))$
7: $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(\theta_i, Z_i)\}$
8: Use $\mathcal{T}'$ to learn the p-value function $\hat{p}(D;\theta)$ using Z as the label for each θ
9: **return** $\hat{p}(D;\theta_0)$

Algorithm A.4 Construct $(1-\alpha)$ confidence set for θ (no nuisance parameters)

Input: simulator F_θ; proposal distribution π over Θ; parameter p of Bernoulli; number of simulations B (test statistic); number of simulations B' (critical values); probabilistic classifier; observations $D = \{x_i^{\text{obs}}\}_{i=1}^n$; level $\alpha \in (0,1)$; size of evaluation grid over parameter space n_{grid}; test statistic λ (ACORE or BFF)

Output: θ evaluation points in confidence set $\widehat{\mathcal{R}}(D)$

// **Estimate odds**
Generate labeled sample $\mathcal{T}$ according to Algorithm A.1
Learn $\widehat{\mathbb{P}}(Y = 1 \mid \theta, X)$ on $\mathcal{T}$ with a probabilistic classifier, for all $\theta \in \Theta$ and $X \in \mathcal{X}$
Let the estimated odds $\widehat{\mathbb{O}}(X;\theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,X)}{\widehat{\mathbb{P}}(Y=0|\theta,X)}$

// **Compute critical values for ACORE or BFF**
if $\lambda ==$ ACORE **then**
    Let $\lambda(\mathcal{D};\theta) \leftarrow \widehat{\Lambda}(\mathcal{D};\theta)$ be the ACORE statistic (Equation (2.8)) with estimated odds
else if $\lambda ==$ BFF **then**
    Let $\lambda(\mathcal{D};\theta) \leftarrow \hat{\tau}(\mathcal{D};\theta)$ be the BFF statistic (Equation (2.10)) with estimated odds
Learn critical values $\widehat{C}_\theta$ according to Algorithm 2.1

// **Confidence sets for θ via Neyman inversion**
Initialize confidence set $\widehat{\mathcal{R}}(D) \leftarrow \varnothing$
Let L_Θ be a lattice over Θ with n_{grid} elements
for $\theta_0 \in \mathrm{L}_\Theta$ **do**
    if $\lambda(D;\theta_0) \geqslant \widehat{C}_{\theta_0}$ **then**
        $\widehat{\mathcal{R}}(D) \leftarrow \widehat{\mathcal{R}}(D) \cup \{\theta_0\}$
return confidence set $\widehat{\mathcal{R}}(D)$

Theorem A.1. *For each $C \in \mathbb{R}$, let $\widehat{\phi}_{B,C}(\mathcal{D})$ be the test based on the ACORE statistic $\widehat{\Lambda}_B$ with critical value C[1] for a number of simulations B in Algorithm A.1. Moreover, let $\phi_C(\mathcal{D})$ be the likelihood ratio test with critical value C. If, for every $\theta \in \Theta$, the probabilistic classifier is such that*

$$\widehat{\mathbb{P}}(Y = 1 \mid \theta, X) \xrightarrow[B\longrightarrow\infty]{P} \mathbb{P}(Y = 1 \mid \theta, X),$$

where $|\Theta| < \infty$, and $\widehat{C}_B$ is chosen such that $\widehat{C}_B \xrightarrow[B\longrightarrow\infty]{D} C$ for a given $C \in \mathbb{R}$, then, for every $\theta \in \Theta$,

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D}) = 1\right) \xrightarrow[B\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}mid\theta}\left(\phi_C(\mathcal{D}) = 1\right).$$

Proof. Because $\widehat{\mathbb{P}}(Y = 1 \mid \theta, X) \xrightarrow[B\longrightarrow\infty]{P} \mathbb{P}(Y = 1 \mid \theta, X)$, it follows directly from the properties of convergence in probability that for every $\theta_0, \theta_1 \in \Theta$

$$\sum_{i=1}^{n} \log\left(\widehat{\mathbb{OR}}(X_i^{\text{obs}};\theta_0,\theta_1)\right) \xrightarrow[B\longrightarrow\infty]{P} \sum_{i=1}^{n} \log\left(\mathbb{OR}(X_i^{\text{obs}};\theta_0,\theta_1)\right).$$

[1] That is, $\widehat{\phi}_{B,C}(\mathcal{D}) = 1 \iff \widehat{\Lambda}_B(\mathcal{D};\Theta_0) < C$.

Algorithm A.5 Construct confidence set for μ with (approximate) coverage $1-\alpha$ under the presence of nuisance parameters

Input: simulator F_θ; proposal distribution π over $\Theta = \mathcal{M} \times \mathcal{N}$; parameter p of Bernoulli; number of simulations B (test statistic); number of simulations B' (critical values); probabilistic classifier; observations $D = \{x_i^{\text{obs}}\}_{i=1}^n$; level $\alpha \in (0,1)$; size of evaluation grid over parameter space, n_{grid}; test statistic λ (ACORE or BFF)

Output: μ evaluation points in confidence set $\widehat{\mathcal{R}}(D)$

// **Estimate odds**
Generate labeled sample $\mathcal{T}$ according to Algorithm A.1
Learn $\widehat{\mathbb{P}}(Y = 1 \mid \theta, X)$ on $\mathcal{T}$ with a probabilistic classifier, $\forall\ \theta = (\mu, \nu) \in \Theta, X \in \mathcal{X}$
Let the estimated odds $\widehat{\mathbb{O}}(X;\theta) \leftarrow \frac{\widehat{\mathbb{P}}(Y=1|\theta,X)}{\widehat{\mathbb{P}}(Y=0|\theta,X)}$

// **Compute (hybrid) critical values for h-ACORE or h-BFF**
if $\lambda ==$ ACORE **then**
    Let $\widehat{\nu}_\mu \leftarrow \arg\max_{\nu \in \mathcal{N}} \prod_{i=1}^n \widehat{\mathbb{O}}(x_i^{\text{obs}}; (\mu, \nu))$ for every μ
    Let $\lambda(\mathcal{D}; \mu) \leftarrow \widehat{\Lambda}\left(\mathcal{D}; (\mu, \widehat{\nu}_\mu)\right)$ be ACORE (Equation (2.8)) with estimated odds
    Generate $\mathcal{T}'$ as in Algorithm 2.1 using the proposal $\pi'((\mu,\nu)) \propto \pi(\mu) \times \delta_{\widehat{\nu}_\mu}(\nu)$
    Learn $\widehat{C}_\mu = \widehat{F}^{-1}_{\lambda(\mathcal{D};\mu)|(\mu,\widehat{\nu}_\mu)}(\alpha)$ for every μ as in Algorithm 2.1 using $\mathcal{T}'$
else if $\lambda ==$ BFF **then**
    Let $\pi_{\mathcal{N}}(\nu)$ be the restriction of proposal distribution π over $\mathcal{N}$
    Let $\lambda(\mathcal{D}; \mu) \leftarrow \widehat{\tau}(\mathcal{D}; \mu)$ be the BFF statistic (Equation (2.10)) with estimated odds
    Learn $\widehat{C}_\mu = \widehat{F}^{-1}_{\lambda(\mathcal{D};\mu)|(\mu)}(\alpha)$ for every μ (no ν) as in Algorithm 2.1

// **Confidence sets for μ via Neyman inversion**
Initialize confidence set $\widehat{\mathcal{R}}(D) \leftarrow \varnothing$
Let $\mathrm{L}_{\mathcal{M}}$ be a lattice over $\mathcal{M}$ with n_{grid} elements
for $\mu_0 \in \mathrm{L}_{\mathcal{M}}$ **do**
    if $\lambda(D; \mu_0) \geqslant \widehat{C}_{\mu_0}$ **then**
        $\widehat{\mathcal{R}}(D) \leftarrow \widehat{\mathcal{R}}(D) \cup \{\mu_0\}$
return confidence set $\widehat{\mathcal{R}}(D)$

The continuous mapping theorem implies that

$$\widehat{\Lambda}_B(\mathcal{D}; \Theta_0) \xrightarrow[B \longrightarrow \infty]{P} \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^n \log\left(\mathbb{OR}(X_i^{\text{obs}}; \theta_0, \theta_1)\right),$$

and therefore $\widehat{\Lambda}_B(\mathcal{D}; \Theta_0)$ converges in distribution to $\sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^n \log\left(\mathbb{OR}(X_i^{\text{obs}}; \theta_0, \theta_1)\right)$. Now, from Slutsky's theorem,

$$\widehat{\Lambda}_B(\mathcal{D}; \Theta_0) - \widehat{C}_B \xrightarrow[B \longrightarrow \infty]{D} \sup_{\theta_0 \in \Theta_0} \inf_{\theta_1 \in \Theta} \sum_{i=1}^n \log\left(\mathbb{OR}(X_i^{\text{obs}}; \theta_0, \theta_1)\right) - C.$$

It follows that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\widehat{\phi}_{B,\widehat{C}_B}(\mathcal{D})=1\right)=\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}\left(\hat{\Lambda}_B(\mathcal{D};\Theta_0)-\widehat{C}_B\leqslant 0\right)\\
&\xrightarrow[B\longrightarrow\infty]{}\mathbb{P}_{\mathcal{D}|\theta}\Big(\sup_{\theta_0\in\Theta_0}\inf_{\theta_1\in\Theta}\sum_{i=1}^{n}\log\left(\mathbb{OR}(X_i^{\text{obs}};\theta_0,\theta_1)\right)-C\leqslant 0\Big)\\
&=\mathbb{P}_{\mathcal{D}|\theta}\left(\phi_C(\mathcal{D})=1\right),
\end{aligned}$$

where the last equality follows from Proposition 2.1. □

A.5 Analysis of Critical Values for Experiments 2.6.1 and 2.6.2

In this section we visualize how critical values vary across the parameter space Θ for the experiments of Sections 2.6.1 and 2.6.2. Figure A.1 compares critical values for the exact LRT of the Gaussian Mixture Model (GMM) example, where the distribution of the test statistic is unknown, using three different methods:

i) The first approach is to compute cutoffs via Monte Carlo (MC) simulations at fixed values of θ. These critical values can be considered the "ground truth", since for this one-dimensional example we were able to use a high-resolution grid and large batches at each grid point. Unfortunately, MC quickly becomes infeasible if the dimensionality of the parameter space increases. In addition, a scientist cannot adopt MC samples in practical settings, where one only has access to a pre-determined data set and not to the simulator itself.

ii) The second approach is to assume that the cutoff is (asymptotically) constant across the parameter space. Here we have computed cutoffs assuming that Wilks' theorem holds and that the limiting distribution is a χ^2-distribution, which is not the case. Indeed, the bottom central panel of Figure 2.3 shows that the χ^2-approximation achieves correct coverage only when $\theta = 0$ (i.e., when the GMM collapses to one Gaussian).

iii) The third approach is to compute the critical values of the (known) test statistic via quantile regression (QR). With a very small calibration set (0.1% of the total simulations used for the MC approach), QR is able to approximate the quantile surface and achieve nominal coverage for all values of θ (see Figure 2.3).

Figure A.2 shows similar results for the HEP example of Section 2.6.2; here we visualize the the critical values of `h-ACORE` (estimated via LF2I) as a function of the parameter of interest μ and the nuisance parameter ν. Again, we see evidence that the quantile surface is far from being constant, and that the test statistic is not pivotal. Hence, there is a need for a quantile regression that adapts to the varying distribution of the test statistic.

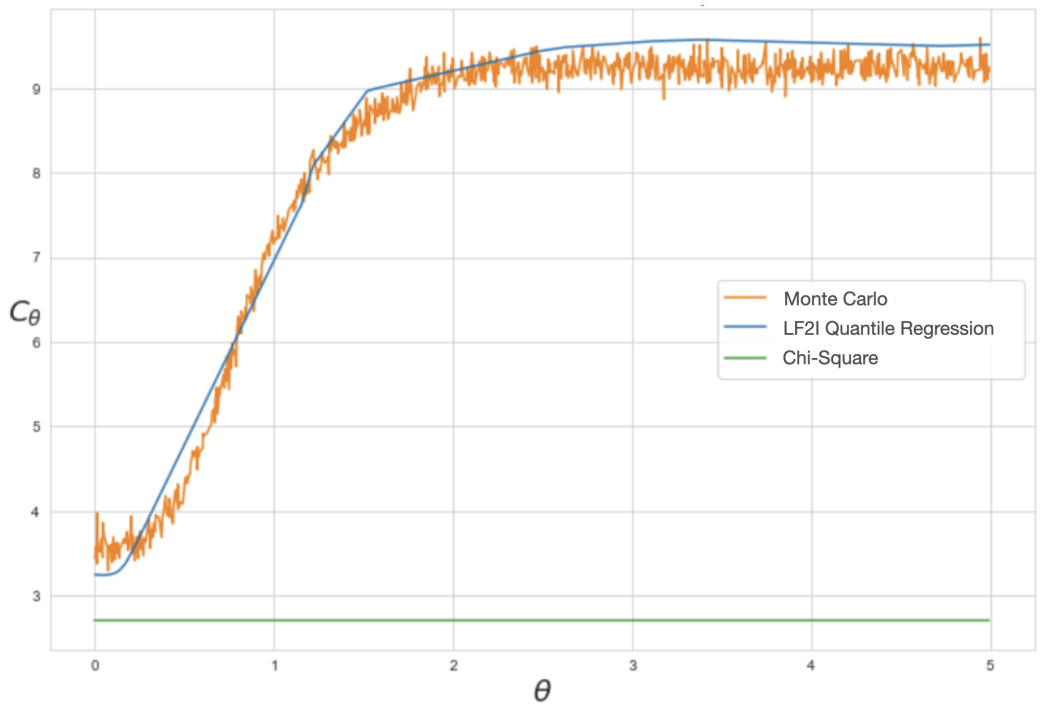


Figure A.1: **Comparison of critical values** obtained via Monte Carlo, the Chi-Square asymptotic assumption of Wilks' Theorem, and LF2I Quantile Regression, for the GMM example of Section 2.6.1.

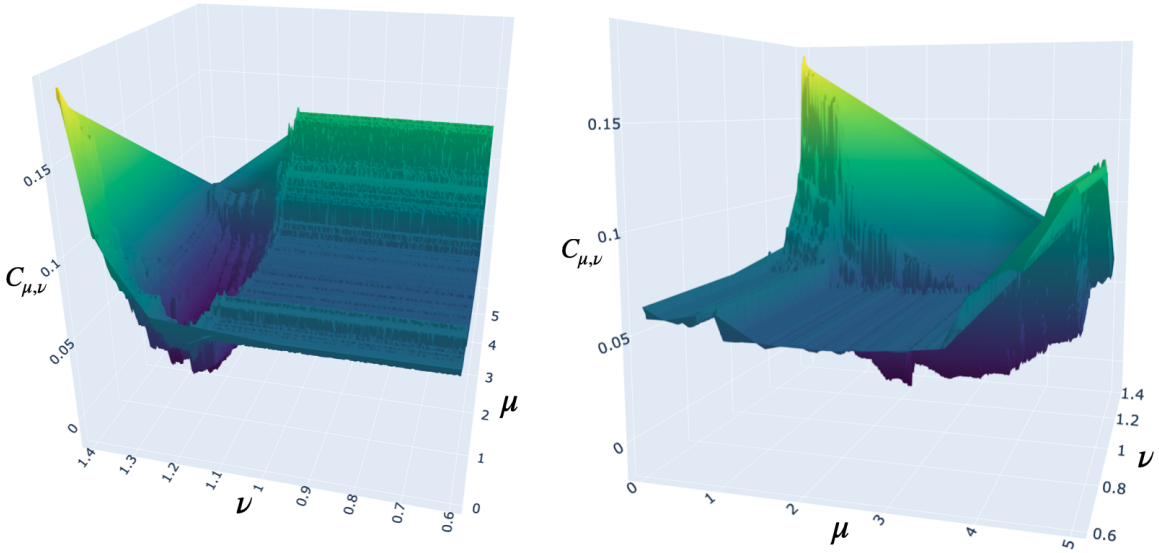


Figure A.2: **Critical values of** `h-ACORE` **estimated via quantile regression** as a function of the parameter of interest μ and the nuisance parameter ν, for the example of Section 2.6.2. The figures show the same 2D surface from two different angles.

A.6 Additional Proofs

Proof of Proposition 2.1. Because the measure ν dominates F_θ, G also dominates F_θ. Let $f(x \mid \theta)$ be the density of F_θ with respect to G. By Bayes rule,

$$\mathbb{O}(x;\theta) := \frac{\mathbb{P}(Y=1 \mid \theta, x)}{\mathbb{P}(Y=0 \mid \theta, x)} = \frac{f(x \mid \theta)p}{(1-p)}.$$

If $\widehat{\mathbb{P}}(Y=1 \mid \theta, x) = \mathbb{P}(Y=1 \mid \theta, x)$, then $\widehat{\mathbb{O}}(x;\theta_0) = \mathbb{O}(x;\theta_0)$. Therefore,

$$\begin{aligned}
\widehat{\tau}(\mathcal{D};\Theta_0) :=& \frac{\int_{\Theta_0} \prod_{i=1}^n \widehat{\mathbb{O}}(X_i^{\text{obs}};\theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \widehat{\mathbb{O}}(X_i^{\text{obs}};\theta)\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n \mathbb{O}(X_i^{\text{obs}};\theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \mathbb{O}(X_i^{\text{obs}};\theta)\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n \frac{f(X_i^{\text{obs}}|\theta)p}{(1-p)}\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n \frac{f(X_i^{\text{obs}}|\theta)p}{(1-p)}\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n f(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n f(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_1(\theta)}
\end{aligned}$$

Moreover, the chain rule implies that $f(x \mid \theta) = p(x \mid \theta)h(x)$, where $h(x) := \frac{\mathrm{d}\nu}{\mathrm{d}G}(x)$. It follows that

$$\begin{aligned}
\widehat{\tau}(\mathcal{D};\Theta_0) =& \frac{\int_{\Theta_0} \prod_{i=1}^n f(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n f(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n p(X_i^{\text{obs}} \mid \theta)h(X_i^{\text{obs}})\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n p(X_i^{\text{obs}} \mid \theta)h(X_i^{\text{obs}})\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \prod_{i=1}^n p(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \prod_{i=1}^n p(X_i^{\text{obs}} \mid \theta)\mathrm{d}\pi_1(\theta)} \\
=& \frac{\int_{\Theta_0} \mathcal{L}(\mathcal{D};\theta)\mathrm{d}\pi_0(\theta)}{\int_{\Theta_1} \mathcal{L}(\mathcal{D};\theta)\mathrm{d}\pi_1(\theta)} \\
=& \mathrm{BF}(\mathcal{D};\Theta_0).
\end{aligned}$$

□

Proof of Theorem 2.3. By definition, for all fixed $c_{B'}$, $\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leqslant c_{B'}) = F(c_{B'} \mid \theta_0)$. It follows that the random variable $\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leqslant C_{B'}) = F(C_{B'} \mid \theta_0)$. Moreover, by construction, $\alpha = \widehat{F}_{B'}(C_{B'} \mid \theta_0)$. It follows that

$$\begin{aligned}
|\mathbb{P}_{\mathcal{D}|\theta_0,C_{B'}}(\lambda(\mathcal{D};\theta_0) \leqslant C_{B'}) - \alpha| &= |F(C_{B'} \mid \theta_0) - \alpha| \\
&= |F(C_{B'} \mid \theta_0) - \widehat{F}_{B'}(C_{B'} \mid \theta_0)| \\
&\leqslant \sup_{\lambda \in \mathbb{R}} |F(\lambda \mid \theta_0) - \widehat{F}_{B'}(\lambda \mid \theta_0)| \xrightarrow[B' \longrightarrow \infty]{P} 0.
\end{aligned}$$

The result follows from the fact that convergence in probability to a constant implies almost sure convergence. □

Proof of Theorem 2.5. The proof follows from applying the convergence rate to the last equation in the proof of Theorem 2.3. □

Assumption A.2 (Uniform consistency in θ and λ)**.** *Let $\widehat{F}_{B'}(\cdot \mid \theta)$ be the estimated cumulative distribution function of the test statistic $\lambda(\mathcal{D}; \Theta_0)$ conditional on θ based on a sample $\mathcal{T}'$ with size B' implied by the quantile regression, and let $F(\cdot \mid \theta)$ be its true distribution given θ. Assume that the quantile regression estimator is such that*

$$\sup_{\theta \in \Theta_0, \lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda \mid \theta) - F(\lambda \mid \theta)| \xrightarrow[B' \longrightarrow \infty]{P} 0.$$

This assumption holds, for instance, for quantile regression forests (Meinshausen, 2006) under additional assumptions (see Proposition A.3).

Proposition A.3. *If, for every $\theta \in \Theta_0$, the quantile regression estimator is such that*

$$\sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda \mid \theta) - F(\lambda \mid \theta)| \xrightarrow[B' \longrightarrow \infty]{P} 0 \tag{A.1}$$

and either

- $|\Theta| < \infty$ *or,*
- Θ *is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta) = \sup_{t \in \mathbb{R}} \mid \widehat{F}_{B'}(t \mid \theta) - F(t \mid \theta)|$ is almost surely continuous in θ and strictly decreasing in B',*

then Assumption A.2 holds.

Proof. If $|\Theta| < \infty$, the union bound and Equation A.1 imply that

$$\sup_{\theta \in \Theta_0} \sup_{\lambda \in \mathbb{R}} |\widehat{F}_{B'}(\lambda|\theta) - F(\lambda|\theta)| \xrightarrow[B' \longrightarrow \infty]{P} 0. \tag{A.2}$$

Similarly, by Dini's theorem, Equation A.2 also holds if Θ is a compact subset of $\mathbb{R}^d$, and the function $g_{B'}(\theta)$ is continuous in θ and strictly decreasing in B'. □

Theorem A.4. *Let $C_{B'} \in \mathbb{R}$ be the critical value of the test based on a absolutely continuous statistic $\lambda(\mathcal{D}; \Theta_0)$ chosen according to Algorithm 2.1 for a fixed $\alpha \in (0, 1)$. If the quantile estimator satisfies Assumption A.2, then*

$$C_{B'} \xrightarrow[B' \longrightarrow \infty]{P} C^*,$$

where C^* is such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_{\mathcal{D}|\theta}(\lambda(\mathcal{D}; \Theta_0) \leqslant C^*) = \alpha.$$

Proof. Assumption A.2 implies that

$$\sup_{\theta\in\Theta_0} |\widehat{F}^{-1}_{B'}(\alpha \mid \theta) - F^{-1}(\alpha \mid \theta)| \xrightarrow[B'\longrightarrow\infty]{P} 0.$$

The result then follows from the fact that

$$\begin{aligned} 0 \leqslant |C_{B'} - C^*| &= |\sup_{\theta\in\Theta_0} \widehat{F}^{-1}_{B'}(\alpha \mid \theta) - \sup_{\theta\in\Theta_0} F^{-1}(\alpha \mid \theta)| \\ &\leqslant \sup_{\theta\in\Theta_0} |\widehat{F}^{-1}_{B'}(\alpha \mid \theta) - F^{-1}(\alpha \mid \theta)|, \end{aligned}$$

and thus

$$|C_{B'} - C^*| \xrightarrow[B'\longrightarrow\infty]{P} 0.$$

□

Lemma A.5. *Let $g_1, g_2, \ldots$ be a sequence of random functions such that $g_i : \mathcal{Z} \longrightarrow \mathbb{R}$, and let Z be a random quantity defined over $\mathcal{Z}$, independent of the random functions. Assume that $g(Z)$ is absolutely continuous with respect to the Lebesgue measure. If, for every $z \in \mathcal{Z}$,*

$$g_m(z) \xrightarrow[m\longrightarrow\infty]{a.s.} g(z),$$

then

$$g_m(Z) \xrightarrow[m\longrightarrow\infty]{\mathcal{L}} g(Z).$$

Proof. Fix $y \in \mathbb{R}$ and let $A_y = \{z \in \mathcal{Z} : g(z) \neq y\}$. Notice that $\mathbb{P}(Z \in A_y) = 1$. Moreover, almost sure convergence of $g_m(z)$ implies its convergence in distribution. It follows that for every $z \in A_y$,

$$\lim_m \mathbb{P}(g_m(z) \leqslant y) = \mathbb{P}\left(g(z) \leqslant y\right). \tag{A.3}$$

Now, using Equation (A.3) and Lebesgue's dominated convergence theorem, notice that

$$\begin{aligned} \lim_m \mathbb{P}(g_m(Z) < y) &= \lim_m \int_{\mathcal{Z}} \mathbb{P}(g_m(Z) < y \mid Z = z)\mathrm{d}\mathbb{P}_Z(z) \\ &= \int_{\mathcal{Z}} \lim_m \mathbb{P}(g_m(Z) < y \mid Z = z)\mathrm{d}\mathbb{P}_Z(z) = \int_{A_z} \lim_m \mathbb{P}(g_m(z) < y)\mathrm{d}\mathbb{P}_Z(z) \\ &= \int_{A_z} \mathbb{P}(g(z) < y)\mathrm{d}\mathbb{P}_Z(z) = \int_{\mathcal{Z}} \mathbb{P}(g(Z) < y \mid Z = z)\mathrm{d}\mathbb{P}_Z(z) \\ &= \mathbb{P}(g(Z) < y), \end{aligned}$$

which concludes the proof. □

Proof of Theorem 2.7. Assumption 2.6 implies that, for every D,

$$\begin{aligned} 0 \leqslant |\hat{p}(D;\Theta_0) - p(D;\Theta_0)| &= |\sup_{\theta\in\Theta_0} \hat{p}(D;\theta) - \sup_{\theta\in\Theta_0} p(D;\theta)| \\ &\leqslant \sup_{\theta\in\Theta_0} |\hat{p}(D;\theta) - p(D;\theta)| \xrightarrow[B'\longrightarrow\infty]{\text{a.s.}} 0, \end{aligned}$$

and therefore $\hat{p}(D;\Theta_0)$ converges almost surely to $p(D;\Theta_0)$. It follows from Lemma A.5 that $\hat{p}(\mathcal{D};\Theta_0)$ converges in distribution to $p(\mathcal{D};\Theta_0)$. We then conclude that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\hat{p}(\mathcal{D};\Theta_0) \leqslant \alpha) = F_{\hat{p}(\mathcal{D};\Theta_0)|\theta}(\alpha) \xrightarrow[B'\longrightarrow\infty]{} F_{p(\mathcal{D};\Theta_0)|\theta}(\alpha) = \mathbb{P}_{\mathcal{D}|\theta}(p(D;\Theta_0) \leqslant \alpha),$$

where F_Z denotes the cumulative distribution function of the random variable Z. □

Proof of Corollary 2.8. Fix $\theta \in \Theta$. Because F_θ is continuous, the definition of $p(\mathcal{D};\theta)$ implies that its distribution is uniform under the null. Thus $\mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\theta) \leqslant \alpha) = \alpha$. Theorem 2.7 therefore implies that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\hat{p}(\mathcal{D};\theta) \leqslant \alpha) \xrightarrow[B'\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\theta) \leqslant \alpha) = \alpha. \tag{A.4}$$

Now, for any $\theta \in \Theta_0$, uniformity of the p-value implies that

$$\mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\Theta_0) \leqslant \alpha) = \mathbb{P}_{\mathcal{D}|\theta}\left(\sup_{\theta_0\in\Theta_0} p(\mathcal{D};\theta_0) \leqslant \alpha\right) \leqslant \mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\theta) \leqslant \alpha) = \alpha.$$

Conclude from Theorem 2.7 that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}'|\theta}(\hat{p}(\mathcal{D};\Theta_0) \leqslant \alpha) \xrightarrow[B'\longrightarrow\infty]{} \mathbb{P}_{\mathcal{D}|\theta}(p(\mathcal{D};\Theta_0) \leqslant \alpha) \leqslant \alpha. \tag{A.5}$$

The conclusion follows from putting together Equations (A.4) and (A.5). □

Proof of Theorem 2.10.

$$\begin{aligned}
|\hat{p}(D;\Theta_0) - p(D;\Theta_0)| &= |\sup_{\theta\in\Theta_0} \hat{p}(D;\theta) - \sup_{\theta\in\Theta_0} p(D;\theta)| \\
&\leqslant \sup_{\theta\in\Theta_0} |\hat{p}(D;\theta) - p(D;\theta)| \\
&= O_P\left(\left(\frac{1}{B'}\right)^r\right),
\end{aligned}$$

where the last line follows from Assumption 2.9 □

Lemma A.6. *Under Assumption 2.11, for every $\theta, \theta_0 \in \Theta$*

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}\left[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|\right] \leqslant M^2 \int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0))^2 \mathrm{d}G(x).$$

Proof. For every $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] &= \left(\int |\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)| \ \mathrm{d}F(x \mid \theta)\right)^2 \\
&= \left(\int |\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0)| \ \mathrm{d}F(x \mid \theta)\right)^2 \\
&= \left(\int |\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0)| \mathbb{O}(x;\theta) \mathrm{d}G(x)\right)^2 \\
&\leqslant \left(\int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0)^2 \mathrm{d}G(x)\right) \left(\int \mathbb{O}^2(x;\theta) \mathrm{d}G(x)\right),
\end{aligned}$$

where the last inequality follows from Cauchy-Schwarz. Assumption 2.11 implies that

$$\int \mathbb{O}^2(x;\theta)\mathrm{d}G(x) \leqslant M^2,$$

from which we conclude that

$$\mathbb{E}^2_{\mathcal{D}|\theta,T}[|\tau(\mathcal{D};\theta_0) - \widehat{\tau}_B(\mathcal{D};\theta_0)|] \leqslant M^2 \int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0))^2 \mathrm{d}G(x).$$

□

Lemma A.7. *For fixed $c \in \mathbb{R}$, let $\phi_{\tau;\theta_0}(\mathcal{D}) = \mathbb{1}\left(\tau(\mathcal{D};\theta_0) < c\right)$ and $\phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D}) = \mathbb{1}(\widehat{\tau}_B(\mathcal{D};\theta_0) < c)$ be the testing procedures for testing $H_{0,\theta_0} : \theta = \theta_0$ obtained using τ and $\widehat{\tau}_B$. Under Assumptions 2.11-2.12, for every $0 < \epsilon < 1$,*

$$\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau}_B;\theta_0}(\mathcal{D})) \leqslant \frac{2MC_L \cdot \sqrt{\int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0))^2 \mathrm{d}G(x)}}{\epsilon} + \epsilon.$$

Proof of Lemma A.7. It follows from Markov's inequality and Lemma A.6 that with probability at least $1-\epsilon$, $\mathcal{D}$ is such that

$$|\tau(\mathcal{D};\theta_0) - \widehat{\tau}(\mathcal{D};\theta_0)| \leqslant \frac{M \cdot \sqrt{\int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0))^2 \mathrm{d}G(x)}}{\epsilon}. \tag{A.6}$$

Now we upper bound $\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau};\theta_0}(\mathcal{D}))$. Define A as the event that Equation (A.6) happens and let $h(\theta_0) := \int (\mathbb{O}(x;\theta_0) - \widehat{\mathbb{O}}(x;\theta_0))^2 \mathrm{d}G(x)$. Then:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau};\theta_0}(\mathcal{D})) &\leqslant \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau};\theta_0}(\mathcal{D}), A) + \mathbb{P}_{\theta}(A^c) \\
&\leqslant \mathbb{P}_{\mathcal{D}|\theta,T}\left(\mathbb{1}\left(\tau(\mathcal{D};\theta_0) < c\right) \neq \mathbb{1}\left(\widehat{\tau}(\mathcal{D};\theta_0) < c\right), A\right) + \epsilon \\
&\leqslant \mathbb{P}_{\mathcal{D}|\theta,T}\left(c - \frac{M\sqrt{h(\theta_0)}}{\epsilon} < \tau(\mathcal{D};\theta_0) < c + \frac{M\sqrt{h(\theta_0)}}{\epsilon}\right) + \epsilon.
\end{aligned}$$

Assumption 2.12 then implies that

$$\mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\widehat{\tau};\theta_0}(\mathcal{D})) \leqslant \frac{K' \cdot \sqrt{h(\theta_0)}}{\epsilon} + \epsilon$$

where $K\prime = 2MC_L$, which concludes the proof. □

Proof of Theorem 2.13. Follows directly from Lemma A.7 and Jensen's inequality. □

Lemma A.8. *Under Assumptions 2.11-2.15, there exists $C > 0$ such that*

$$\mathbb{E}_{\mathcal{T}}\left[L(\widehat{\mathbb{O}}, \mathbb{O})\right] \leqslant CB^{-\kappa/((\kappa+d+p))}.$$

Proof. Let $\hat{p} = \widehat{\mathbb{P}}(Y = 1 \mid x, \theta)$ and $p = \mathbb{P}(Y = 1 \mid x, \theta)$ be the probabilistic classifier and true classification function, respectively, on the training sample $\mathcal{T}$. Let $h(y) = \frac{y}{1-y}$ for $0 < y < 1$. A Taylor expansion of h implies that

$$(h(\hat{p}) - h(p))^2 = (h(p) + R_1(\hat{p}) - h(p))^2 = R_1(\hat{p})^2,$$

where $R_1(\hat{p}) = h'(\xi)(\hat{p} - p)$ for some ξ between p and $\hat{p}$. Also note that due to Assumption 2.11,

$$\exists a > 0 \text{ s.t. } p, \hat{p} > a, \ \forall x \in \mathcal{X}, \theta \in \Theta.$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{T}}\left[\iint (h(\hat{p}) - h(p))^2 \, \mathrm{d}G(x)\mathrm{d}\pi(\theta)\right] \\
&= \mathbb{E}_{\mathcal{T}}\left[\iint \frac{1}{(1-\xi)^4} (\hat{p} - p)^2 \, \mathrm{d}G(x)\mathrm{d}\pi(\theta)\right] \\
&\leqslant \frac{1}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\iint (\hat{p} - p)^2 \, \mathrm{d}G(x)\mathrm{d}\pi(\theta)\right] \\
&= \frac{1}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y = 1 \mid x, \theta) - \mathbb{P}(Y = 1 \mid x, \theta)\right)^2 h'(x, \theta)\mathrm{d}H(x, \theta)\right] \\
&\leqslant \frac{\gamma}{(1-a)^4}\mathbb{E}_{\mathcal{T}}\left[\int \left(\widehat{\mathbb{P}}(Y = 1 \mid x, \theta) - \mathbb{P}(Y = 1 \mid x, \theta)\right)^2 \mathrm{d}H(x, \theta)\right] \\
&= \mathcal{O}\left(B^{-\kappa/(\kappa+d+p)}\right).
\end{aligned}$$

□

Proof of Theorem 2.16. It follows from Theorem 2.13 that

$$\begin{aligned}
\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\pi(\theta_0) &= \mathbb{E}_{\mathcal{T}}\left[\int \mathbb{P}_{\mathcal{D}|\theta,T}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\pi(\theta_0)\right] \\
&\leqslant \frac{2MC_L \cdot \mathbb{E}_{\mathcal{T}}\left[\sqrt{L(\widehat{\mathbb{O}}, \mathbb{O})}\right]}{\epsilon} + \epsilon \\
&\leqslant \frac{2MC_L \cdot \sqrt{\mathbb{E}_{\mathcal{T}}\left[L(\widehat{\mathbb{O}}, \mathbb{O})\right]}}{\epsilon} + \epsilon,
\end{aligned}$$

where the last step follows from Jensen's inequality. It follows from this and Lemma A.8 that

$$\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\pi(\theta_0) \leqslant \frac{KB^{-\kappa/(2(\kappa+d+p))}}{\epsilon} + \epsilon,$$

where $K = 2MC_L\sqrt{C}$. Notice that taking $\epsilon^* = \sqrt{K}B^{-\kappa/(4(\kappa+d+p))}$ optimizes the bound and gives the result. □

Proof of Corollary 2.17. The result follows from noticing that

$$\mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}) = 1) \geqslant \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1) - \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\hat{\tau}_B;\theta_0}(\mathcal{D})),$$

and therefore

$$\begin{aligned}\int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}) = 1)\mathrm{d}\theta_0 &\geqslant \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1)\mathrm{d}\theta_0 - \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) \neq \phi_{\hat{\tau}_B;\theta_0}(\mathcal{D}))\mathrm{d}\theta_0 \\ &\geqslant \int \mathbb{P}_{\mathcal{D},\mathcal{T}|\theta}(\phi_{\tau;\theta_0}(\mathcal{D}) = 1)\mathrm{d}\theta_0 - K' B^{-\kappa/(4(\kappa+d+p))},\end{aligned}$$

where the last inequality follows from Theorem 2.16. □

A.7 Loss Functions

In this work, we use the cross-entropy loss to train probabilistic classifiers. Consider a sample point $\{\theta, x, y\}$ generated according to Algorithm A.1. Let p be a Bernoulli(y) distribution, and q be a Bernoulli $\left(\widehat{\mathbb{P}}(Y = 1 \mid \theta, x)\right)$ = Bernoulli $\left(\frac{\widehat{\mathbb{O}}(x;\theta)}{1+\widehat{\mathbb{O}}(x;\theta)}\right)$ distribution. The *cross-entropy* between p and q is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}}(\widehat{\mathbb{O}}; \{\theta, x, y\}) &= -y \log\left(\frac{\widehat{\mathbb{O}}(x;\theta)}{1+\widehat{\mathbb{O}}(x;\theta)}\right) - (1-y)\log\left(\frac{1}{1+\widehat{\mathbb{O}}(x;\theta)}\right) \\ &= -y \log\left(\widehat{\mathbb{O}}(x;\theta)\right) + \log\left(1+\widehat{\mathbb{O}}(x;\theta)\right). \qquad \text{(A.7)}\end{aligned}$$

For every x and θ, the expected cross-entropy $\mathbb{E}[L_{\mathrm{CE}}(\widehat{\mathbb{O}}; \{\theta, x, y\})]$ is minimized by $\widehat{\mathbb{O}}(x;\theta) = \mathbb{O}(x;\theta)$. If the probabilistic classifier attains the minimum of the cross-entropy loss, then the estimated `ACORE` statistic $\widehat{\Lambda}(\mathcal{D}; \Theta_0)$ will be equal to the likelihood ratio statistic in Equation (2.5), as shown in Dalmasso et al. (2020). Similarly, as stated in Proposition 2.1, at the minimum, the estimated `BFF` statistic $\widehat{\tau}(\mathcal{D}; \Theta_0)$ is equal to the Bayes factor in Equation (2.6).

B

Additional Results for Chapter 3

B.1 Additional Experiments

B.1.1 Property III: Estimating the Conditional Variance Matters

We complete the exposition of the statistical properties of `Waldo` (Section 3.3.3) by demonstrating the importance of estimating the conditional variance in the test statistic $\tau_{\texttt{Waldo}}$. Recall that in principle any test statistic defined in an LFI setting could be used for our framework. One could then define a simpler "unstandardized" test statistic $\tau_{\texttt{Waldo-novar}}(\mathcal{D}; \theta_0) = (\mathbb{E}[\theta \mid \mathcal{D}] - \theta_0)^T (\mathbb{E}[\theta \mid \mathcal{D}] - \theta_0)$ which does not require estimation of $\mathbb{V}[\theta \mid \mathcal{D}]$. It turns out that estimating $\mathbb{V}[\theta \mid \mathcal{D}]$ and using $\tau_{\texttt{Waldo}}$ is actually of crucial importance, as it leads to confidence regions of smaller or equal expected size, especially

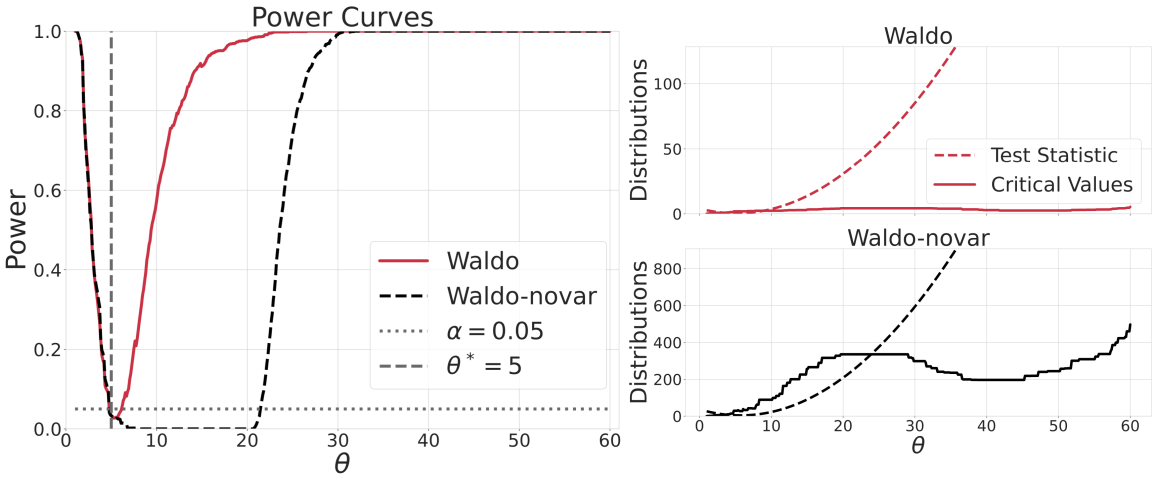
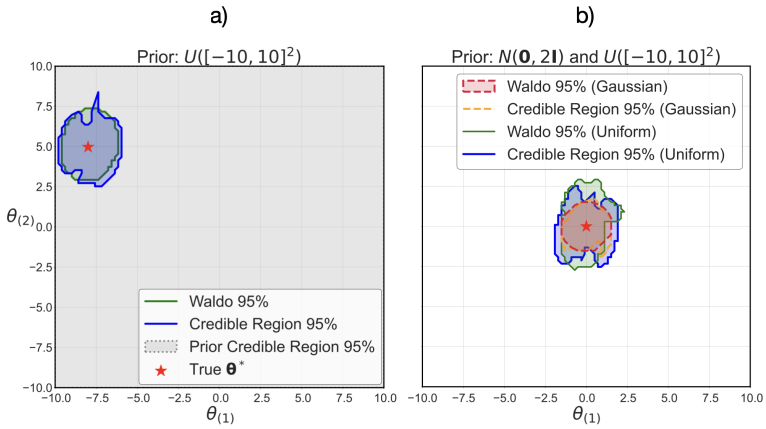


Figure B.1: **Property III: Estimating the conditional variance matters. Left:** Power curves at 95% confidence level when the true Pareto shape $\theta^* = 5$, implying a very skewed data distribution. **Right:** Test statistics and critical values as a function of θ. In this example, we set $n = 10$.



Method	Area (Uniform Prior)	Area (Gaussian Prior
WALDO	2.71% ± 0.11%	2.10% ± 0.34%
Credible Regions	2.75% ± 0.08%	2.12% ± 0.18%

Figure B.2: **a) When the prior is uninformative, `Waldo` can still correct for possible approximation errors in the estimated posterior. b)-c) When the prior is consistent with the data, `Waldo` tightens the confidence sets, improving the precision with respect to the case using a Uniform prior. a) and b)** Posterior credible regions and `Waldo` confidence sets using different priors. **c)** Average area of credible regions and `Waldo` confidence sets across 100 independent samples, reported as the percentage of points retained among those in the evaluation grid.

in settings where the conditional variance varies significantly as a function of θ. Consider, for example, the problem of estimating the shape of a Pareto distribution with fixed scale $x_{\min} = 1$ and true unknown shape $\theta^* = 5$, which yields a strongly right-skewed data distribution. Figure B.1 shows that $\tau_{\texttt{Waldo}}$ has much higher power than $\tau_{\texttt{Waldo-novar}}$ for inferring θ. Dividing by the conditional variance effectively stabilizes the test statistic and makes its distribution over $\mathcal{D}$ pivotal, i.e., independent of θ. This implies that the critical values will be relatively constant over θ (see top right panel for `Waldo`), which yields tighter parameter regions due to the curvature of the test statistic.

B.1.2 Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture

The results of Figure 3.5 in Chapter 3 showed that `Waldo` is able to leverage an estimated posterior to construct conditionally valid confidence regions, even when the prior is at odds with the data. On the other side, when no prior information is available, it is common to sample θ according to a uniform distribution over the parameter space. In this case, we observe that confidence sets and posterior credible regions largely overlap. Nonetheless, if the latter happen to suffer from approximation errors, as is common for neural posteriors in high dimensions, this could hinder the statistical reliability of the estimated region. `Waldo` can correct even for this problem and guarantee conditional coverage, as we can see from panel *a)* in Figure B.2.

Figure B.3 shows the output of the diagnostics procedure when using a uniform prior to train the posterior estimator (compare with Figure 3.5, right column, in Chapter 3, which used a Gaussian prior). We achieve correct conditional coverage for `Waldo` but not for

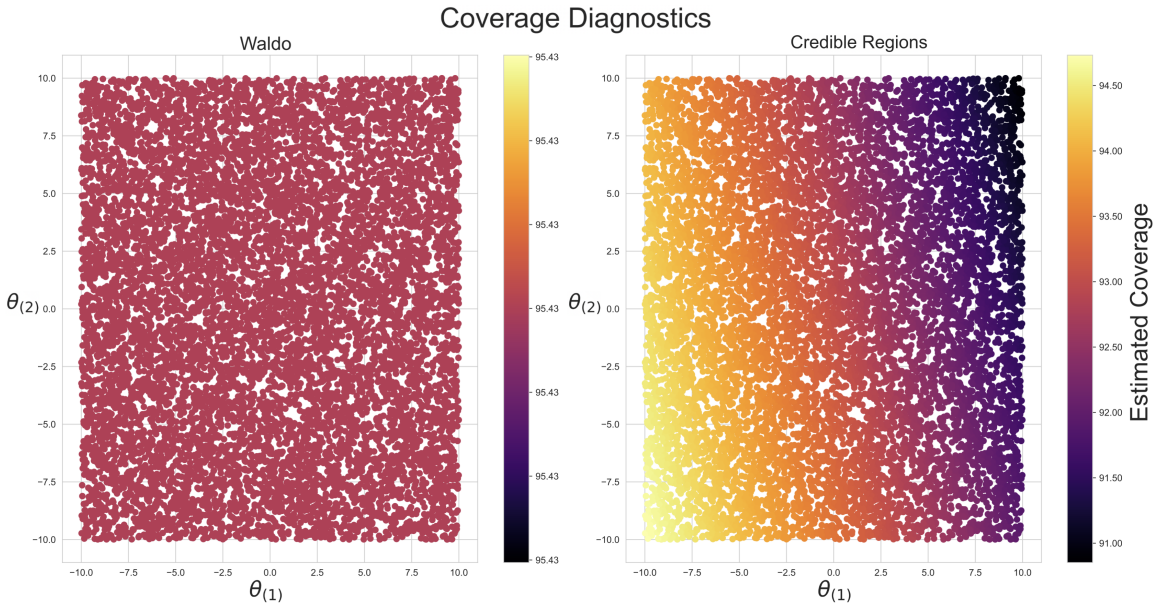


Figure B.3: **Coverage diagnostics for Gaussian mixture model example with uniform prior**. We achieve correct conditional coverage for `Waldo` (**left**) but not for credible regions (**right**) even though the prior is is uniform, due to estimation and approximation errors, which `Waldo` can correct via recalibration.

credible regions even though the prior is is uniform, due to estimation and approximation errors in the posterior, which `Waldo` can correct using quantile regression to calibrate the test statistics.

B.1.3 Confidence Sets for Muon Energies using CNN Predictions

Figure B.4 compares confidence sets and prediction sets for the full calorimeter data, showing clearly the bias in the prediction sets and the correction applied by Waldo. These results explain the observed patterns in Figure 3.6 in Chapter 3: prediction sets are centered around the point prediction, which is downward biased at high energies, mainly due to the nonlinearity of the response at high energies.

B.2 Details on Models, Training, and Computational Resources

B.2.1 Synthetic Examples for Statistical Properties

See Section 3.3.3 in Chapter 3 and Appendix B.1.1 for descriptions of the experiments. For **Property I** and **Property II**, we used the implementation of local linear regression available in Seabold and Perktold (2010) to estimate conditional mean and conditional variance within a prediction setting, with $B = 20{,}000$. For **Property III**, instead, we used a simple feedforward neural network with one hidden layer and $B = 50{,}000$. In all

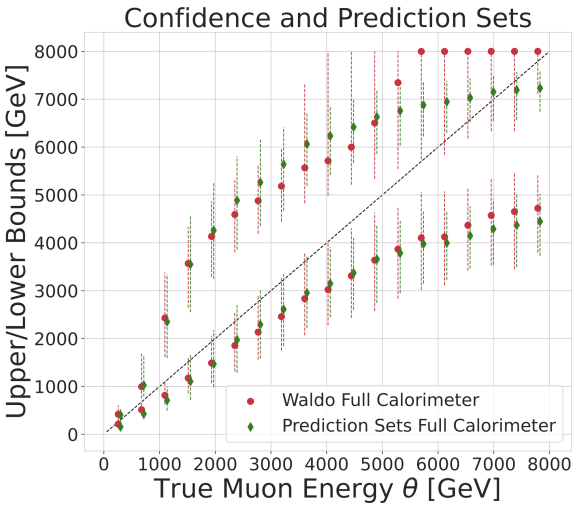


Figure B.4: **Confidence and prediction sets for the muon energy reconstruction experiment**. Boxplots of the upper and lower bounds of prediction sets (green) versus `Waldo` confidence sets (red) for full the calorimeter data, all divided in 19 bins over true energy. We clearly see the bias occurring in the prediction sets (especially at high energies) and the correction applied by `Waldo`.

cases, for quantile regression we used quantile gradient boosted trees as implemented in `scikit-learn` (Pedregosa et al., 2011), with $B' = 20{,}000$ for **Property I** and **Property II**, and $B' = 50{,}000$ for **Property III**. All models were trained on a MacBook Pro M1Pro (CPU only).

B.2.2 Synthetic Example for Computational Properties

See Section 3.3.4 in Chapter 3 for a description of the experiment. To compute the test statistic $\tau_{\texttt{Waldo}}$, we approximated conditional mean and conditional variance through a posterior distribution estimated via normalizing flows (Tejero-Cantero et al., 2020), with $B = 20{,}000$ for $d = 1$ and $B = 200{,}000$ for $d = 10$. To construct the confidence sets, critical values were then estimated both via quantile regression using quantile gradient boosted trees as implemented in `scikit-learn` (Pedregosa et al., 2011) with varying values of B', and via Monte Carlo by simulating many times for each θ and retaining the $(1-\alpha)$ quantile of the computed test statistics. The evaluation set was made of 1,000 samples over $\Theta = [-1, 1]^d$. To make the comparison fair, if quantile regression used $B' = 50{,}000$, then Monte Carlo had access to 50 simulations for each of the 1,000 samples in the evaluation set. The estimated coverage probability for both methods was then estimated using the implementation of Generalized Additive Models (GAMs) with thin plate splines available in the `MGCV` package (Wood, 2015) of `R`, with $B'' = 30{,}000$.

B.2.3 Confidence Sets from Neural Posteriors: Two-Dimensional Gaussian Mixture

See Section 3.4.1 in Chapter 3 and Appendix B.1.2 for descriptions of the experiments and details on the algorithms and sample sizes used. Training was done on a MacBook Pro M1Pro (CPU only); it took approximately 15-20 minutes to train the posterior estimator, and an additional $\approx$2 minutes for the quantile neural network to estimate the critical values. Note that the latter step requires computing the conditional mean, the conditional variance and the Waldo statistic over all sample points in $\mathcal{T}'$. The posterior was sampled multiple times for each $X \in \mathcal{T}'$ to approximate $\mathbb{E}(\theta \mid X)$ and $\mathbb{V}(\theta \mid X)$ via Monte Carlo; this procedure took a total of $\approx$45 minutes (but could potentially be optimized through vectorizations in the future).

B.2.4 Confidence Sets for Muon Energies using CNN Predictions

See Section 3.4.2 and Appendix B.1.3 for descriptions of the experiment and details on the algorithms and sample sizes used. We collected 886,716 simulated muons in total; roughly 200,000 muons were used to estimate the critical values, $\approx$ 24,000 muons to construct the final confidence sets and diagnostics, and the rest was used to estimate the conditional mean and variance via the custom 3D CNN from Kieseler et al. (2022). Training the latter CNN took approximately 20 hours for the conditional mean and another 20 hours for the conditional variance, using an NVIDIA V100 GPU on an Azure cloud computing machine. Estimating the critical values via quantile gradient boosted trees in `scikit-learn` (Pedregosa et al., 2011) took approximately 2 minutes.

C

Additional Results for Chapter 4

C.1 Relation to Other Methodology

Classical statistical inference and approximate likelihood methods. Our approach builds on the classical construction of confidence sets via inversion of hypothesis tests, which dates back to Neyman's seminal work (Neyman, 1935b). While this method has a long-standing tradition in scientific inference, it initially required tractable likelihoods and closed-form critical values, limiting its applicability. More recent advancements, especially within high-energy physics (HEP), have extended the Neyman construction to likelihood-free inference (LFI) scenarios (Feldman and Cousins, 1998; Cowan et al., 2011b; Cranmer, 2015; Schafer and Stark, 2009). These pioneering efforts highlighted critical open problems, such as efficiently constructing Neyman confidence sets in general settings, evaluating coverage without prohibitive computational costs, and effectively implementing hybrid statistical techniques (Cousins, 2006, 2018). Building upon these foundations, several recent machine-learning-based techniques approximate the likelihood-ratio test (LRT) statistic and rely on asymptotic χ^2 cutoffs to form confidence sets (Cranmer et al., 2015). While these approaches have shown promising performance in fields like HEP, they struggle with small-sample sizes or irregularities introduced by complex likelihoods (Algeri et al., 2019) and the use of neural density estimators.

To overcome these limitations, Dalmasso et al. (2020) developed `ACORE`, a method that estimates LRT cutoffs directly without resorting to asymptotic approximations, improving performance in limited-data scenarios. Subsequently, Dalmasso* et al. (2024) proposed LF2I, a flexible framework generalizing Neyman's inversion for likelihood-free inference and suitable for any test statistic, thereby opening the way for the usage of a wide array of machine learning methods to obtain confidence sets with frequentist guarantees. Within this framework, they introduced `BFF`, which leverages the Bayes Factor as a frequentist test statistic. In contrast, this chapter exploits highest-posterior-density regions derived directly from estimates of posterior distributions: not only this allows to take advantage of recent advancements in the AI literature that are now popular in LFI, but it also enables domain scientists to construct valid *and* optimal (i.e., as small as possible under suitable conditions; cfr. Section C.3) confidence sets. More traditional techniques in the LFI literature that are

based on posterior estimates usually fall under Approximate Bayesian Computation (ABC) methods. While they have been very popular in different scientific fields — see for example Beaumont and Rannala (2004), Beaumont (2010) and Sunnåker et al. (2013) — they do not guarantee validity nor optimality of the resulting credible regions.

Bayesian SBI and Conformal Inference Recent advancements in SBI have primarily come from cross-pollination with the machine learning literature Cranmer et al. (2020); Bürkner et al. (2025). Several works have proposed learning algorithms that leverage novel neural density estimators such as normalizing flows (e.g., Papamakarios and Murray 2016; Lueckmann et al. 2017; Greenberg et al. 2019; Miller et al. 2021; Radev et al. 2023a), diffusion models (e.g., Geffner et al. 2022; Sharrock et al. 2022; Linhart et al. 2023), flow matching (e.g., Wildberger et al. 2024; Holzschuh and Thuerey 2024) and consistency models (e.g., Schmitt et al. 2024). These methods are enabling a revolution in the inference capabilities available to domain scientists, but unfortunately they are not equipped with the necessary statistical guarantees required by the rigor of the scientific method, as it has been shown by Hermans et al. (2021) and Dalmasso* et al. (2024). The work of Delaunoy et al. (2022) successfully alleviates this issue by enforcing a balancing condition that yields more conservative posteriors, resulting in highest-posterior-density regions with approximate *expected* coverage. Nonetheless, a posterior estimator that largely under-covers in some regions of the parameter space and largely over-covers in other regions would still be considered valid under this notion of *marginal* coverage. Our work targets the stronger notion of validity defined in Equation (1.1), which ensures (conditional) coverage *point-wise* across the entire parameter space.

Besides SBI-specific techniques, conformal methods have also become extremely popular in the machine learning community and beyond. Although conformal methods were originally developed for predictive problems, they can also enhance the marginal coverage properties of approximate Bayesian methods (see, e.g., Baragatti et al. 2024 and Patel et al. 2023). However, they do not guarantee frequentist (conditional) coverage across all parameter values.

Inference based on predictions: `WALDO` and Prediction-Powered Inference Several studies have employed predictions methods on simulated datasets for inference on real observations, often without incorporating the necessary corrections to ensure valid uncertainty quantification (e.g., Dorigo et al. (2022); Gerber and Nychka (2021); Ho et al. (2021)). To address this issue, Masserano et al. (2023) introduced `WALDO`, a method that can take predictions from any machine learning algorithm and transform them into confidence sets with frequentist guarantees. Our approach differs in that we estimate the full posterior distribution from simulated data rather than just point predictions, allowing us to derive confidence sets that are typically smaller and more accurate than those obtained through `WALDO`, particularly in cases where the posterior is multimodal or asymmetric.

Prediction-powered inference (Angelopoulos et al., 2023a) has also emerged as a promising framework that leverages both labeled training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and additional unlabeled covariates $X_{n+1}, \ldots, X_{n+m}$ to enhance inference. However, this approach funda-

mentally differs from our setting, as its primary goal is to infer global parameters characterizing the data-generating process of the entire set, rather than constructing confidence sets for individual instances.

Bridging Bayesian and frequentist approaches. The interplay between Bayesian and frequentist methodologies has been explored in various contexts. Good (1992) proposed using the Bayes Factor as a frequentist test statistic, but only in scenarios where likelihoods are tractable. Similarly, Pratt (1961), Yu and Hoff (2018) and Hoff (2023) showed that, when the likelihood is available, confidence sets derived from posterior distributions tend to be more efficient (in terms of expected volume) than those based purely on likelihood ratios. Our work extends these results to LFI settings, where likelihoods are intractable and confidence sets are constructed from posterior estimates obtained via generative models.

In addition, Wasserman (2011) and Fong and Holmes (2021) showed that conformal inference can be applied to Bayesian models to construct prediction sets with valid frequentist coverage. Concretely, in that setting, one models the Bayesian predictive distribution $Y_{n+1} \mid x_{n+1}, (x_n, y_n), \ldots, (x_1, y_1)$ starting from a statistical model for $Y \mid \theta, X$. However, as already mentioned in the previous paragraph, conformal methods only guarantee *marginal* coverage over θ, which does not imply valid confidence sets for every parameter value. As a result, conformal procedures that exhibit sever under-coverage in some regions and strong over-coverage in others might still satisfy conformal guarantees, but would fail within our setting. In contrast, our method provides confidence sets that maintain validity point-wise across the entire parameter space, offering stronger guarantees for inference in scientific settings where one has to ensure the reliability of conclusions regardless of the specific source that generated an observation.

C.2 Constructing Confidence Procedures with Frequentist Coverage

Notation and problem set-up. Our assumption (well borne by the fundamental science use cases that we target) is that labeled data encode *the same* physical process as target data. Hence, we also assume that the likelihood function $\mathcal{L}(\theta; x) = p(x \mid \theta)$ with $\theta \in \Theta$ and $x \in \mathcal{X}$, which describes the data-generating process, is the same for train and target data. We refer to the label distribution $\pi(\theta)$ on the train data as our *prior distribution*. The *reference distribution* $r(\theta)$ on the universal set is a distribution that dominates the prior distribution, $r \gg \pi$. The prior $\pi(\theta)$ can be different from the label distribution $p_{\text{obs}}(\theta)$ of the target data, as well as different from the reference distribution $r(\theta)$ of the universal set. See Section 4.3.1 for our experimental set-up.

Now let $p(x) := \int \mathcal{L}(\theta; x)\pi(\theta)\mathrm{d}\theta$ be the marginal probability density function of X on train data. Our *posterior distribution* is then defined as $\pi(\theta \mid x) := \mathcal{L}(\theta; x)\pi(\theta)/p(x)$; that is, the posterior is the conditional density of θ given x on train data.

Definition C.1 (Confidence procedure)**.** *Let $\mathcal{A}$ denote the space of all measurable sets, $\mathcal{A} \subseteq \mathcal{X} \times \Theta$. A* confidence procedure *is a set $\mathbf{C}$ in the space $\mathcal{A}$ defined as*

$$\{(x, \theta) : (x, \theta) \in \mathbf{C}\}.$$

For fixed x, we define the confidence set or θ-section as

$$C(x) = \{\theta : (x, \theta) \in \mathbf{C}\}.$$

For fixed θ, we define the acceptance region or x-section as

$$C_\theta = \{x : (x, \theta) \in \mathbf{C}\}.$$

A $(1 - \alpha)$ confidence procedure is valid *if, for every $\theta \in \Theta$ and every miscoverage level $0 \leqslant \alpha \leqslant 1$,*

$$\mathbb{P}_{X|\theta} \left(\theta \in C(X)\right) \geqslant 1 - \alpha.$$

C.2.1 Fast Construction of Confidence Procedures from Posterior Estimates

Let $\widehat{\pi}(\theta \mid X)$ be a posterior approximation based on the train data

$$\mathcal{T}_{\text{train}} = \{(\theta_1, X_1) \ldots (\theta_B, X_B)\} \sim \pi(\theta)\mathcal{L}(\theta; x).$$

Once we have $\widehat{\pi}(\theta \mid X)$, it is straightforward to construct Bayesian credible regions for fixed x by computing high-posterior density (HPD) level sets

$$H_c(x) \coloneqq \{\theta : \widehat{\pi}(\theta \mid x) > c\}, \tag{C.1}$$

where $\int_{H_c(x)} \widehat{\pi}(\theta \mid x) \mathrm{d}\theta = 1 - \alpha$. These HPD sets however do not result in a valid confidence procedure (according to Definition C.1) for train *or* target data.[1]

In this chapter, we propose a new approach that constructs confidence procedures that mirror the style of HPD level sets in Bayesian inference, while providing frequentist coverage properties for every $\theta \in \Theta$, regardless of $\pi(\theta)$. We apply a monotonic transformation g_θ to the posterior, so that the level sets $B_\alpha(x) = \{\theta : h(x; \theta) > \alpha\}$, where $h(x; \theta) \coloneqq g_\theta(\widehat{\pi}(\theta \mid x))$ control the type-I error at level α for any $\theta \in \Theta$ and $0 < \alpha < 1$. In Section C.2.1, we outline the construction of one such procedure that estimates $h(x; \theta)$ from the universal set

$$\mathcal{T}_{\text{univ}} = \{(\theta'_1, X'_1) \ldots (\theta'_{B'}, X'_{B'})\} \sim r(\theta)\mathcal{L}(\theta; x),$$

where the likelihood $\mathcal{L}(x; \theta)$ is the same as for the train data, and $r \gg \pi$.

[1] In addition, in terms of average or marginal coverage, HPD sets are by construction only valid for the train distribution: $\int_\Theta \mathbb{P}_{X|\theta} \left(\theta \in H(X)\right) \pi(\theta) \mathrm{d}\theta = \int_\Theta \left(\int_{H_\theta} p(x \mid \theta) \mathrm{d}x\right) \pi(\theta) \mathrm{d}\theta = \int_\mathcal{X} \left(\int_{H(x)} \pi(\theta \mid x) \mathrm{d}\theta\right) p(x) \mathrm{d}x \approx \int_\mathcal{X} \left(\int_{H(x)} \widehat{\pi}(\theta \mid x) \mathrm{d}\theta\right) p(x) \mathrm{d}x = 1 - \alpha$, where H_θ is the x-section of a HPD confidence procedure with $1 - \alpha$ credible sets $H(x)$ at every $x \in \mathcal{X}$.

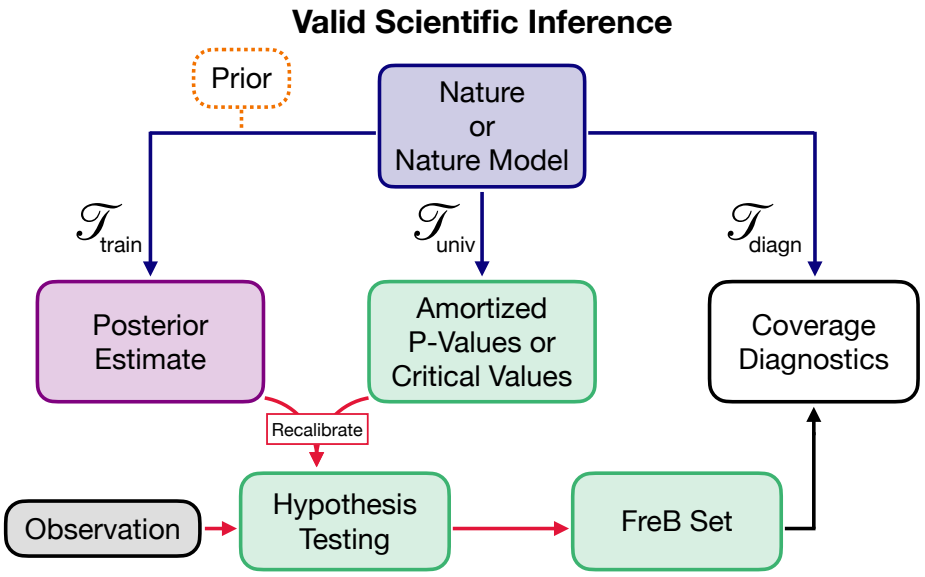


Figure C.1: **The three-branch modular framework for valid scientific inference with neural density estimators (NDE). Left branch:** Leverage a NDE to learn the posterior distribution $\pi(\theta \mid X)$ from a labeled training set $\mathcal{T}$. **Center branch:** From a universal labeled set $\mathcal{T}'$, learn amortized p-values to allow amortization for all miscoverage levels. Alternatively, learn critical values at a fixed level α. **Left + Center:** Given a new datapoint x, construct Frequentist-Bayes sets by taking level sets of the amortized p-value function, or by retaining all the values of θ for which $\widehat{\pi}(\theta \mid X)$ is larger than the corresponding critical value. **Right branch:** The coverage diagnostics branch independently checks whether the instance-wise coverage $\mathbb{P}_{X|\theta}(\theta \in B_\alpha(X))$ of the confidence set is indeed correct across the entire parameter space.

In Section C.2.1, we show how confidence procedures can be constructed for all levels of miscoverage α simultaneously from an estimate of g_θ. Our procedure can be seen as a generalization of *confidence distributions* (Schweder and Hjort, 2002; Xie and Singh, 2013; Nadarajah et al., 2015; Cui and Xie, 2023; Thornton and Xie, 2024) from one-dimensional to multidimensional parameter spaces Θ. However, for many practical applications, researchers are only interested in constructing valid and precise confidence procedures for a *fixed prespecified* miscoverage level α. In the latter case, one can reduce the complexity of the numerical estimation problem via an α-level quantile regression of the test statistic on θ, as shown in Chapter 2.

Rejection Probability Across the Entire Parameter Space

At the heart of our construction is the relationship between frequentist confidence sets $C(X)$ and acceptance regions C_{θ_0} for tests of $H_{0,\theta_0} : \theta = \theta_0$ at all $\theta_0 \in \Theta$. Below we define the rejection probability function W for an arbitrary test statistic λ that rejects H_{0,θ_0} for small values of the test statistic λ.

Definition C.2 (Rejection Probability)**.** *Let λ be any test statistic; such as the estimated*

posterior, $\lambda(X;\theta_0) = \widehat{\pi}(\theta_0 \mid X)$*. The rejection probability of the test* H_{0,θ_0} *is defined as*

$$W_\lambda(t;\theta,\theta_0) \coloneqq \mathbb{P}_{X|\theta}\left(\lambda(X;\theta_0) \leqslant t\right), \tag{C.2}$$

where $\theta, \theta_0 \in \Theta$ *and* $t \in \mathbb{R}$*.*

We can learn the rejection probability function using a monotone regression that enforces the rejection probability to be a nondecreasing function of t. The computation is straightforward when $\theta = \theta_0$. In this chapter, we propose a fast procedure for estimating the cumulative distribution function

$$F_\lambda(t;\theta_0) \coloneqq W_\lambda(t;\theta_0,\theta_0) = \mathbb{P}_{X|\theta_0}\left(\lambda(X;\theta_0) \leqslant t\right) \tag{C.3}$$

of the test statistic λ as a function of the cut-off t and the parameter value $\theta_0 \in \Theta$. For each point i $(i = 1, \ldots, B')$ in the universal set $\mathcal{T}_{\text{univ}} = \{(\theta'_1, X'_1) \ldots (\theta'_{B'}, X'_{B'})\} \sim r(\theta)\mathcal{L}(x;\theta)$, we draw a sample of cutoffs K according to the empirical distribution of the test statistic λ. Then, we regress the indicator variable

$$Y_{i,j} \coloneqq \mathbb{1}\left(\lambda(X'_i;\theta'_i) \leqslant t_j\right) \tag{C.4}$$

on θ'_i and $t_{i,j}$ $(= t_j)$ using the "augmented" calibration sample $\widetilde{\mathcal{T}}_{\text{univ}} = \{(\theta'_i, t_{i,j}, Y_{i,j})\}_{i,j}$, for $i = 1, \ldots, B'$ and $j = 1, \ldots, K$, where K is our augmentation factor. See Algorithm C.1 for more details.

Amortized P-Values for Constructing Confidence Procedures

For any test statistic λ and null hypothesis $H_{0,\theta_0} : \theta = \theta_0$, we can define a new test statistic h via a monotonic transformation,

$$\begin{aligned} h(X;\theta_0) &\coloneqq F_\lambda(\lambda(X;\theta_0);\theta_0), \\ &= \mathbb{P}_{X|\theta_0}\left(\lambda(X;\theta_0) < \lambda(x;\theta_0)\right), \end{aligned} \tag{C.5}$$

Algorithm C.1 `Learning the Rejection Probability Function`

Input: test statistic λ; calibration data $\mathcal{T}_{\text{univ}} = \{(\theta'_1, X'_1), \ldots, (\theta'_{B'}, X'_{B'})\}$; re-sampled cutoffs $G = \{t_1, \ldots, t_K\}$; evaluation points $\mathcal{V} \subset \Theta$

Output: Estimate of rejection probability $F_\lambda(t;\theta)$ when $\theta = \theta_0$, for all $t \in G$ and $\theta \in \mathcal{V}$

// **Learn rejection probability from augmented calibration data** $\widetilde{\mathcal{T}}_{\text{univ}}$
Set $\widetilde{\mathcal{T}}_{\text{univ}} \leftarrow \varnothing$
for i in $\{1, ..., B'\}$ **do**
    for j in $\{1, ..., K\}$ **do**
        Compute $Y_{i,j} \leftarrow \mathbb{1}\left(\lambda(X'_i;\theta'_i) \leqslant t_j\right)$
        Let $\widetilde{\mathcal{T}}_{\text{univ}} \leftarrow \widetilde{\mathcal{T}}_{\text{univ}} \cup \{(\theta'_i, t_j, Y_{i,j})\}$
Estimate $F_\lambda(t;\theta) \coloneqq \mathbb{P}_{X|\theta}\left(\lambda(X;\theta) \leqslant t\right)$ from $\widetilde{\mathcal{T}}_{\text{univ}}$ via a regression of Y on θ and t, which is monotonic in t.
return estimated rejection probabilities $\widehat{F}_\lambda(t;\theta)$, for $t \in G$, $\theta \in \mathcal{V}$

and then a corresponding family of confidence sets of θ by taking level sets,

$$B_\alpha(X) = \{\theta_0 \in \Theta \mid h(X;\theta_0) > \alpha\},$$

where $0 \leqslant \alpha \leqslant 1$. The following theorem shows that F_λ in Equation (C.3) is the only monotonic transformation that controls type-I errors; that is, it makes $h(X;\theta_0)$ a *valid* p-value with level sets $B_\alpha(X)$ that are confidence sets with frequentist level-α coverage.

Theorem C.3. *Let $\lambda(x;\theta)$ be any test statistic. For every fixed $\theta \in \Theta$, let $g_\theta : \mathbb{R} \longrightarrow \mathbb{R}$ be a monotonic transformation of $\lambda(x;\theta)$. Then*

$$\mathbb{P}_{X|\theta}\left(g_\theta(\lambda(X;\theta)) > \alpha\right) = 1-\alpha \text{ for every } \alpha \in (0,1) \text{ and } \theta \in \Theta$$

if, and only if, $g_\theta(\lambda(x;\theta)) = F_\lambda(\lambda(x;\theta);\theta)$.

Proof. $\Rightarrow$ direction: Fix θ and let g_θ be any monotonic transformation for λ as stated in the theorem. Then

$$\begin{aligned}
&\mathbb{P}_{X|\theta}\left(g_\theta(\lambda(X;\theta)) > \alpha\right) = 1-\alpha, \ \forall \alpha \in (0,1) \\
\iff &\mathbb{P}_{X|\theta}\left(\lambda(X;\theta) > g_\theta^{-1}(\alpha)\right) = 1-\alpha, \ \forall \alpha \in (0,1) \\
\iff &\mathbb{P}_{X|\theta}\left(\lambda(X;\theta) \leqslant g_\theta^{-1}(\alpha)\right) = \alpha, \ \forall \alpha \in (0,1) \\
\iff &F_\lambda(g_\theta^{-1}(\alpha);\theta) = \alpha, \ \forall \alpha \in (0,1) \\
\iff &g_\theta^{-1}(\alpha) = F_\lambda^{-1}(\alpha;\theta), \ \forall \alpha \in (0,1) \\
\iff &g_\theta(\lambda(x;\theta)) = F_\lambda(\lambda(x;\theta);\theta), \ \forall x \in \mathcal{X}.
\end{aligned}$$

$\Leftarrow$ direction: Let $g_\theta(\lambda(x;\theta)) = F_\lambda(\lambda(x;\theta);\theta)$. Notice that

$$\begin{aligned}
\mathbb{P}_{X|\theta}\left(g_\theta(\lambda(X;\theta)) > \alpha\right) &= \mathbb{P}_{X|\theta}\left(F_\lambda(\lambda(X;\theta);\theta) > \alpha\right) \\
&= \mathbb{P}_{X|\theta}\left(\lambda(X;\theta) > F_\lambda^{-1}(\alpha;\theta)\right) \\
&= 1 - \mathbb{P}_{X|\theta}\left(\lambda(X;\theta) \leqslant F_\lambda^{-1}(\alpha;\theta)\right) \\
&= 1 - F_\lambda(F_\lambda^{-1}(\alpha;\theta);\theta) \\
&= 1-\alpha.
\end{aligned}$$

□

From Amortized P-Values to Confidence Procedures at all Levels α Simultaneously

Algorithm C.1 offers a means to computing p-values $\widehat{h}(x;\theta_0) \coloneqq \widehat{F}_\lambda(\lambda(x;\theta_0);\theta_0)$ and the entire family of confidence sets $\widehat{B}_\alpha(x) \coloneqq \left\{\theta \in \Theta \mid \widehat{h}(x;\theta_0) > \alpha\right\}$, which is fully amortized with respect to observed data $x \in \mathcal{X}$, the parameter $\theta_0 \in \Theta$, and the miscoverage level $0 \leqslant \alpha \leqslant 1$. That is, once we have the test statistic $\lambda(x;\theta_0)$ and the rejection probability $\widehat{F}(t;\theta_0)$ as a function of all $t \in \mathbb{R}$ and $\theta_0 \in \Theta$ (via Algorithm C.1), we can perform inference for new data without retraining for all miscoverage levels α simultaneously.

Alternative Construction of Confidence Procedures at a Fixed Prespecified Level α

For many practical applications, researchers are only interested in constructing valid and precise confidence procedures with

$$\begin{aligned}\widehat{B}_\alpha(x) &:= \left\{\theta \in \Theta \mid \widehat{F}_\lambda\left(\lambda(x;\theta);\theta\right) > \alpha\right\} \\ &= \left\{\theta \in \Theta \mid \lambda(x;\theta) > \widehat{F}_\lambda^{-1}(\alpha;\theta)\right\} \end{aligned} \tag{C.6}$$

for some pre-specified miscoverage level $\alpha \in (0,1)$. In such cases, we only need to estimate the critical values $t_{\theta_0} := F_\lambda^{-1}(\alpha;\theta_0)$ for a fixed level-α test of $H_0 : \theta = \theta_0,\ \forall \theta_0 \in \Theta$. We refer the reader to Chapter 2 for more details on this method.

C.2.2 Validity of Frequentist Bayes Procedure

P-Value Estimation

The method of estimating the p-value described in Section C.2.1 is consistent. Below we adapt the general LF2I results of Chapter 2 which hold in general, even for fully amortized procedures (Algorithm C.1). The proofs are equivalent.

Assumption C.4 (Uniform consistency)**.** *The regression estimator used in Algorithm C.1 is such that*

$$\sup_{\theta,t} \mid \hat{\mathbb{E}}_{B'}[Y \mid \theta, t] - \mathbb{E}[Y \mid \theta, t]| \xrightarrow[B' \longrightarrow \infty]{a.s.} 0.$$

If Θ is continuous and the Lebesgue measure dominates r, then the estimators described, e.g., in Bierens (1983); Hardle et al. (1984); Liero (1989); Girard et al. (2014) satisfy this assumption.

Theorem C.5. *Fix $\theta_0 \in \Theta$. Under Assumption C.4 and if $h(X;\theta_0)$ is an absolutely continuous random variable then, for every $\theta \in \Theta$,*

$$\widehat{h}(X;\theta_0) \xrightarrow[B' \longrightarrow \infty]{a.s.} h(X;\theta_0)$$

and

$$\mathbb{P}_{X,\mathcal{T}_{univ}|\theta}\left(\hat{h}\left(X;\theta_0\right) \leqslant \alpha\right) \xrightarrow{B' \longrightarrow \infty} \mathbb{P}_{X|\theta}(h(X;\theta_0) \leqslant \alpha).$$

In particular,

$$\mathbb{P}_{X,\mathcal{T}_{univ}|\theta_0}\left(\hat{h}\left(X;\theta_0\right) \leqslant \alpha\right) \xrightarrow{B' \longrightarrow \infty} \alpha.$$

Assumption C.6 (Convergence rate of the regression estimator)**.** *The regression estimator is such that*

$$\sup_{\theta,t} \mid \hat{\mathbb{E}}[Z \mid \theta, t] - \mathbb{E}[Z \mid \theta, t]| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

for some $r > 0$.

Examples of regression estimators that satisfy Assumption C.6 when Θ is continuous and the Lebesgue measure dominates r can be found in Stone (1982); Hardle et al. (1984); Donoho (1994); Yang et al. (2017).

Theorem C.7. *Under Assumption C.6,*

$$|\hat{h}(X;\theta_0) - h(X;\theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

Proof of Theorem C.7. The result follows directly from Assumption C.6 and the fact that $\widehat{h}(x;\theta_0) \coloneqq \widehat{F}_\lambda(\lambda(x;\theta_0);\theta_0) = \widehat{\mathbb{E}}\left[Z \mid \theta_0, \lambda(x;\theta_0)\right]$. □

Critical Value Estimation

Our procedure for choosing critical values leads to valid hypothesis tests (that is, tests that control the type-I error probability), as long as the number of simulations B' in Algorithm 2.1 is sufficiently large. See Dalmasso* et al. (2024, Sec. 4.1) Chapter 2 for details.

Assumption C.8 (Uniform consistency)**.** *Let $\hat{F}_{B'}(\lambda;\theta)$ be the estimated distribution function of the test statistics λ indexed by θ, implied by Algorithm 2.1. Assume that the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}} |\hat{F}_{B'}(\lambda;\theta_0) - F(\lambda;\theta_0)| \xrightarrow[B'\longrightarrow\infty]{P} 0.$$

Assumption C.8 holds, for instance, for quantile regression forests (Meinshausen, 2006). Next, we show that Algorithm 2.1 yields a valid hypothesis test as $B' \to \infty$.

Theorem C.9. *Let $C_{B'} = \hat{F}_{B'}(\alpha;\theta_0)$. If the quantile estimator satisfies Assumption C.8, then, for every $\theta_0 \in \Theta$,*

$$\mathbb{P}_{X|\theta_0,C_{B'}}(\lambda(X;\theta_0) \leqslant C_{B'}) \xrightarrow[B'\longrightarrow\infty]{a.s.} \alpha,$$

where $\mathbb{P}_{X|\theta_0,C_{B'}}$ denotes the probability integrated over $X \sim p(x \mid \theta_0)$ and conditional on the random variable $C_{B'}$.

If the convergence rate of the quantile regression estimator is known (Assumption C.10), Theorem C.11 provides a finite-B' guarantee on how far the type-I error of the test will be from the nominal level.

Assumption C.10 (Convergence rate of the quantile regression estimator)**.** *Using the notation of Assumption C.8, assume that the quantile regression estimator is such that*

$$\sup_{\lambda\in\mathbb{R}} |\hat{F}_{B'}(\lambda;\theta_0) - F(\lambda;\theta_0)| = O_P\left(\left(\frac{1}{B'}\right)^r\right)$$

for some $r > 0$.

Theorem C.11. *With the notation and assumptions of Theorem C.9, and if Assumption C.10 also holds, then,*

$$|\mathbb{P}_{X|\theta_0,C_{B'}}(\lambda(X;\theta_0) \leqslant C_{B'}) - \alpha| = O_P\left(\left(\frac{1}{B'}\right)^r\right).$$

C.3 Power of Frequentist Bayes Procedure

Consider a confidence procedure $\mathbf{B} \in \Theta \times \mathcal{X}$ with θ-sections at fixed $x \in \mathcal{X}$ and $\alpha \in (0,1)$ defined by

$$B_\alpha(x) = \{\theta \in \Theta \mid h(x;\theta) > \alpha\}, \tag{C.7}$$

where $h(x;\theta)$ is the p-value (Equation C.5) for the test statistic $\lambda(x;\theta) = \pi(\theta \mid x)$. In Appendix C.2.2, we show that $\mathbf{B}$ is a valid confidence procedure on both train and target data, regardless of the choice of prior $\pi(\theta)$, satisfying $\mathbb{P}_{X|\theta}(\theta \in B_\alpha(X)) = 1-\alpha,\ \forall\theta \in \Theta$. In this section, we show that $B_\alpha(x)$ has a small expected size

$$\mathbb{E}\left(|B_\alpha(X)|\right) \coloneqq \int_{\mathcal{X}} \left(\int_{B_\alpha(x)} \mathrm{d}\theta \right) p(x)\mathrm{d}x$$

with respect to the marginal distribution $p(x) = \int \mathcal{L}(\theta;x)\pi(\theta)\mathrm{d}\theta$. Different versions of this theorem have appeared in e.g. Pratt (1961); Yu and Hoff (2018); Hoff (2023) for continuous Θ, as well as Sadinle et al. (2019) when Θ is finite.

In other words, if the design prior π is "well-specified" and places a high mass around the true parameter value θ for the target data according to $\pi(\theta) = p_{\text{obs}}(\theta)$, then the frequentist Bayes sets $B_\alpha(x)$ will not only achieve nominal coverage across the parameter space Θ; they will also on average be smaller than any other valid confidence sets with respect to the marginal distribution $p(x)$ of the train data, which is defined by the prior $\pi(\theta)$. However, if the prior is different from the (unknown) label distribution or "true prior" $p_{\text{obs}}(\theta)$ of the target data, then frequentist Bayes sets will not have optimal average constraining power with respect to $p_{\text{obs}}(x)$.

Lemma C.12 (Neyman-Pearson Lemma)**.** *Let $\mu(z)$ and $\nu(z)$ be nonnegative functions in L_1. Fix $\alpha \in (0,1)$, and assume that there exists t such that the set $A^* = \{z : \mu(z)/\nu(z) \geqslant t\}$ satisfies $\mu(A^*) = 1-\alpha$. Then A^* is the solution to the following optimization problem:*

$$\min_A \int_A \nu(z)dz \quad \textit{subject to } \int_A \mu(z)dz \geqslant 1-\alpha.$$

Theorem C.13. *Let $\mathcal{A}$ denote the space of all measurable sets $A \subseteq \Theta \times \mathcal{X}$, and let $A(x) = \{\theta : (\theta,x) \in A\}$ be the θ-section of A, and let $|A(X)| = \int_{A(X)} d\theta$ be the size of $A(X)$. Let A^* be the solution to the following minimization problem:*

$$\min_{A\in\mathcal{A}} \mathbb{E}\left[|A(X)|\right] \quad \textit{subject to } \mathbb{P}_{X|\theta}(\theta \in A(X)) \geqslant 1-\alpha,\ \forall\theta \in \Theta,$$

where the expectation is taken with respect to the marginal distribution $p(x) = \int p(x \mid \theta)\pi(\theta)d\theta$. Then, $A^*(x) = B_\alpha(x)$ (Equation C.7).

Proof. Let $A_\theta = \{x : (\theta,x) \in A\}$ be the x-section of A. Notice that the optimization problem is equivalent to

$$\min_{A\in\mathcal{A}} \int \left[\int_{A(x)} 1\mathrm{d}\theta \right] p(x)\mathrm{d}x \ \text{ subject to } \int_{A_\theta} p(x \mid \theta)\mathrm{d}x \geqslant 1-\alpha\ \forall\theta \in \Theta,$$

which is further equivalent to

$$\min_{A \in \mathcal{A}} \int \left[\int_{A_\theta} p(x) \mathrm{d}x \right] \mathrm{d}\theta \;\; \text{subject to} \;\; \int_{A_\theta} p(x \mid \theta) \mathrm{d}x \geqslant 1 - \alpha \;\; \forall \theta \in \Theta,$$

which is equivalent to a point-wise optimization problem for any given θ:

$$\min_{A_\theta} \int_{A_\theta} p(x) \mathrm{d}x \;\; \text{subject to} \;\; \int_{A_\theta} p(x \mid \theta) \mathrm{d}x \geqslant 1 - \alpha.$$

Lemma C.12 implies that the optimal solution is

$$A^*_\theta = \{x : p(x \mid \theta)/p(x) \geqslant t_\theta\},$$

where t_θ satisfies $\mathbb{P}_{X|\theta}(\theta \in A^*(X)) = 1 - \alpha$. The optimal set is then (using the fact that $p(x \mid \theta)/p(x) = \pi(\theta \mid x)/\pi(\theta)$)

$$A^* = \{(\theta, x) : \pi(\theta \mid x)/\pi(\theta) \geqslant t_\theta\},$$

or, equivalently,

$$A^* = \{(\theta, x) : \pi(\theta \mid x) \geqslant t'_\theta\},$$

where $t'_\theta = t_\theta \pi(\theta)$. □

C.4 Details on Synthetic Examples of Section 4.1

C.4.1 Synthetic Example of Figure 4.2

The synthetic example of Figure 4.2 (Panels C and D) leverages a simple setting to showcase the main components of our framework for trustworthy scientific inference. We assume that all data is generated from an (unknown) Gaussian likelihood $p(X \mid \theta) = \mathcal{N}(\theta, 1)$ and proceed as follows:

1. We construct a training set $\mathcal{T}_{\text{train}} = \{(\theta_i, X_i)\}_{i=1}^{B} \sim p(X \mid \theta)\pi(\theta)$ with $B = 100{,}000$ and $\pi(\theta) = \mathcal{N}(0, 1)$ to learn $\hat{\pi}(\theta \mid X)$ through a generative model. For this example, we use a simple masked autoregressive flow (Papamakarios and Murray, 2016; Lueckmann et al., 2017) as implemented in the `SBI` library (Tejero-Cantero et al., 2020), using default hyper-parameters;

2. We construct a "universal" calibration set $\mathcal{T}_{\text{univ}} = \{(\theta_i, X_i)\}_{i=1}^{B'} \sim p(X \mid \theta)r(\theta)$ with $B' = 50{,}000$ and $r(\theta) = \mathcal{U}(-10, 10)$ to learn a monotonic transformation $\hat{F}(\hat{\pi}(\theta \mid x); \theta)$ of the estimated posterior. Here, we estimate an amortized p-value function $\mathbb{P}_{X|\theta}\left(\hat{\pi}(\theta \mid X) < \hat{\pi}(\theta_0 \mid x)\right)$ according to Algorithm C.1 by setting the number of resampled cutoffs to $K = 10$ and leveraging a tree-based gradient-boosted probabilistic classifier as implemented in the `CatBoost` library (Prokhorenkova et al., 2018). We only optimize the number of trees and the maximum depth, which are finally set to 1000 and 9, respectively;

3. We generate $x_{\text{target}} \sim p(X \mid \theta^\star = 4)$ and construct an HPD set according to Equation C.1 and a FreB set as shown in Section 4.3.2 and Section C.2.1. Note that we only observe a single sample to infer $\theta^\star$, i.e., $n = 1$;

4. Finally, we check instance-wise coverage as detailed in Chapter 2 by first generating a diagnostic set $\mathcal{T}_{\text{diagn}} = \{(\theta_i, X_i)\}_{i=1}^{B''} \sim p(X \mid \theta) r(\theta)$ with $B'' = 50{,}000$ and $r(\theta) = \mathcal{U}((-10, 10))$ and then learning a probabilistic classifier via a univariate Generalized Additive Model (GAM) with thin plate splines as implemented in the `MGCV` library in `R` Wood (2015).

C.4.2 Synthetic Example of Figure 4.6

The synthetic example of Figure 4.6 showcases the main properties of our framework — i.e., reliability (in the form of correct coverage) and precision (in the form of optimal constraining power) — for an inference task that was introduced in Sisson et al. (2007) and has become a standard benchmark in the SBI literature (Clarté et al., 2021; Toni et al., 2009; Simola et al., 2021; Lueckmann et al., 2021). It consists of estimating the (common) mean of the components of a two-dimensional Gaussian mixture, with one component having much broader covariance: $X \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, I) + \frac{1}{2}\mathcal{N}(\theta, 0.01 \cdot I)$, where $\theta \in \mathbb{R}^2$ and $n = 1$. We proceed as follows:

1. We construct a training set $\mathcal{T}_{\text{train}} = \{(\theta_i, X_i)\}_{i=1}^{B} \sim p(X \mid \theta)\pi(\theta)$ with $B = 50{,}000$ and $\pi(\theta) = \mathcal{N}(0, 2I)$ to learn $\hat{\pi}(\theta \mid X)$ through a generative model. For this example, we use a flow matching posterior estimator, whose idea was first introduced in (Lipman et al., 2022) and then adapted for simulation-based inference settings in (Wildberger et al., 2024). We leverage the implementation available in the `SBI` library (Tejero-Cantero et al., 2020), using default hyper-parameters;

2. We construct a "universal" calibration set $\mathcal{T}_{\text{univ}} = \{(\theta_i, X_i)\}_{i=1}^{B'} \sim p(X \mid \theta) r(\theta)$ with $B' = 30{,}000$ and $r(\theta) = \mathcal{U}([-10, 10] \times [-10, 10])$ to learn a monotonic transformation $\hat{F}(\hat{\pi}(\theta \mid x); \theta)$ of the estimated posterior. Here, we again estimate an amortized p-value function $\mathbb{P}_{X|\theta}\left(\hat{\pi}(\theta \mid X) < \hat{\pi}(\theta_0 \mid x)\right)$ according to Algorithm C.1 by setting the number of resampled cutoffs to $K = 10$ and leveraging a tree-based gradient-boosted probabilistic classifier as implemented in the `CatBoost` library (Prokhorenkova et al., 2018). We only optimize the number of trees and the maximum depth, which are finally set to 1000 and 9, respectively;

3. We then generate two observations to represent poor alignment with the prior distribution — $x_{1,\text{target}} \sim p(X \mid \theta^\star = [8.5, -8.5])$ and $x_{2,\text{target}} \sim p(X \mid \theta^\star = [-8.5, -8.5])$ — and one observation to represent good alignment with the prior distribution — $x_{3,\text{target}} \sim p(X \mid \theta^\star = [0, 0])$ — for which we again construct HPD sets according to Equation C.1 and FreB sets as shown in Section 4.3.2 and Section C.2.1. As in the previous example, we only observe a single sample to infer $\theta^\star$, i.e., $n = 1$;

4. We check instance-wise coverage as detailed in Chapter 2 by first generating a diagnostic set $\mathcal{T}_{\text{diagn}} = \{(\theta_i, X_i)\}_{i=1}^{B''} \sim p(X \mid \theta) r(\theta)$ with $B'' = 20{,}000$ and $r(\theta) = \mathcal{U}([-10, 10] \times [-10, 10])$ and then learning a probabilistic classifier via a bivariate

Generalized Additive Model (GAM) with thin plate splines as implemented in the `MGCV` library in `R` Wood (2015).

C.5 Supplement for Case Study I

C.5.1 Experimental Setup

Training and target data sets for this case study have been created as a proof-of-concept. We base the parameter distributions of the simulated air showers on the following three gamma-ray sources:

- **Crab Nebula:** A pulsar-wind nebula emitting the brightest and stable TeV signal in the northern hemisphere sky, for the past 970 years.
- **Markarian 421 (Mrk421):** A blazar located about 397 million light years from earth. Blazars and other active galactic nuclei emit intense electromagnetic radiation, facilitating the discovery of otherwise faint distant galaxies (Abdo and Others, 2011).
- **Dark Matter (DM) Annihilation:** Similar to matter-antimatter annihilation, some theories of dark matter propose an annihilation mechanism for dark matter particles, which emit gamma rays following a certain energy spectrum (Jueid et al., 2021). Gamma-ray measurements from regions of space thought to contain dark matter (e.g. around galaxies) can put these theories to the test.

Note that Mrk421 is a point source much like the Crab Nebula, but the DM Annihilation source is a theorized mechanism that could happen anywhere in the cosmos. As such, we treat DM as a diffuse source of gamma ray events that hit the Earth from all directions. We only consider the zenith component of the point source trajectories, azimuth distributed uniformly, for direct comparison between sources. The zenith distribution along the *Crab* and *Mrk421* trajectories relative to the zenith distribution in the pre-simulated `CORSIKA` data is used to assign weights to individual gamma ray events. All trajectory calculations are performed using `astropy` (Astropy Collaboration and Astropy Project Contributors, 2022). Each source's theoretical energy spectrum assign weights to individual gamma ray events in the pre-simulated set. For the *Crab*, we use the log-parabola fit proposed by Aleksić et al. (2015). For *Mrk421*, we perform a custom fit to observational data that accounts for attenuation of gamma-ray flux due to extragalactic background light (EBL). For the DM source, we use `gammapy` (Donath and Others, 2023) to generate the dark-matter annihilation spectrum for very heavy DM particles (100 TeV). We do not attenuate this spectrum using EBL. Most cosmic rays observed from Earth are actually hadrons (specifically protons). Because hadrons also produce an atmospheric shower observable by ground detectors, a preliminary step in reconstructing gamma-ray events from ground detector data is to first determine if an observed shower is a gamma ray or a hadron. We do not perform this initial classification step in this case study and focus only on the reconstruction of gamma ray events, and we refer the reader to existing approaches to this initial classification such as Alfaro et al. (2022).

C.5.2 Data

Our data set consists of a large number of labeled gamma-ray events (E_i, Z_i, A_i, x_i). For each event i:

1. E_i is the energy of the original gamma ray in GeV
2. Z_i is the zenith angle, defined as the angle that the gamma ray's source makes with the vertical. A source directly overhead would have a zenith angle of 0.
3. A_i is the azimuthal angle, defined as the angle between the source and the true north, measured clockwise. For example, a source directly east of the observer would have azimuthal angle of 90 degrees
4. x_i is the data collected by ground detectors by the resulting atmospheric shower

Our data come from the CORSIKA (Heck et al., 1998) simulator. We make three splits from the data:

1. Training set (B = 1,072,821) used to train our posterior estimator $p(\theta_i \mid x_i)$
2. Universal set (B' = 98,765) used to train our FreB quantile regression
3. Diagnostic set (B'' = 42,270) used to evaluate the performance of our confidence set procedures

For observed detector data x_i, we assume full ground coverage in a 4km x 4km square, where each detector is 2m x 2m. For a given shower, we assume that each detector is capable of recording the identity and timing of every secondary particle that passes through it. The number of secondary particles per shower can range from less than 10 for low-energy gamma rays to up to 100 million for very high-energy gamma rays. Figure C.2 shows an example of the data recorded for a single gamma ray air shower. Although many types of secondary particles may appear in an atmospheric shower, we consider only two broad groups (photons/electrons/positrons versus everything else) for ease of analysis.

We remove all gamma-ray events in all data splits where less than 10 ground detectors recorded secondary particle hits. We weight our filtered training data to resemble the Crab Nebula in terms of its energy spectrum Aleksić et al. (2015). We also weight the training data to resemble a fixed reference distribution in zenith. This reference distribution is a combination of a uniform distribution over the sphere and atmospheric effects at high zenith angles. We assume that $p(x_i \mid \theta_i)$ exhibits azimuthal symmetry.

We place our observer at 19 degrees north. This latitude corresponds to the current location of the operational HAWC observatory (Abeysekara et al., 2023) and provides a better view of the Crab Nebula versus the proposed SWGO site in the southern hemisphere.

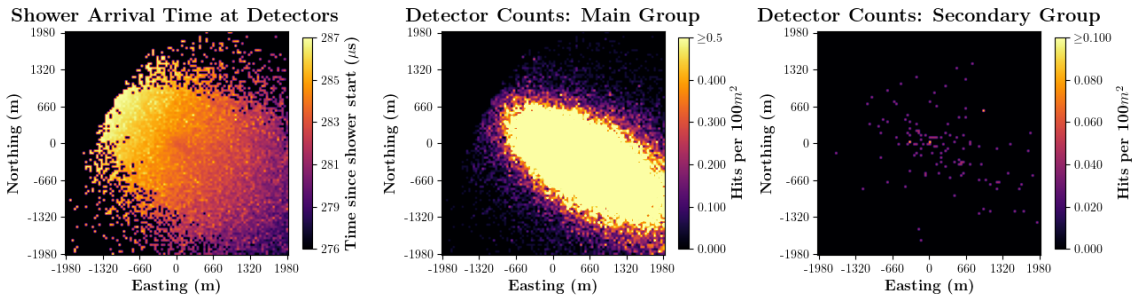


Figure C.2: **Example features collected for a single gamma-ray event.** For each detector (represented by the pixels in each figure), we plot three measurements of the induced atmospheric shower. **(Left)** Average arrival time of secondary shower particles. **(Center)** Number of detections of "main" shower particles (photons, electrons, and positrons). **(Right)** Number of detections of "secondary" shower particles (muons, all other possible shower particles).

C.5.3 Details on Training

We train a Flow Matching Posterior Estimator (Wildberger et al., 2024), a diffusion-based model with training-based acceleration, to obtain an estimate of the posterior $\hat{p}(\theta_i \mid x_i)$. We use the `sbi` Python package v0.23.2 (Tejero-Cantero et al., 2020) to implement the flow matching model. We use the default model architecture in `sbi`, but use a custom context model to convert our high- dimensional x_i into a low-dimensional context vector:s

1. x_i has initial shape 3x2000x2000
2. Max pooling for timing channel and Average pooling for counts channels with kernel size/stride of 20
3. 2D Convolution with max pooling and batch normalization
4. 2D Convolution with max pooling and batch normalization
5. Flatting and fully connected layer to a fixed sized context vector

Additional hyperparameters can be found on the `sbi` GitHub repository and our GitHub repository (https://github.com/lee-group-cmu/vsi). We use default training parameters from the `sbi` Python package.

C.6 Supplement for Case Study II

C.6.1 Experimental Set-Up

A galactic model is a representation of the galaxy as a mixture of three components, the "thick disk," "thin disk," and "stellar halo." These components represent fields of stellar

	$[Fe/H]$ Halo Mean	$[Fe/H]$ Halo Std. Dev.	$[Fe/H]\|Age$ Ctr.	$[Fe/H]\|Age$ Scale
Model H	-2.25	0.5	0.0	0.4
Model D	-0.6	0.2	-0.72	0.58

Table C.1: Galactic model parameters

$T_{\rm eff}$ $[10^3 K]$	$\log g$ [cgs]	$[Fe/H]$ [dex]	$[Fe/H]_{\rm surf}$ [dex]	L $[L_\odot]$
7.13	2.85	-2.80	-2.76	7.87

Dist. [kpc]	$M_{\rm ini}$ $[M_\odot]$	Age [Gyr]	EEP
0.842	1.30	2.48	696

Table C.2: True stellar parameters for the displayed star in Section 4.2.2

objects which comprise the majority of the galaxy's stellar objects. Lines of sight along different galactic coordinates slice through these components in different proportions. In this case study, we identify stars along the $(\ell, b) = (70^\circ, 30^\circ)$ (in Galactic coordinates) line of sight because it amply includes both disk and halo components. We further identify a few sources which one may plausibly measure and for which one may like to conduct inference.

To obtain **Model H**, we decrease the default mean and increase the default variance of the age distribution in the galactic halo component from `Brutus`. To obtain **Model D**, we increase the mean of the conditional metallicity-given-age distribution according to Table C.1. These hyperparameters affect the `Brutus` model which is encoded as a collection of PDFs which can be evaluated directly. See (Speagle et al., 2025, Section 2.4) for further details on the `Brutus` prior. The true parameters of the star displayed in Figure 4.4 are given in Table C.2.

C.6.2 Data

`Brutus` is an open-source Python package designed to quickly estimate stellar properties, distances, and reddening based on photometric and astrometric data Speagle et al. (2025). It operates using grids of stellar models within a Bayesian framework that incorporates Galactic models, enabling efficient parameter estimation. Brutus accepts photometric and astrometric data as inputs, and it outputs derived stellar properties, including 3D positions, effective temperatures, distances, and extinction values. It uses empirical corrections for better accuracy and can rapidly process large datasets, making it suitable for studies requiring quick stellar parameter recovery.

Our dataset consists of a large number of labeled stellar objects drawn from a prior over the log-scale gravitational constant ($\log g$), effective object temperature ($T_{\rm effective}$), surface metallicity ($[Fe/H]_{\rm surface}$), luminosity (L), distance (d), dust extinction (A_V), and differential extinction (R_V). The parameters of interest of the model are

$$\theta = (\log g, T_{\rm effective}, [Fe/h]_{\rm surface}, L) \in \mathbb{R}^5.$$

Note that we treat A_V, R_V, and d as nuisance parameters, i.e. unavailable for inference in this setting. To report our inference on θ, d is included along with θ in posterior estimation as it is known to be strongly informative of the expected measurements whereas A_V and R_V are ignored.

The estimated filtered spectra for those objects are then hypothetically obtained under the Two Micron All-Sky Survey (2MASS) J, H, and K_S filters (Skrutskie et al., 2006) and the Panoramic Survey Telescopic And Rapid Response System (PS) 'grizy' filters (Bolden and Kervin, 2010). Our likelihood processes the raw magnitudes m_i of these filtered spectra with noiseless and noisy components. First, the magnitudes m_i for the eight photometric bands are estimated noiselessly,

$$m_i \coloneqq f_i(\theta) + \mu(d) + A_V \cdot (R_i(\theta) + R_V \cdot R'_i(\theta)),$$

where $\mu(d) = 5\log(d/10)$ is the distance modulus in parsecs (pc) and f, R, and R' are deterministic functions available in the `Brutus` library parameterizing spectral generation and reddening. Then some random noise is added on the flux scale,

$$F_i \sim \mathcal{N}\left(\exp\left(-\frac{2}{5}m_i\right), 0.2\right).$$

The final noised magnitudes $M_i = -\frac{5}{2}\log(F_i)$ are normalized and the normalized measurements together with the raw magnitude norm give the final measurements,

$$x = (\tilde{M}_1, \tilde{M}_2, \ldots, \tilde{M}_8, M) \in \mathbb{R}^9,$$

where M is such that $\tilde{M}_i = M_i/M$.

C.6.3 Details on Training

We train a posterior estimator $\hat{\pi}(\theta \mid x)$ using a normalizing flow model with the masked autoregressive flow (Papamakarios and Murray, 2016; Lueckmann et al., 2017) architecture as implemented in the `SBI` library (Tejero-Cantero et al., 2020) with 50 hidden features over 5 hidden layers. Quantile regression for calibration of the FreB method was implemented using Python's `CatBoost` library (Prokhorenkova et al., 2018). We used $B = 500{,}000$ for training, $B' = 500{,}000$ for calibration, and $B'' = 25{,}000$ for evaluation.

C.6.4 Additional Results

For completeness, we show an additional illustrative example in Figure C.3.

C.7 Supplement for Case Study III

C.7.1 Data

Our data consists of a set of 202,970 Gaia XP spectra (Gaia Collaboration et al., 2023) cross-matched with APOGEE (Majewski et al., 2017) derived stellar labels that have been

T_{eff} $[10^3 K]$	$\log g$ [cgs]	$[Fe/H]$ [dex]	$[Fe/H]_{\text{surf}}$ [dex]	L $[L_\odot]$
4.74	3.61	0.25	0.30	1.66

Dist. [kpc]	M_{ini} $[M_\odot]$	Age [Gyr]	EEP
1.63	1.05	1.12	478

Table C.3: True stellar parameters for the additional example star in Section C.6.4

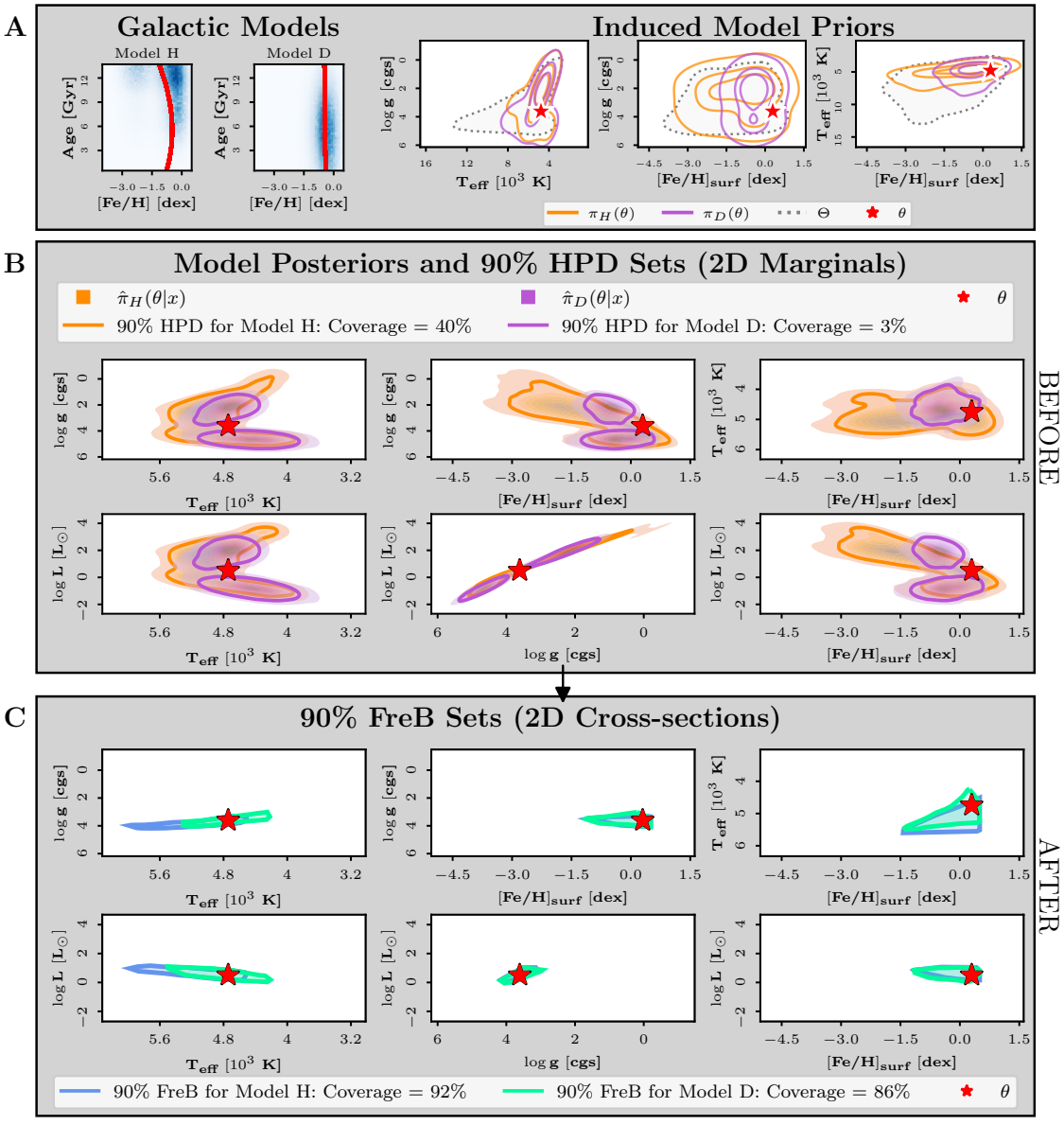


Figure C.3: See caption of Figure 4.4.

observed across the Milky Way galaxy (see Table 4.1, row III). The stellar labels refer to stellar properties like effective temperature (T_{eff}), surface gravity ($\log g$), and metallicity (Fe/H)—all of which were derived from the high-resolution APOGEE spectra.

This cross-match between the two catalogs was originally compiled by Laroche and Speagle (2024) to train a scatter variational auto-encoder that was used to generate XP spectra. The "full" cross-match catalog contained 502,311 stars, but after implementing filters to ensure a high signal-to-noise ratio for reliable labels for training, we were left with the "good" labels set of 202,970 stars. These filter ranges for signal-to-noise ratios and measurement errors were placed on measurements including T_{eff}, $\log g$, metallicity, and $BP - RP$ (see Laroche and Speagle (2024) for details).

No PPS			
	Train	Calibration	Diagnostics
GB-stars	85001 (83.8%)	51001 (83.8%)	838 (83.8%)
MS-stars	16484 (16.2%)	9890 (16.2%)	162 (16.2%)
Total	101485	60891	1000

PPS MS-dominated			
	Train	Calibration	Diagnostics
GB-stars	0	51001 (83.8%)	1000
MS-stars	16484	9890 (16.2%)	0
Total	16484	60891	1000

PPS GB-dominated			
	Train	Calibration	Diagnostics
GB-stars	85001	51001 (83.8%)	0
MS-stars	0	9890 (16.2%)	1000
Total	85001	60891	1000

D

Additional Results for Chapter 5

D.1 The Bayes Factor as a Frequentist Test Statistic

In this work, we treat the Bayes factor as a frequentist test statistic, similar to the Bayes Frequentist Factor (`BFF`) method in Dalmasso* et al. (2024). Consider the composite-versus-composite hypothesis test:

$$H_{0,y} : \theta \in \Theta_0 \quad \text{versus} \quad H_{1,y} : \theta \in \Theta_1 \tag{D.1}$$

where $\Theta_0 = \{y\} \times \mathcal{N}$, $\Theta_1 = \{y\}^c \times \mathcal{N}$, and $y \in \{0, 1\}$. The Bayes factor of the test is defined as

$$\tau_y(x) \coloneqq \frac{\mathbb{P}'(x \mid H_{0,y})}{\mathbb{P}'(x \mid H_{1,y})} = \frac{\int_{\mathcal{N}} \mathcal{L}(x; y, \nu)\, p'(\nu \mid y)\, \mathrm{d}\nu}{\int_{\mathcal{N}} \mathcal{L}(x; 1-y, \nu)\, p'(\nu \mid 1-y)\, \mathrm{d}\nu}$$

By Bayes theorem,

$$\begin{aligned} \tau_y(x) &= \frac{\int_{\mathcal{N}} \frac{p'(y,\nu|x)}{p'(y,\nu)} p'(\nu \mid y)\, \mathrm{d}\nu}{\int_{\mathcal{N}} \frac{p'(1-y,\nu|x)}{p'(1-y,\nu)} p'(\nu \mid 1-y)\, \mathrm{d}\nu} = \frac{\int_{\mathcal{N}} \frac{p'(y,\nu|x)}{\mathbb{P}'(Y=y)}\, \mathrm{d}\nu}{\int_{\mathcal{N}} \frac{p'(1-y,\nu|x)}{\mathbb{P}'(Y=1-y)}\, \mathrm{d}\nu} \\ &= \frac{\mathbb{P}'(Y = y \mid x)\, \mathbb{P}'(Y = 1-y)}{\mathbb{P}'(Y = 1-y \mid x)\, \mathbb{P}'(Y = y)}. \end{aligned} \tag{D.2}$$

However, unlike `BFF`, we are not estimating the likelihood or odds from simulated data, but instead directly evaluate a pretrained classifier $\mathbb{P}'(Y = y \mid x)$.

D.2 Proofs

For simplicity in notation, we will henceforth omit the "train" and "target" subscripts in $\mathbb{P}$. The symbol $\mathbb{P}'$ will represent the training distribution, while $\mathbb{P}$ will denote the target distribution.

Proof of Lemma 5.3. This follows from the fact that $W_\lambda(C; y, \nu)$ only depends on the conditional randomness of $X \mid y, \nu$, which, under GLS, is the same on both train and target data. □

Proof of Theorem 5.5. Notice that

$$\begin{aligned}\mathbb{P}(\lambda(X) \leqslant C^*_{\alpha,y}(X) \mid y, \nu) &= \mathbb{P}(\lambda(X) \leqslant C^*_{\alpha,y}(X), \nu \in S_y(X;\gamma) \mid y, \nu) \\ &+ \mathbb{P}(\lambda(X) \leqslant C^*_{\alpha,y}(X), \nu \notin S_y(X;\gamma) \mid y, \nu) \\ &\leqslant \mathbb{P}(\lambda(X) \leqslant W_\lambda^{-1}(\beta; y, \nu) \mid y, \nu) + \mathbb{P}(\nu \notin S_y(X;\gamma) \mid y, \nu) \\ &\leqslant \beta + \gamma = \alpha,\end{aligned}$$

which proves the first part of the result. Similarly,

$$\begin{aligned}\mathbb{P}(\lambda(X) \geqslant \widetilde{C}^*_{\alpha,y}(X) \mid 1-y, \nu) &= \mathbb{P}(\lambda(X) \geqslant \widetilde{C}^*_{\alpha,y}, \nu \in S_{1-y}(X;\gamma) \mid 1-y, \nu) \\ &+ \mathbb{P}(\lambda(X) \geqslant \widetilde{C}^*_{\alpha,y}, \nu \notin S_{1-y}(X;\gamma) \mid 1-y, \nu) \\ &\leqslant \mathbb{P}(\lambda(X) \geqslant W_\lambda^{-1}(\beta; 1-y, \nu) \mid 1-y, \nu) \\ &+ \mathbb{P}(\nu \notin S_{1-y}(X;\gamma) \mid 1-y, \nu) \\ &\leqslant 1 - \beta + \gamma = 1 - \alpha,\end{aligned}$$

and therefore

$$\mathbb{P}(\lambda(X) \leqslant \widetilde{C}^*_{\alpha,y}(X) \mid 1-y, \nu) \geqslant \alpha,$$

which concludes the proof. □

Proof of Theorem 5.6. By construction

$$\begin{aligned}\mathbb{P}(Y \in \mathbf{H}(X;\alpha) \mid y, \nu) &= \mathbb{P}\left(\widehat{\tau}_y(X) > C^*_{\alpha,y}(X) \mid y, \nu\right) \\ &= 1 - \mathbb{P}\left(\widehat{\tau}_y(X) \leqslant C^*_{\alpha,y}(X) \mid y, \nu\right) \\ &\geqslant 1 - \alpha,\end{aligned}$$

where the last inequality follows from Lemma 5.5. This proves the first statement of the theorem. To prove the second statement, notice that

$$\begin{aligned}\mathbb{P}(Y \in \mathbf{H}(X;\alpha)) &= \int \mathbb{P}(Y \in \mathbf{H}(X;\alpha) \mid y, \nu) \mathrm{d}\mu(y, \nu) \\ &\geqslant \int (1-\alpha) \mathrm{d}\mu(y, \nu) \\ &= 1 - \alpha,\end{aligned}$$

where $\mu(y, \nu)$ denotes the measure on (Y, ν) on the target set. □

D.3 Estimating the Rejection Probability Function

We learn $W_\lambda(C; y, \nu)$ using a monotone regression that enforces the rejection probability to be a non-decreasing function of C. For each point $i = 1, \ldots, B'$ in the calibration set $\mathcal{T}' = \{(Y_1, \nu_1, X_1), \ldots, (Y_{B'}, \nu_{B'}, X_{B'})\}$ drawn from $p_{\text{train}}(\theta)\mathcal{L}(x;\theta)$ where $\theta = (Y, \nu)$, we sample a set of K cutoffs according to the empirical distribution of the test statistic λ. Then, we regress the random variable

$$Z_{i,j} := \mathbb{1}\left(\lambda(X_i) \leqslant C_j\right) \tag{D.3}$$

on Y_i, ν_i and $C_{i,j}$ $(= C_j)$ using the "augmented" calibration set $\mathcal{T}'' = \{(Y_i, \nu_i, C_{i,j}, Z_{i,j})\}_{i,j}$, for $i = 1, \ldots, B'$ and $j = 1, \ldots, K$, where K is the augmentation factor. See Algorithm D.1 for details.

Algorithm D.1 `Learning the Rejection Probability Function`

Input: test statistic λ; calibration data $\mathcal{T}' = \{(Y_1, \nu_1, X_1), \ldots, (Y_{B'}, \nu_{B'}, X_{B'})\}$; sampled cutoffs $G = \{C_1, \ldots, C_K\}$

Output: Estimate of rejection probability $W_\lambda(C; y, \nu)$ for all $C \in G$, $y \in \{0, 1\}$ and $\nu \in \mathcal{N}$

1: // **Learn rejection probability from augmented calibration data $\mathcal{T}''$**
2: Set $\mathcal{T}'' \leftarrow \varnothing$
3: **for** i in $\{1, ..., B'\}$ **do**
4: **for** j in $\{1, ..., K\}$ **do**
5: Compute $Y_{i,j} \leftarrow \mathbb{1}(\lambda(X_i) \leqslant C_j)$
6: Let $\mathcal{T}'' \leftarrow \mathcal{T}'' \cup \{(Y_i, \nu_i, C_j, Z_{i,j})\}$
7: Estimate $W_\lambda(C; y, \nu) \coloneqq \mathbb{P}_{y,\nu}(\lambda(X) \leqslant C)$ from $\mathcal{T}''$ via a regression of Z on Y, ν and C, which is monotonic in C.
8: **return** Estimated rejection probabilities $\widehat{W}_\lambda(C; y, \nu)$, for $C \in G$, $y \in \{0, 1\}$ and $\nu \in \mathcal{N}$

D.4 Diagnostics of Estimated ROC Curves

Here we describe how to evaluate goodness-of-fit of an estimate of the rejection probability function. This is inspired by methods that use the Probability Integral Transform (PIT) to assess conditional density estimators (Cook et al., 2006a; Freeman et al., 2017; Izbicki et al., 2017; D'Isanto and Polsterer, 2018).

If $W_\lambda(C; y, \nu) = \mathbb{P}_{\text{target}}(\lambda(X) \leqslant C \mid y, \nu) = F_{\lambda(X)|y,\nu}(C)$ is well estimated, then the random variable$W_\lambda(\lambda(X'); y, \nu) \sim \text{Uniform}(0, 1)$, where X' is drawn from the simulator using (y, ν) as parameters. This suggests we assess the performance of our estimator of W, $\widehat{W}$, via a P-P plot comparing $\widehat{W}(\lambda(X_1); Y_1, \nu_1), \ldots, \widehat{W}(\lambda(X_B); Y_B, \nu_B)$ to a Uniform$(0, 1)$ distribution, where $(\lambda(X_1); Y_1, \nu_1), \ldots, (\lambda(X_B); Y_B, \nu_B)$ denote an evaluation sample drawn from the simulator. The distribution of these statistics can however be uniform even if $\widehat{W}$ is not a good estimate (Zhao et al., 2021, Theorem 1). Here, we avoid this problem by dividing the parameter space Θ into bins and constructing separate distribution plots for samples within each bin.

D.5 The Standard Bayes Classifier

Lemma D.1 (Bayes classifier)**.** *Let $h : \mathcal{X} \to \{0, 1\}$ be a classification rule. Define the weighted loss*

$$W = c_1 \mathbb{1}_{\{1\}}(Y)\, \mathbb{1}_{\{0\}}(h(X)) + c_0 \mathbb{1}_{\{0\}}(Y)\, \mathbb{1}_{\{1\}}(h(X)), \tag{D.4}$$

where c_k is the cost of mis-classifying a $Y = k$ observation, for $k = 0, 1$. The Bayes (that is, optimal) classifier that minimizes the error rate $\mathbb{E}_{target}(W)$ averaged over both X and Y is given by

$$h^*(x) = \begin{cases} 1 & \text{if } \mathbb{P}_{target}(Y = 1 \mid x) > \alpha^*, \\ 0 & \text{if } \mathbb{P}_{target}(Y = 1 \mid x) < \alpha^*, \\ \textit{arbitrary} & \text{if } \mathbb{P}_{target}(Y = 1 \mid x) = \alpha^*, \end{cases} \tag{D.5}$$

where $\alpha^* := \frac{c_0}{c_0+c_1}$.

Remark D.2 (Balanced accuracy)**.** *If there is no shift between the train and target sets, a common choice for the loss (Equation* (D.4)*) is* $c_1 = 1/\mathbb{P}_{train}(Y = 1)$ *and* $c_0 = 1/\mathbb{P}_{train}(Y = 0)$*. This yields the balanced error rate*

$$\mathbb{E}_{train}(W) = \mathbb{P}_{train}(h(X) = 0 \mid Y = 1) + \mathbb{P}_{train}(h(X) = 1 \mid Y = 0)$$

and the cut-off $\alpha^* = \mathbb{P}_{train}(Y = 1)$ *for the Bayes classifier (Equation* (D.5)*).*

Remark D.3 (Bayes classifier under GLS)**.** *Under GLS, there is no monotonic relationship between* $\mathbb{P}_{target}(Y = 1 \mid x)$ *and* $\mathbb{P}_{train}(Y = 1 \mid x)$*. Thus, it is not possible to use* $\mathbb{P}_{train}(Y = 1 \mid x)$ *to recover* $\mathbb{P}_{target}(Y = 1 \mid x)$ *using standard label shift corrections (Saerens et al., 2002; Lipton et al., 2018).*

Remark D.4 (Bayes classifier under the presence of nuisance parameters but no GLS)**.** *If there is no GLS,* $\mathbb{P}_{train}(Y = 1 \mid x) = \mathbb{P}_{target}(Y = 1 \mid x)$*. However, without a nuisance-aware cutoff, the Bayes classifier is usually calibrated to control type-I error marginally over* ν*. NAPS instead controls this error for all* $\nu \in \mathcal{N}$*.*

D.6 Additional Results and Details on Cosmic Ray Experiment

D.6.1 Experimental Set-Up with Ground-Based Detector Arrays

The data used in this paper are generated via the CORSIKA cosmic ray simulator (Heck et al., 1998). CORSIKA is a Monte Carlo simulation program that models the interactions of primary cosmic rays with the Earth's atmosphere. Given values of the parameters μ, E, Z, A, which define the primary cosmic ray identity, energy, zenith and azimuth angle, respectively, CORSIKA outputs the identities, momenta, positions, and arrival times of all secondary particles generated in the atmospheric shower, that eventually reach the ground and that are mostly muons, electrons and photons at gamma-ray energies with minor abundance of heavier particles.

The measured data x in our analysis does not incorporate the full shower footprint, as this level of information cannot be captured in any realistic scenario. Instead, we simulate a simple 6×6 detector grid, where each detector covers a 2×2 m^2 area, with 48 m detector spacing. Information for a secondary particle of a particular shower footprint is incorporated into the analysis only if that secondary particle lands within the area of a detector. See Figure D.1 (right) for a simplified representation of the detector grid.

We assume 100% detector efficiency and that all secondary particles types are detectable. We also assume that showers always originate at the center of the detector grid. Finally, we assume that both the zenith and azimuth angles Z and A are known due to the relative ease with which they can be estimate from observed footprint data. Thus, our only nuisance parameter for inference on μ is the energy E of the cosmic ray.

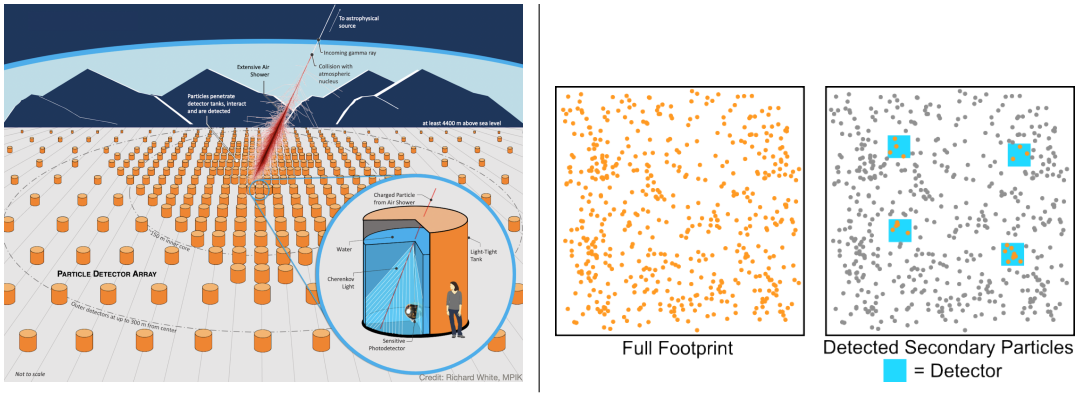


Figure D.1: **Left:** Artistic representation of the SWGO array. The inlay shows the individual detector unit. **Right:** Although we have access to all secondary particles in our simulated cosmic ray showers, we only include the particles that hit our simulated detector setup (blue rectangles) in the analysis. This layout pictured here is an illustrative example.

The data used to estimate the test statistic are drawn according to the following distribution (which may be different from that of actual astrophysical sources):

1. Gamma ray to Hadron ratio $1:1$ (whereas actual observed ratios are in the range $1:1{,}000 - 1:100{,}000$)
2. Energy between 100 TeV and 10 PeV, with probability density proportional to E^{-1} for gamma rays and E^{-2} for hadrons (with standard astrophysical sources closer to between $-2:-4$)
3. Zenith uniformly distributed between 0 and 65 degrees
4. Azimuth uniformly distributed between -180 and 180 degrees

To derive x_i, we first define four secondary particle groups: photons (neutral); electrons and positrons; muons (charged); and all other secondary particle types. Then for each simulated detector, we record the count of particles in each group that hit the detector. This results in a vector of length $4 \cdot 36 = 144$ for each primary cosmic ray that represents the detector data. We construct x_i by concatenating the detector data with Z_i and A_i.

For the calibration and test sets, we use the same reference distribution.

D.6.2 Details on the algorithms used in Section 5.5.3

We used gradient boosted probabilistic classifiers as implemented in `CatBoost` (Prokhorenkova et al., 2018) to estimate both $\mathbb{P}(Y \mid X)$ and $W_\lambda(C; y, \nu)$. For the latter, `CatBoost` allows to easily enforce monotonicity constraints on the features, which we used on C. To compute cutoffs, we used the `brentq` routine (Brent, 2013) to calculate the inverse and the differential evolution global optimization algorithm (Storn and Price, 1997) to find the infimum. Both are implemented in `SciPy` (Virtanen et al., 2020). To obtain confidence sets for ν, we

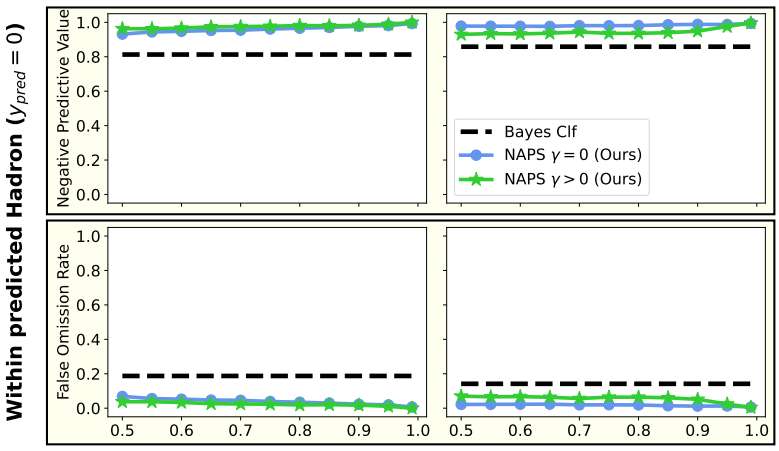


Figure D.2: **Classification metrics within predicted Hadrons ($y_{\text{pred}} = 0$).** Results are binned according to whether the shower energy is below (left) or above (right) the median value. Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue) achieve high precision and low false discovery rates (FDR), especially at high confidence levels. In addition, by constraining the nuisance parameters $\nu = (E, A, Z)$, we see performance (NAPS $\gamma > 0$; green) increase in the lower energy bin but with a corresponding tradeoff in the higher energy bins. Both approaches yield better results relative to the oracle Bayes classifier (black dashed line).

used the method developed by Masserano et al. (2023) with a masked autoregressive flow (Papamakarios et al., 2017) since it guarantees that the constructed region contains the true value of ν at the desired confidence level for all $\nu \in \mathcal{N}$.

D.6.3 Additional Results

Figures D.2 and D.3 mirror the results for Figure 5.5, focusing on cosmic rays predicted to be hadrons and true hadron cosmic rays repectively. Identifying hadrons is of lesser scientific value than identifying gamma rays, so the results here are presented mainly for reference.

D.7 Additional Results and Details on the RNA Sequencing experiment

D.7.1 Data Simulation Procedure

The `scDesign3` simulator for RNA-Seq constructs a new simulated dataset through the following steps

1. The user chooses a model type (e.g. linear with Gaussian noise) and specification to model the relationship between cell gene counts and cell features.
2. `scDesign3` estimates model parameters on the reference data.
3. The user supplies a matrix of all features of all cells in for the new simulated data.

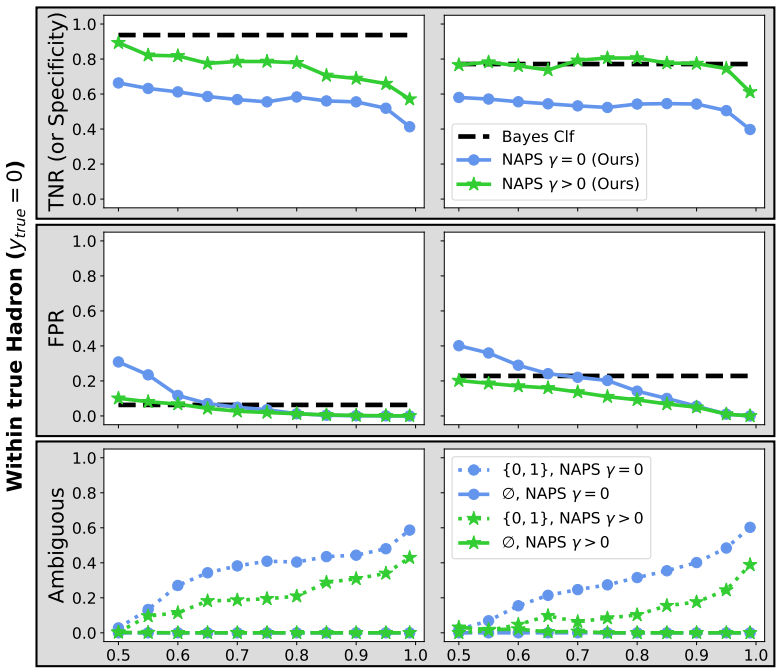


Figure D.3: **Classification metrics within true Hadrons ($y = 0$)**. Results are binned according to whether the shower energy is below (left) or above (right) the median value. Our set-valued classifier makes explicit its level of uncertainty on the label y by returning ambiguous prediction sets (bottom row) for hard-to-classify x_{target}. Even so, NAPS with $\gamma > 0$ is able to achieve a comparable number true negatives in the higher energy bins and lower number of false positives in both energy bins relative to the Bayes classifier. Here $\gamma = \alpha \times 0.3$

4. `scDesign3` outputs the gene counts for these cells by sampling from the estimated model.

In our paper, we use a negative binomial GLM with cell type and batch protocol indicator as the only features:

$$\log \mathbb{E}[X_{i,j} \mid Y_j, B_j] = \alpha_i + \beta_i Y_j + \gamma_i B_j,$$

where

1. $X_{i,j}$ are the observed counts for gene i for cell j
2. $Y_j \in \{0, 1\}$ is the cell type for cell j, `CD4`$^+$ T-cells (Y = 1) or `Cytotoxic` T-cells (Y = 0)
3. B_j is which of the 4 protocols was used to process cell j, with a separate model coefficient for each protocol excluding the baseline (represented by the vector $\gamma_i \in \mathbb{R}^3$)

We also restrict our analysis to 100 genes chosen randomly from the approximately 6000 genes in the reference dataset. Although each gene count receives its own set of model parameters, new gene counts are generated in a way that captures the correlation between

gene counts in the reference data. See (Song et al., 2023) for more details.

The reference data used in our analysis contains two experimental protocols. One is used as a baseline to derive $\widehat{\alpha}_i$. The second is used to fit the first entry of each $\widehat{\gamma}_i$, denoted $\widehat{\gamma}_{i,1}$. The last two entries $\widehat{\gamma}_{i,2}$ and $\widehat{\gamma}_{i,3}$ are constructed in this way:

1. Each $\widehat{\gamma}_{i,2}$ is sampled with replacement from $\{\widehat{\gamma}_{i,1} : |\widehat{\gamma}_{i,1}| < \text{median}(\{|\widehat{\gamma}_{j,1}|,\ j \in [100]\})\}$
2. Each $\widehat{\gamma}_{i,3}$ is sampled with replacement from $\{\widehat{\gamma}_{i,1} : |\widehat{\gamma}_{i,1}| \geqslant \text{median}(\{|\widehat{\gamma}_{j,1}|,\ j \in [100]\})\}$

These last two batch protocols are meant to emulate a weak and stronger batch effect respectively than the different between the two original experimental protocols, while keeping realistic estimates for the effects on gene counts.

D.7.2 Details on the algorithms used in Section 5.5.2

We used gradient boosting probabilistic classifiers as implemented in `CatBoost` (Prokhorenkova et al., 2018) to estimate both $\mathbb{P}(Y \mid X)$ and $W_\lambda(C; y, \nu)$. For the latter, `CatBoost` allows to easily enforce monotonicity constraints on the features, which we used on C. To compute cutoffs, we used the `brentq` routine (Brent, 2013) to calculate the inverse and the differential evolution global optimization algorithm (Storn and Price, 1997) to find the infimum. Both are implemented in `SciPy` (Virtanen et al., 2020). The three baselines against which we compare NAPS were computed from the same base probabilistic classifier (also used for NAPS). After training it, we calibrated it on the same set used for NAPS via isotonic regression, but only for the baselines (our method has a separate calibration procedure as described in Section 5.3. Then we computed cutoffs as described in Sadinle et al. (2019); Romano et al. (2020).

D.7.3 Additional Results

Taking $\texttt{CD4}^+$ T-cells (Y = 1) to be the positive class, Figures D.4, D.5, D.6, D.7 show various performance metrics for four prediction set methodologies: standard prediction sets (Sadinle et al., 2019, Theorem 1), class-conditional prediction sets (Sadinle et al., 2019), conformal adaptive prediction sets (APS; Romano et al. (2020)), and NAPS with $\gamma = 0$. For many of the metrics like precision and NPV, each method achieves very good performance (perhaps due to the ease of the underlying inference problem). For TPR, we see that each method has differing strength for each of the protocols. We also notice that at very high levels of confidence, conformal APS starts outputting $\{0, 1\}$ for every observation, leading to a sharp drop in performance across all metrics.

D.8 Computational Analysis: Training and Inference Times

Table D.1 reports training and inference times for NAPS under the Single-Cell RNA Sequencing (Section 5.5.2) and Atmospheric Cosmic-Ray Showers (Section 5.5.3) experiments. Dataset sizes are the proportions included in the training, calibration and inference sets out of the total number of simulations indicated in Sections 5.5.2 and Section 5.5.3. For calibration,

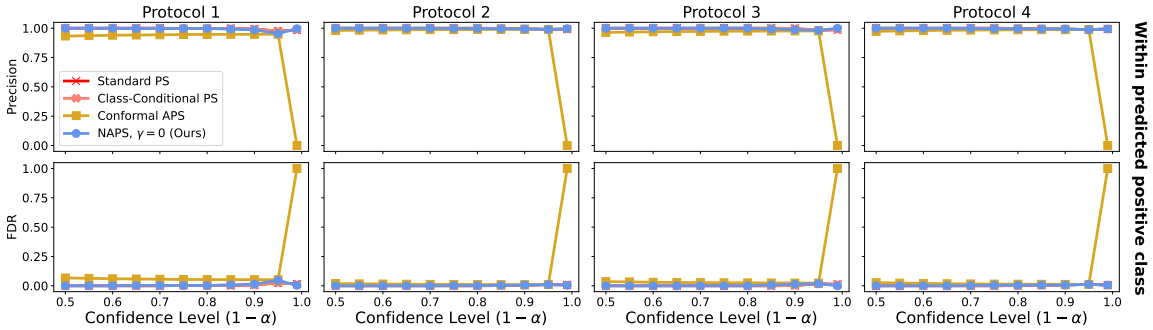


Figure D.4: **Classification metrics within predicted positive class**: Precision (top) and FDR (bottom) for observations predicted to be $\texttt{CD4}^+$ T-cells (i.e. prediction set output is $\{1\}$), additionally separated by protocol (columns). Metrics are shown for nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value.

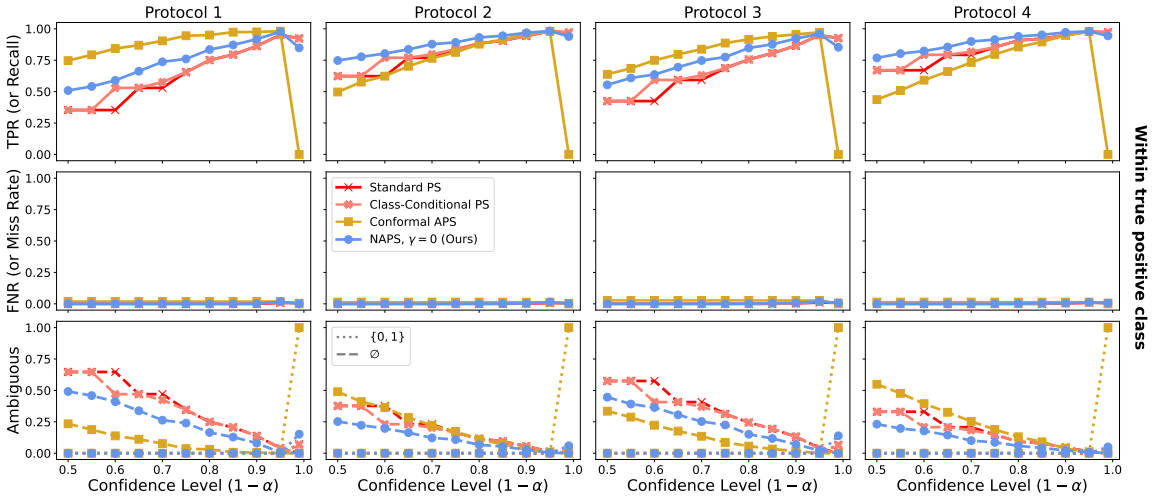


Figure D.5: **Classification metrics within true positive class**: TPR (top), FNR (middle) and proportion of ambiguous sets (bottom) for true $\texttt{CD4}^+$ T-cells, additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value.

we report the time needed to estimate ROC curves from the augmented calibration set,

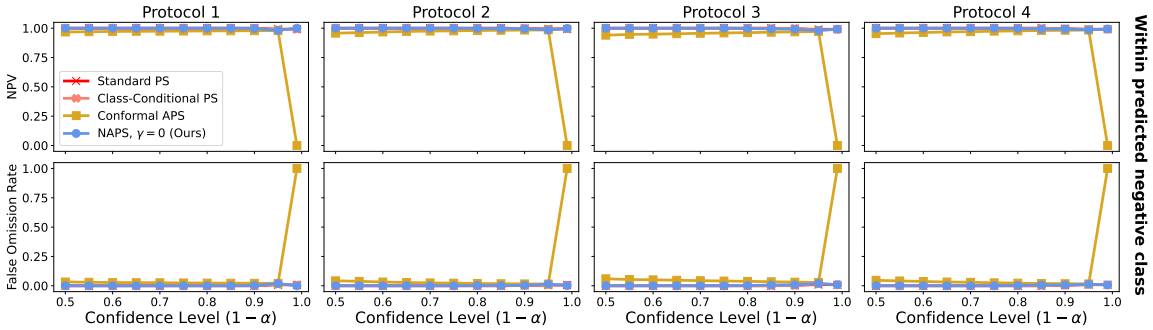


Figure D.6: **Classification metrics within predicted negative class**: NPV (top) and False Omission Rate (bottom) for observations predicted to be `Cytotoxic` T-cells (i.e. prediction set output is $\{0\}$), additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value.

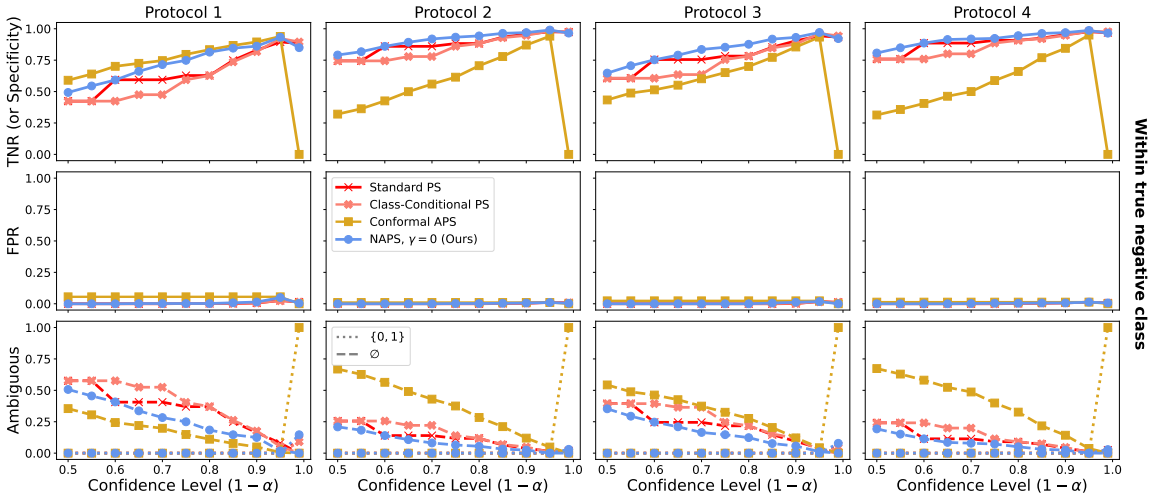


Figure D.7: **Classification metrics within true negative class**: TNR (top), FPR (middle) and proportion of ambiguous sets (bottom) for true `Cytotoxic` T-cells, additionally separated by protocol (columns). Metrics are shown for Nuisance-aware prediction sets (NAPS $\gamma = 0$; blue), standard prediction sets (red), class-conditional prediction sets (pink), and conformal adaptive prediction sets (APS) (gold). At high levels of confidence, conformal APS outputs $\{0, 1\}$ for all points in the test set; the corresponding metrics that require the prediction set to have one element have been set to their worst-case value.

including "re-calibration" of the estimated rejection probabilities via isotonic regression.

Table D.1: Training and inference times for NAPS for the experiments of Sections 5.5.2 and 5.5.3.

Experiment	Dataset Size	Training	Calibration	Inference ($\gamma = 0$)	Inference ($\gamma > 0$)
RNA-Seq	$0.6, 0.35, 0.5$	6 minutes	30 minutes	1 second	/
Cosmic Rays	$0.45, 0.45, 0.1$	8 minutes	65 minutes	6 seconds	4 seconds per-obs

For NAPS with $\gamma > 0$ (only performed in Section 5.5.3), inference times are measured per-observation (on average) since cutoffs are data-dependent and need to be computed for each x. For NAPS with $\gamma = 0$, we report the total time needed to compute cutoffs, as they can then be applied to any new observation x (i.e., they are amortized with respect to observations). Once this is done, constructing the prediction sets takes only a few milliseconds. All times are computed for inference at a single level α. Classifier training and the calibration procedure only need to be estimated once (here we report times that include five-fold cross-validation). All computations were performed on a MacBook Pro M1Pro with 16 GB of RAM.

D.9 Synthetic Example: Deep Dive

D.9.1 Impact of the Nuisance Parameter

As mentioned in the main text, we consider a process that generates events (Y_i, X_i), where $Y_i \in \{0, 1\}$ determines the type or label of the event, and $X_i \in [0, 1]$ is the sole feature of the event. The distribution of events is defined as follows:

1. $\mathbb{P}(Y_i = 0) = \mathbb{P}(Y_i = 1) = 1/2$
2. Conditional density for $Y = 1$: $p(x_i \mid Y_i = 1) = \dfrac{e^{x_i}}{e - 1}$
3. Conditional density for $Y = 0$: $p(x_i \mid Y_i = 0, \nu_i) = \dfrac{\nu_i e^{-\nu_i x_i}}{1 - e^{-\nu_i}}$,

where ν is an additional nuisance parameter that influences the density of X for $Y = 0$ events. ν_i is assumed to be drawn from some distribution independently for each $Y = 0$ event. We are interested in inferring Y given observed X and unobserved ν. Figure D.8 shows how the presence of the nuisance parameter affects this inference task.

The top left of Figure D.8 demonstrates how the shape of the density of X for $Y = 0$ events can vary dramatically depending on the value of ν. Assuming any prior of ν can yield a density of X that does not depend on ν, but it may not closely resemble the conditional densities of X given ν for all values of ν. The top right panel shows how this variation in the shape of the densities subsequently affects the behavior of the posterior probabilities of Y given X and ν. Again, we can derive a posterior that does not depend on ν, with the same caveat as before. We also observe that the posterior probabilities are always monotonic in x, therefore any classifier or prediction set that uses cutoffs on posterior probabilities can be

equivalently defined using cutoffs on x directly. The bottom left figure shows how the ROC for the Bayes Classifier (i.e. directly using the posterior probabilities to classify events) can vary under fixed ν or a prior on ν. These ROC curves demonstrate why ignoring nuisance parameters can yield biased or otherwise unreliable results. Every fixed value of ν as well as every prior on ν yields a completely different relationship between FPR and TPR. The bottom right figure shows that if our goal is valid FPR control for our inference task, we must take the nuisance parameter into account. Because the ultimate FPR for any cutoff depends on the value of ν for each observation, the selection of an cutoff that controls FPR must properly account for the influence of the nuisance parameter.

D.9.2 Additional Results

Figures D.9, D.10, and D.11 show additional results from the synthetic examples for both standard prediction sets and class-specific prediction sets used in the cosmic ray application. All prediction sets are formed under the training prior $\nu \sim \text{Uniform}(1, 10)$, which is the same prior used to compute metrics under the "No GLS" setting. "With GLS" changes the target prior to $\nu \sim \mathcal{N}(4, 0.1)$ without modifying the training prior. Coverage for $Y = 1$ events, power for $Y = 0$ events (defined as $\mathbb{P}(1 \notin \text{Prediction Set} \mid Y = 0)$), and precision for $\{0\}$ outputs do not vary significantly across methodologies due to the fact that $p(x_i \mid Y_i = 1)$ does not depend on the distribution of ν_i. As seen in the text, our methods achieve validity regardless of the presence of GLS. We also achieve higher precision than standard or class-specific prediction sets, although we do sacrifice power compared to those methods. However, careful selection of γ in the NAPS framework can help increase power without losing validity.

D.9.3 ν-Conditional Coverage and validity under GLS

Figure D.12 below explores coverage of different prediction set methods conditional on Y *and* ν, under the training prior $\nu \sim \text{Uniform}(0, 1)$. We compare 4 methods:

1. Standard prediction sets that target marginal coverage only
2. Class-conditional prediction sets that target coverage conditional on Y
3. Class-conditional prediction sets that additionally use the posterior mean $\widehat{\nu}(x) = \int_{\mathcal{N}} \nu \, p(\nu \mid x) \mathrm{d}\nu$ as an point estimate of ν to evaluate the posterior. Specifically, $P(Y = 1 \mid X, \nu = \widehat{\nu}(X))$ is used instead of $P(Y = 1 \mid X)$, where the latter integrates over the prior on ν
4. NAPS with $\gamma = 0$

Method 3 is added as a possible alternative to forming confidence sets on ν within the NAPS framework. The figure shows that, although standard and class-conditional prediction sets achieve marginal and class-conditional validity respectively, they do not maintain validity when conditioning on all values of ν. This is the fundamental reason that these methods do not achieve validity under GLS. Whereas, NAPS achieves validity conditional on both Y and ν, resulting in robustness to GLS. We note that method 3 achieves neither marginal

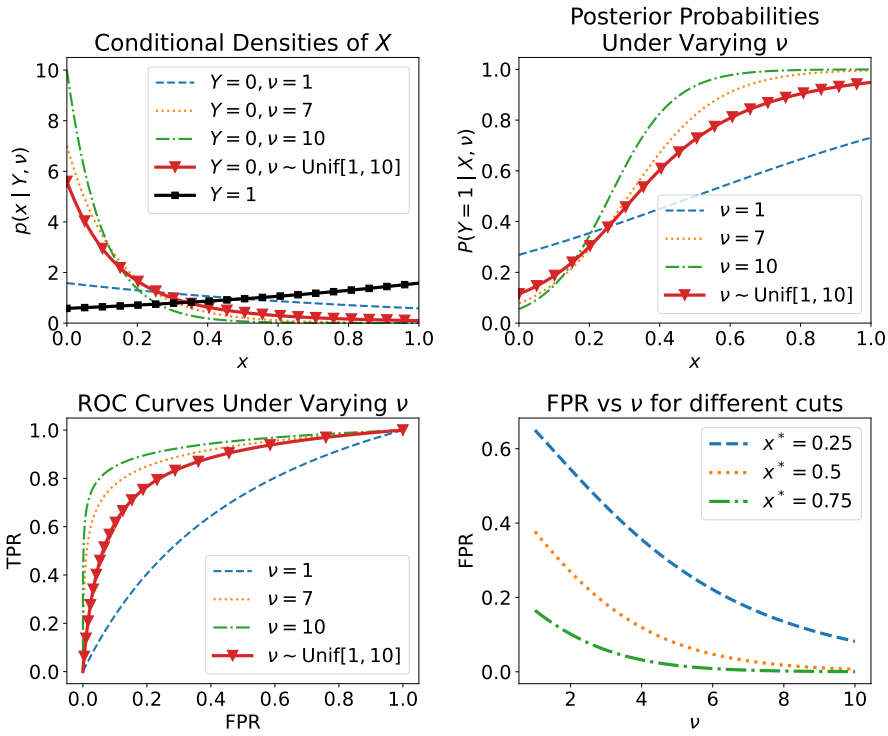


Figure D.8: **Impacts of Nuisance Parameters on the Inference Task Top Left**: Conditional densities $p(x \mid Y, \nu)$ for various values of Y and ν according to the problem setup. The marginal density $p(x \mid Y = 0)$ shown in red is induced by a Uniform$(1, 10)$ prior on ν. **Top Right**: Posterior probability $P(Y = 1 \mid X, \nu)$ as a function of X for different values of the nuisance parameter ν. The marginal posterior $P(Y = 1 \mid X)$ is shown in red for a Uniform$(1, 10)$ prior on ν. **Bottom Left**: ROC curves for the Bayes Classifier holding ν fixed (blue, orange, and green curves) and for a Uniform$(1, 10)$ prior on ν (red). $Y = 1$ is taken to be the positive class. **Bottom Right**: Under the classification rule that $\widehat{y}_i = 1$ if $x_i > x^*$, this figure shows how the FPR of that classifier will vary with ν. Each curve represents a different cut x^* for the classification rule.

nor class-conditional validity, indicating that even well-formed point estimates of ν are insufficient to reach nominal coverage levels.

D.9.4 When does $\gamma > 0$ for NAPS increase power?

The γ parameter for NAPS gives us the option to first form a confidence set for ν on a new observation x before optimizing the cutoffs for our test statistic (see Section 5.4). Because the test statistic is monotonic in the posterior probabilities, we can derive cutoffs on x directly based on the confidence set for ν. Specifically, we can simplify the procedure in Lemma 5.5 to the following

$$x_0(\nu; \alpha, \gamma) = x \quad \text{s.t.} \quad \mathbb{P}(X \geqslant x \mid Y = 0, \nu) = \alpha - \gamma$$

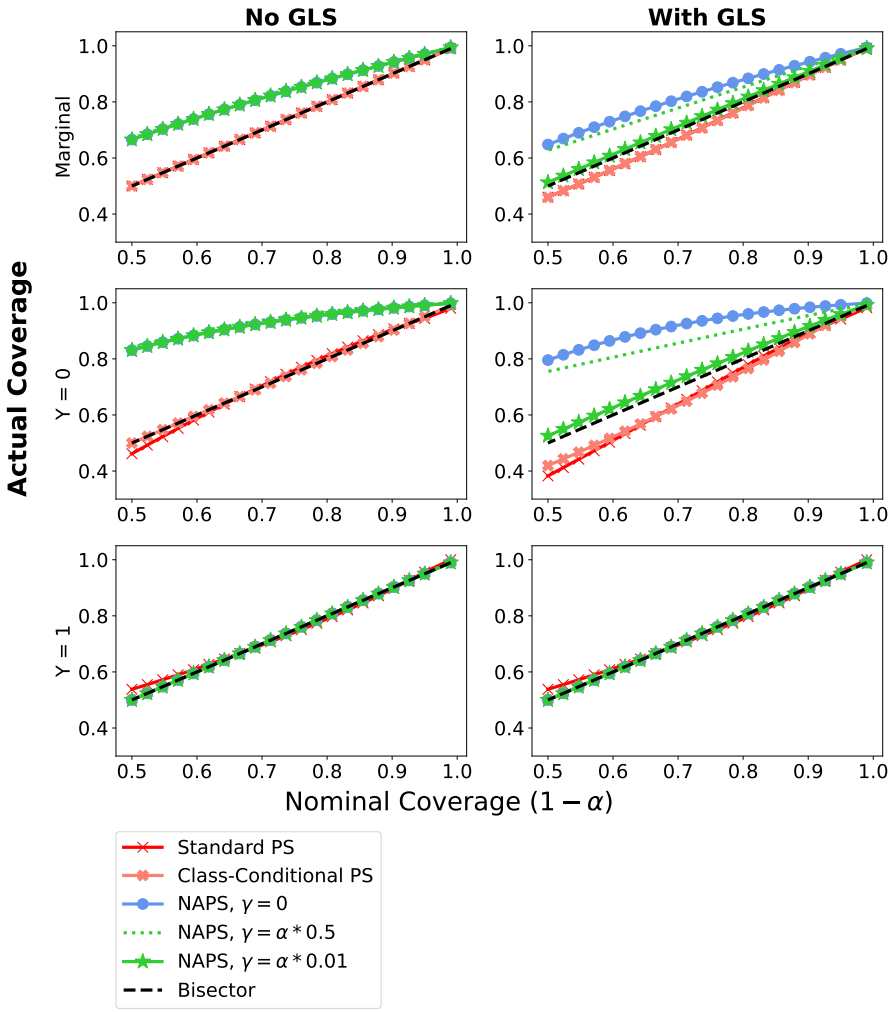


Figure D.9: **Actual vs Nominal Coverage for Several Prediction Set Methods**: We compare the actual coverage of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). We show marginal coverage (top), and conditional coverage for $Y = 0$ events (middle) and $Y = 1$ events (bottom)

$$x_0^*(\alpha) = \sup_{\nu \in S_0(x;\gamma)} x_0(\nu, \alpha)$$

$$x_1^*(\alpha) = x \quad \text{s.t.} \quad \mathbb{P}(X \leqslant x \mid Y = 1) = \alpha,$$

where $S_0(x;\gamma)$ is a $1-\gamma$ confidence set on ν given $Y = 0$. Then, our prediction set becomes

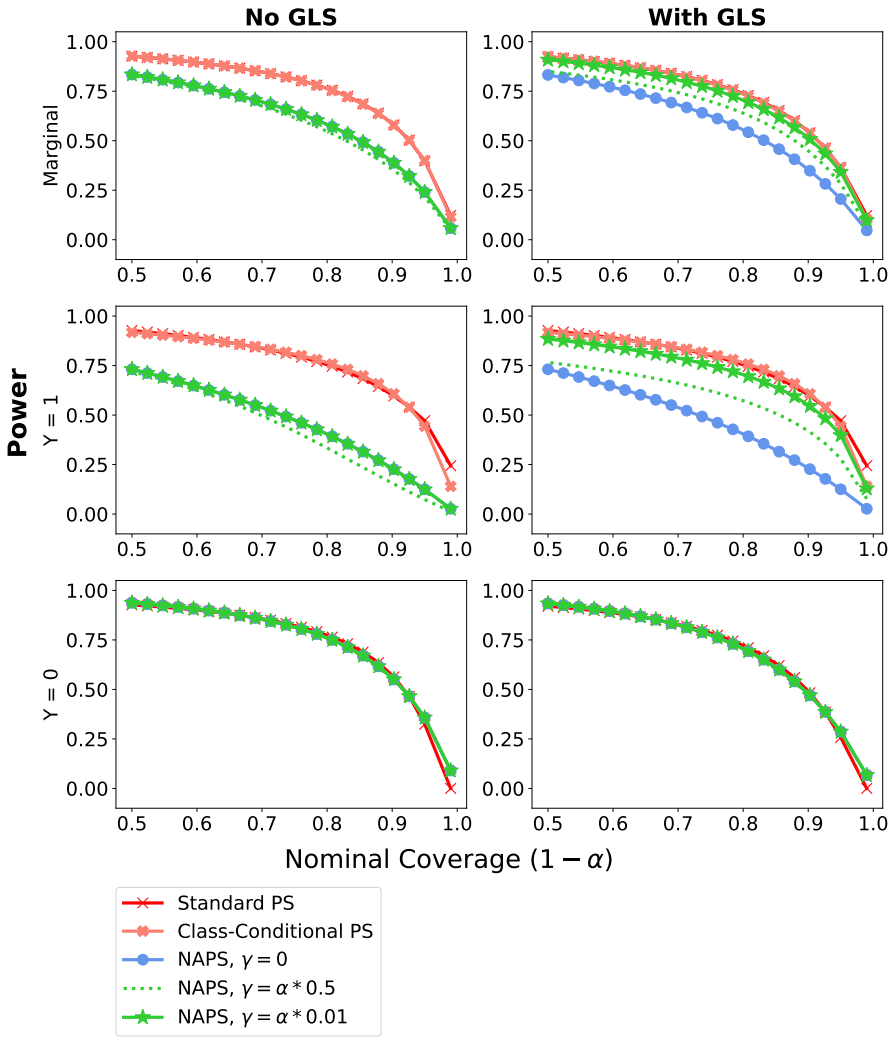


Figure D.10: **Power vs Nominal Coverage for Several Prediction Set Methods**: We compare the power of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). Power for $Y = 0$ events (bottom) is defined as $\mathbb{P}(1 \notin \text{Prediction Set} \mid Y = 0)$ and vice versa for $Y = 1$ (middle). Marginal power (top) is the sum of these two power metrics weighted by $\mathbb{P}(Y = 1)$.

$$0 \in \mathbf{H}(x; \alpha) \text{ if } x < x_0^*(\alpha)$$
$$1 \in \mathbf{H}(x; \alpha) \text{ if } x > x_1^*(\alpha).$$

We note that $x_1^*(\alpha)$ does not depend on our choice of γ, so we focus on $x_0^*(\alpha)$. We also note that lower values of $x_0^*(\alpha)$ result in higher power of the final NAPS. Figure D.13 below

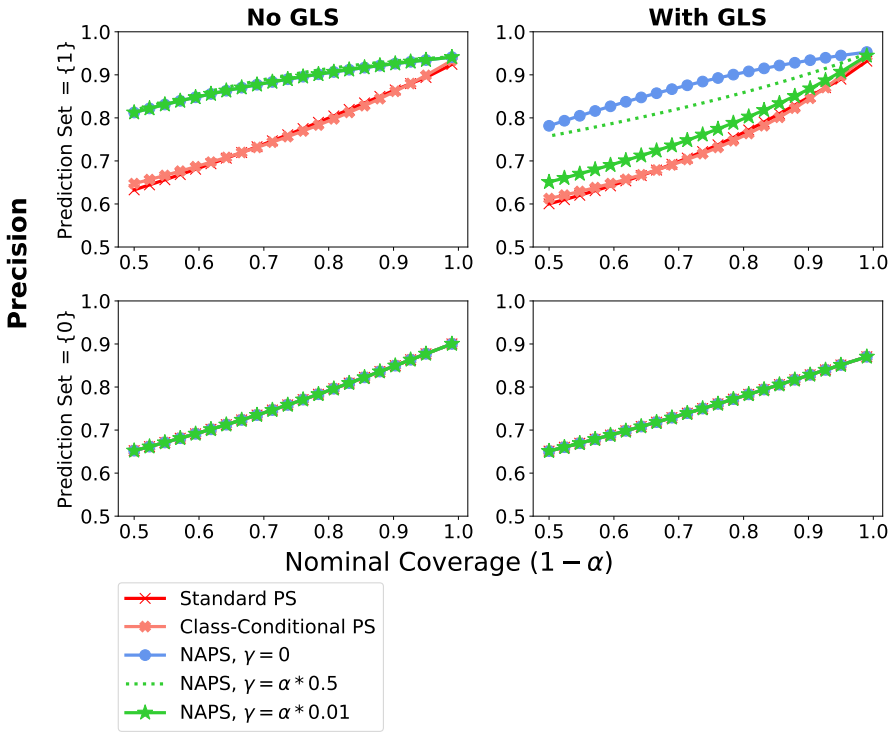


Figure D.11: **Precision vs Nominal Coverage for Several Prediction Set Methods**: We compare the precision of standard prediction sets (red), class-specific prediction sets (pink), and NAPS under different γ values under no GLS (left) and with GLS (right). We define precision for prediction set $= \{0\}$ as $\mathbb{P}(Y = 0 \mid \text{prediction set} = \{0\})$ and vice versa for prediction set $= \{1\}$ outputs. Events where prediction set $= \{0, 1\}$ or prediction set $= \varnothing$ are not considered here.

shows how the choice of γ can affect the power of the resulting NAPS.

The left panel demonstrates the tradeoff inherent in selection a value of γ. Fixing ν and α, $x_0(\nu; \alpha, \gamma)$ is increasing in γ (illustrated by the green curve being always higher than the blue curve), so the cutoff at every ν will always be higher (and power subsequently lower). However, constraining ν to $S_0(x; \gamma)$ may avoid optimizing over regions of ν where $x_0(\nu; \alpha, \gamma)$ is relatively high (i.e. small values of ν). In the synthetic example, the most power is gained when S_0 constrains ν to a region where ν is much larger than 1 (the value of ν that yields $x_0^*(\alpha)$ when $\gamma = 0$). This is illustrated by the fact that $S_0(x_2; \gamma = 0.0025)$ yields a $x_0^*(\alpha)$ value (green star) much lower than the value obtained when $\gamma = 0$ (blue star). However, setting $\gamma > 0$ can sometimes result in power *loss* if S_0 contains small values of ν. This is illustrated by the fact that $S_0(x_1; \gamma = 0.0025)$ yields an even higher $x_0^*(\alpha)$ value

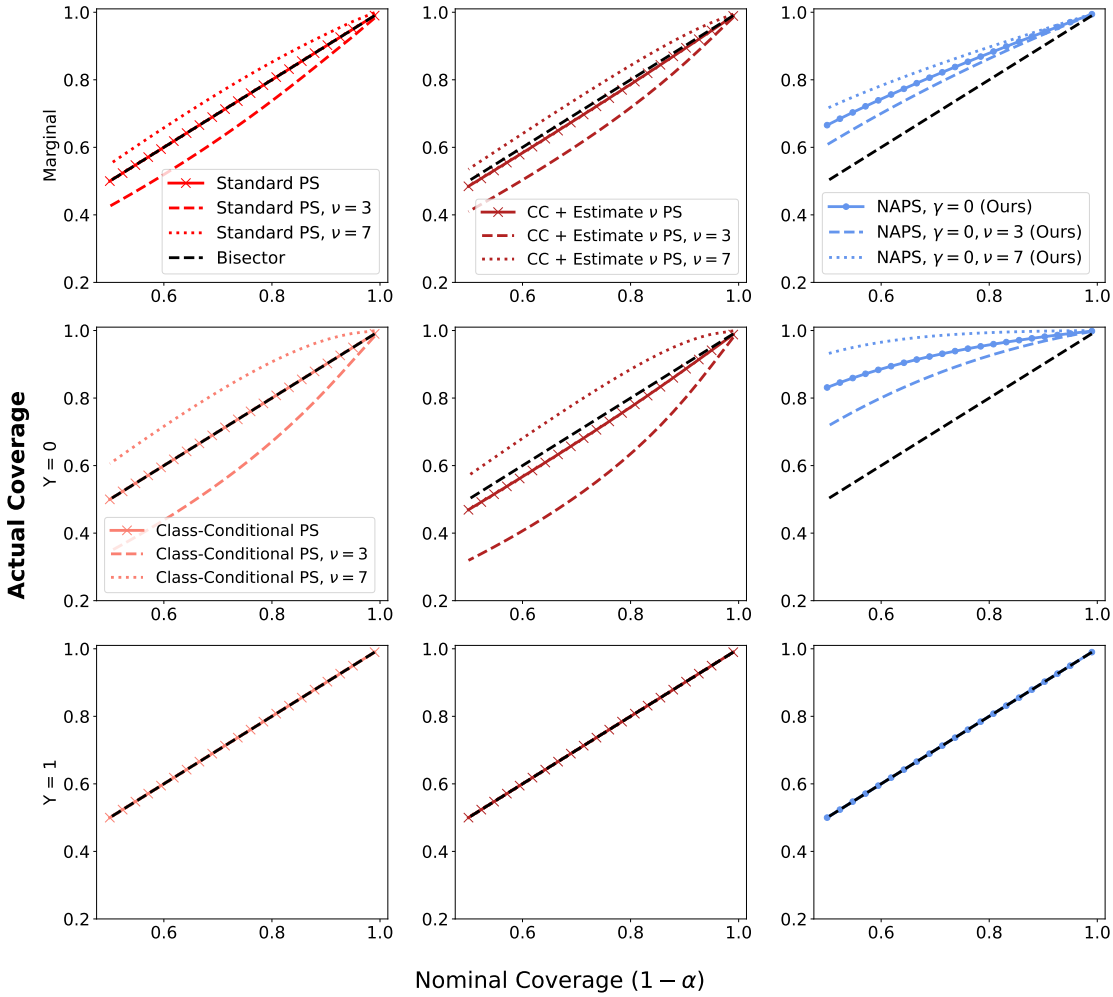


Figure D.12: **Marginal, Class-conditional, and ν-conditional Coverage of Several Prediction Set Methods**: We examine marginal coverage under the training prior on ν (top), $Y = 0$ conditional coverage (middle) and $Y = 1$ conditional coverage (bottom) for standard prediction sets (red, top right only), class-conditional prediction sets (pink, middle left and bottom left), class-conditional prediction sets with estimated ν (dark red, middle column), and NAPS (blue, right column). In each figure, we also show coverage when additionally conditioning on certain values of ν (dotted and dashed lines)

(red star) than the case when $\gamma = 0$. The right panel shows that, in our simple synthetic example, there is a relatively clear optimal value for γ which is non-zero.

In general, the distribution of the nuisance parameter(s) and the efficiency of the confidence sets on those NPs will determine which value of γ is optimal. If most data points have nuisance parameter values in "favorable" regions of the NP space, then it may be worth

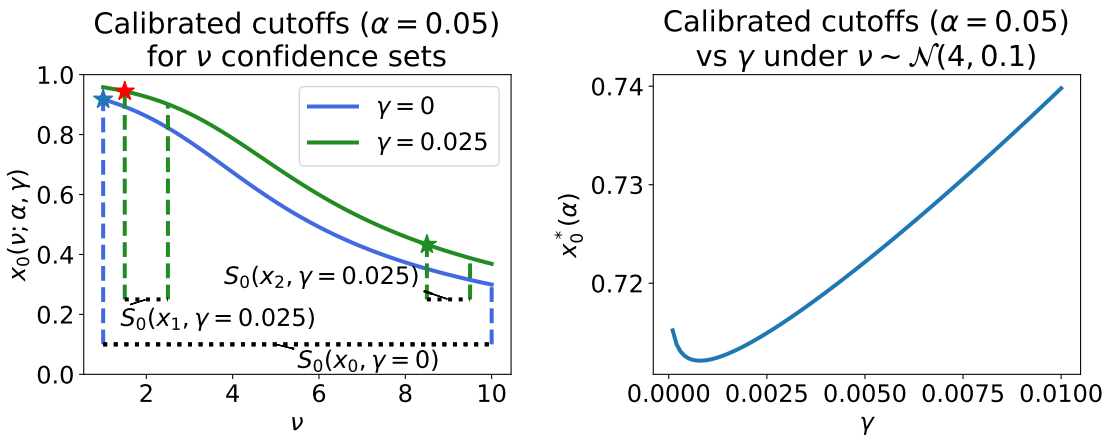


Figure D.13: **Effect of γ on NAPS Power. Left**: We show how the optimization of $x_0(\nu; \alpha, \gamma)$ depends on γ and $S_0(x; \gamma)$. The two curves show the relationship between $x_0(\nu; \alpha, \gamma)$ and ν under two values of γ. When $\gamma = 0$, we must optimize over the entire space of ν to derive $x_0^*(\alpha)$ (or equivalently, $S_0(x; \gamma = 0) = [1, 10]$ for all x. This leads to a $x_0^*(\alpha)$ value indicated by the blue star. When $\gamma = 0.0025$, we consider two hypothetical confidence sets $S_0(x_1; \gamma)$ and $S_0(x_2; \gamma)$ for ν, indicated by the two pairs of green dotted lines. In each case, we only optimize $x_0(\nu; \alpha, \gamma)$ over the values of ν in the confidence set; however, to maintain coverage at $1 - \alpha$, optimization is done over the green curve instead of the blue curve. Optimization over $S_0(x_1; \gamma)$ yields $x_0^*(\alpha)$ indicated by the red star, while optimization over $S_0(x_2; \gamma)$ yields $x_0^*(\alpha)$ indicated by the green star. **Right**: When $S_0(x; \gamma)$ is taken to be the $(\gamma/2, 1 - \gamma/2)$ quantiles of the truncated $\mathcal{N}(4, 0.1)$ distribution for all x, we can derive a relationship between $x_0^*(\alpha)$ and γ. In this case, the calibrated cutoff is minimized at $\gamma \approx 0.001$.

setting $\gamma > 0$ to form confidence sets. In other cases, letting $\gamma = 0$ may be the optimal choice.

D.9.5 Performance of NAPS under SLS

In the synthetic example, we assumed that the distribution of labels $\mathbb{P}(Y = 1)$ was the same for the training and target data. However, the distribution of ν is not the same, which leads to $p_{\text{train}}(x \mid Y) \neq p_{\text{target}}(x \mid Y)$, since

$$p(x \mid Y = y) = \int p(x \mid Y = y, \nu) \pi(\nu \mid Y = y) \, \mathrm{d}\nu$$

and we explicitly allow for a change in $\pi(\nu \mid Y = y)$ under GLS. This setup is essentially the reverse of the Standard Label Shift (SLS) setup. Under SLS, we would assume that $\mathbb{P}_{\text{train}}(Y = 1) \neq \mathbb{P}_{\text{train}}(Y = 1)$, but that $p_{\text{train}}(x \mid Y) = p_{\text{target}}(x \mid Y)$, which is most directly achieved when the distribution of ν does not change between the training and target data.

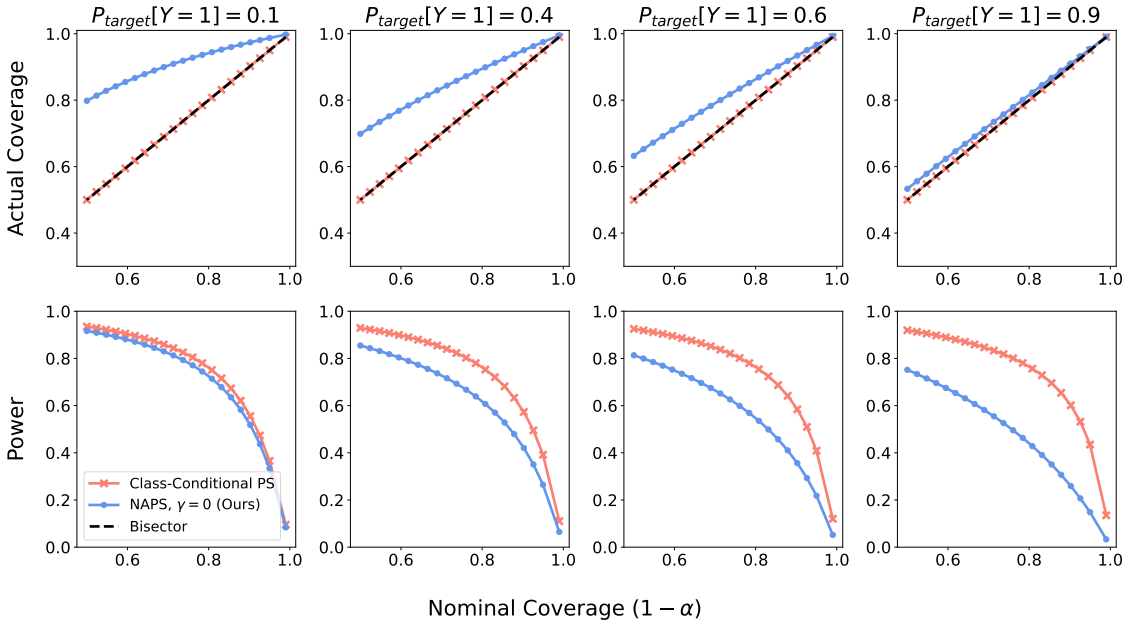


Figure D.14: **Comparison of NAPS and Class-Conditional Prediction Sets under Standard Label Shift**: We plot the test set marginal coverage (top row) and marginal power (bottom row, defined as $\mathbb{P}_{\text{target}}(1 - Y \notin \text{Prediction set})$). We compare NAPS (blue) to Class-Conditional PS (pink). This comparison is done for several levels of SLS (columns), where we shift the distribution Y in the evaluation set from $\mathbb{P}_{\text{train}}(Y = 1) = 0.5$. The distribution of the nuisance parameter ν is *the same* for training versus target data; that is, we have an SLS setting.

We have shown that class-conditional prediction sets (designed to maintain coverage under SLS) do not maintain coverage under GLS due to the violation of the assumption that $p_{\text{train}}(x \mid Y) = p_{\text{target}}(x \mid Y)$. In this section, we explore how NAPS performs in the SLS setting relative to class-conditional prediction sets. We expect NAPS coverage guarantees to hold, with a decrease in power due to NAPS enforcing nominal coverage at every point in the nuisance parameter space. Figure D.14 shows the results of our experiments under SLS.

In all SLS scenarios we tested, NAPS over-covers and achieves lower levels of power compared to class-conditional prediction sets, demonstrating the theoretical tradeoff described above. Looking at coverage, we see that as $\mathbb{P}_{\text{target}}(Y = 1)$ increases, the level of overcoverage for NAPS decreases. This is expected, since the nuisance parameter ν only affects the distribution of features for $Y = 0$ events and causes NAPS to exclude 0 from the prediction set less often. Unsurprisingly, class-conditional prediction sets exactly achieve nominal coverage under every SLS scenario.

Looking at power, we note that class-conditional prediction sets achieve similar (but not identical) power across all SLS scenarios. Power for NAPS appears to decrease as $\mathbb{P}_{\text{target}}(Y = 1)$ increases. This is a consequence of the same fact that ν only affects $Y = 0$

events; because NAPS will exclude 0 from its prediction sets less often, it will suffer a performance loss when there are relatively more $Y = 1$ events in the data. In this particular case, NAPS appears to perform best relative to class-conditional prediction sets when $\mathbb{P}_{\text{target}}(Y = 1)$ is low, but results may vary in other settings where the relationship between the nuisance parameter(s) and labels may be more complex. However, we do not expect NAPS to outperform class-conditional prediction sets (or any method developed for SLS) under SLS-only scenarios.