

Advances in the Methodology and Theory of Neural Conditional Density Models

Pierre Glaser

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Gatsby Computational Neuroscience Unit
University College London

December 17, 2025

I, Pierre Glaser, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis presents advances, made by its author and collaborators, on the theory, methodology and applications of conditional probability models. The motivation behind these is the emergence of the field of “AI for science”, which uses modern AI algorithms to produce predictions and inferences on the features of scientific data, systems and models. This field, which has already produced Nobel Prize-worthy discoveries such as AlphaFold, heavily relies on the use of conditional probability models; improving their expressiveness, training, and evaluation has the potential to further improve the quantity and quality of AI for science discoveries, and is the main challenge that this thesis addresses.

The contributions are split into two parts, containing three works each. The first part focuses on training; there, we theoretically elucidate the properties of important training methods for conditional and unconditional models. We show that (i) certain methods commonly used in Simulation-Based Inference (SBI, a subfield of AI for science) may suffer from poor simulation efficiency in certain scenarios, while (ii) other methods, commonly used for training unconditional models, are nearly optimal in that regard. Inspired by these results, we propose (iii) a new SBI method with strong theoretical guarantees on its simulation efficiency. In the second part of this thesis, we focus on building rigorous evaluation pipelines for conditional probability models. We investigate two statistical properties of such models, namely (conditional) goodness-of-fit, and reliability. Building on the existing framework of model

evaluation using reproducing kernel Hilbert spaces, we designed a (i) new reliability metric, particularly well-suited to unnormalized conditional models, samples to be estimated, (ii) evaluation metrics suited to predictive models of biological sequences, and (iii) a unified implementation of these and other kernel-based evaluation metrics.

Impact Statement

Conditional probability models play a key role in AI for science, and advances in the former can be used, as we show to improve the latter. As such, our contributions have the potential to play the role in the acceleration of scientific discoveries brought about by AI for science, which may in turn have a positive impact in fields with a direct societal impact, such as healthcare or environmental science.

In the first part of this thesis, we sought to construct training methods for conditional models that are more sample-efficient. Creating sample-efficient methods has the potential to reduce the data collection or simulation burden, which can in turn reduce the human, financial, environmental and time cost of using these methods. Moreover, many of our works contain theoretical investigations of the object we study, paving the way for more principled development of future methods.

Robust metrics of model quality, as introduced in the second part of this thesis help scientists understand how far a scientific system is from being well captured by a given model. Their output can thus be used to inform the decision of whether to invest more resources in further data collection, or deploy a model in a real-world scenario. Our contributions to reliability address the issue of calibration of AI models, which has been an important pain point even for frontier models like ChatGPT, making contributions in this area particularly timely. Finally, by providing an open-source implementation of these and other kernel-based evaluation metrics, we ensure that

these tools will be accessible to the broader scientific community.

Finally, some works presented in this thesis are the result of collaborations between machine learning researchers and domain scientists. These collaborations encourage the cross-pollination of ideas between fields, creating a more unified scientific front, helping mathematicians to focus on the right problem and scientists use better tools.

Acknowledgements

First and foremost, I would like to thank my supervisor Arthur, for his guidance and support throughout my PhD; thank you patiently sharing your expertise, putting me in contact with your amazing network of collaborators, and for making research always a fun, open-minded and positive activity. I thank Michael Arbel, who co-supervised my research internship at Gatsby in 2019, for teaching me the ropes of research without counting his time. Watching you re-write my first conference submissions until (too?) late at night was one of the greatest teaching signals I had during my time in Gatsby.

I also thank all my (other) coauthors and collaborators, Samo, Arnaud, David, Fredrik, Andrew Bolton, Vikash Mansinghka, Hudson, Aratrika, Anna, Bharath, Steffanie, Alissa, Charlotte, Debora, Alan, Kevin, Tom, Caswell, Claudia, Kim, Hugh, Peter and Ryan for their mentoring, their dedication and energy spent on our projects, and for teaching me so much about their areas of expertise.

I would like to thank my Gatsby colleagues and now friends—in particular, Will, Tom, Kevin, Antonin, Dimitri and Hugh; getting to know you was a great discovery of my PhD, and your presence deeply contributed to the amazing five years I spent in Gatsby. Thank you for your good spirit, humor and support; thank you for all the stimulating discussions we had, ranging from research to... well, almost everything else; and most of all, thank you for being such diverse yet all inspiring individuals.

To other people who played a role in my academic journey: thank you to Thomas Moreau, for his invaluable and continued mentoring since my time at INRIA in 2019; thank you to Olivier Grisel for teaching me the ways of robust software engineering. Thank you to Laelia Vaulot, for encouraging me to pursue a PhD overseas.

Finally, I would like to thank my family, Florence, Pascal, Aleth, Yves, Iris, Victor and Ariane, for their unconditional support throughout my studies. A special thought goes to my mother Florence for her love, and my father Yves, who played an instrumental role in shaping my love for mathematics and science and turning it into my career; having a close family member with whom I could discuss my work and advances in AI and neuroscience (on which you studiously picked up in recent years) is a privilege I will always be grateful for.

Contents

I	Executive Summary	21
I.1	Introduction	21
I.2	Contributions	25
I.3	Structure of the Thesis	31
I.4	Other Contributions	32
II	Background	45
II.1	Inference and Prediction via Conditional Density Estimation	46
II.1.1	(Probabilistic) Prediction	47
II.1.2	(Bayesian) Inference	47
II.1.3	Conditional Density Estimation in Inference and Prediction: Similarities and Differences	49
II.2	(Conditional) Density Estimation: Design Space and Principles	50
II.2.1	Setup and First Definitions	50
II.2.2	Density Estimation via M-estimation	51
II.2.3	A cost-driven approach to designing Density Estimators	53
II.2.4	Statistical efficiency of density estimators	55
II.2.5	Limits on the efficiency of statistical estimators	57
II.3	Comparing Distributions with Kernels	60
II.3.1	Motivation: constructing estimable distances between distributions	61
II.3.2	Reproducing Kernel Hilbert Spaces The Maximum Mean Discrepancy	63
II.3.3	Vector-Valued RKHS	66

I Contributions to the training of conditional density models 69

III A Statistical Analysis for NRE	70
III.1 Introduction and Motivation	72
III.2 Related Work	74
III.3 Background	76
III.3.1 Neural Ratio Estimation	76
III.3.2 Noise Contrastive Estimation	77
III.4 Consistency and Asymptotic Efficiency of NRE	79
III.4.1 Consistency and Asymptotic normality of NRE	79
III.5 Constructing hard distributions for NRE	83
III.5.1 Lower bounds on the asymptotic efficiency of NCE	84
III.5.2 Proof of the lower bounds on the asymptotic efficiency of INRE	85
III.5.3 Proof of Theorem III.5.1 for NRE	86
III.6 Discussion: towards finite sample bounds	96
Appendices	98
III.A Proofs of General Results	98
III.A.1 Additional Notations for NCE	98
III.A.2 General framework	99
III.B Consistency and Asymptotic Efficiency of NCE	105
III.B.1 Proof of Lemma III.B.4	110
III.B.2 Proof of Theorem III.5.2	111
III.C Proofs for Neural Ratio Estimation	112
III.C.1 Auxiliary Lemmas for NRE	112
III.C.2 Proof of Corollary III.4.1	114
III.C.3 Proof of Corollary III.4.2	117
III.C.4 Proof of Lemma III.5.3	123
III.D Auxiliary Lemmas	125
III.D.1 Differentiability of the NCE Loss	125

III.D.2	Important Identities of NCE losses	127
III.E	Background on finite-sample bounds for Logistic Regression	130
IV	Near Optimality of Contrastive Divergence Algorithms	133
IV.1	Introduction	134
IV.2	Contrastive Divergence in Unnormalized Exponential Families . .	137
IV.3	Non-asymptotic analysis of Online CD	140
IV.3.1	Preliminaries and Assumptions	140
IV.3.2	Results	142
IV.4	Non-asymptotic analysis of offline CD	146
IV.4.1	Background: Asymptotic consistency of offline CD	147
IV.4.2	Nonasymptotic consistency of offline CD	148
IV.4.3	Consistency of offline CD: beyond subexponential tails.	152
IV.5	Related Work	153
IV.6	Discussion	154
Appendices		155
IV.A	Notations	156
IV.B	Additional results for offline SGD	157
IV.B.1	An explicit finite-sample bound for SGDw	158
IV.B.2	Results for SGDo	159
IV.B.3	Explicit tail control	160
IV.C	Auxiliary Tools	162
IV.C.1	Properties of φ_γ	162
IV.C.2	Contraction and integrability results	169
IV.C.3	Miscellaneous	171
IV.D	Proofs for Online CD	172
IV.D.1	Auxiliary Lemmas for Online CD	172
IV.D.2	Proof of the SGD recursion (Lemma IV.3.1)	179
IV.D.3	Proof of Online CD convergence	181
IV.D.4	Proof of online CD with averaging (Theorem IV.3.3)	182

CONTENTS 12

IV.E	<i>L₂</i> approximation by auxiliary gradient updates	197
IV.F	Proofs for offline SGD	205
IV.F.1	Proof of Theorem IV.B.1	208
IV.F.2	Proof of Theorem IV.B.2	210
IV.G	Proofs for tail probability bounds in offline SGD	211
V	Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference	215
V.1	Neural SBI and its simulation efficiency	217
V.1.1	Background	217
V.1.2	The simulation efficiency of neural SBI methods	220
V.2	Unnormalized Neural Likelihood Estimation	223
V.2.1	Learning Conditional Energy-Based Models via maximum-likelihood	226
V.2.2	Posterior Sampling	230
V.2.3	Handling Invalid Simulations	231
V.3	Variational Inference Methods for UNLE	232
V.3.1	Variational Amortized UNLE	232
V.3.2	Double Variational Inference	236
V.4	Experiments	239
V.4.1	A toy model with a multi-modal likelihood	239
V.4.2	Results on SBI Benchmark Datasets	240
V.4.3	Using SUNLE in a Real World neuroscience model	242
Appendices		245
V.A	Filtering out invalid simulations	246
V.A.1	A probabilistic view of invalid simulations filtering	246
V.B	Coverage Study	249
V.C	Proof of Lemma V.1.1	253
V.D	Proof of Theorem V.2.1	255

V.D.1	Full statements of the setup, algorithm, assumptions, and of the theorem	255
V.D.2	Proof of Theorem 2.1	262
V.E	Proof for Double Variational Inference	295
V.E.1	Proof of Proposition V.3.2	295
V.E.2	Proof of Proposition V.3.3	297
V.E.3	Empirical improvements to DVI	297
V.F	Additional Experimental and Inferential Details	299
V.F.1	On the performance metric used for the Pyloric network .	299
V.F.2	Posterior pairplots on benchmark Problems	300
V.F.3	Manifestation of the short-run effect in UNLE	301
V.F.4	Validating the (Z, θ) -uniformization of AUNLE’s posterior in practice	302
V.F.5	Computational Cost Analysis	303
V.F.6	Experimental setup for SNLE and SMNLE	304
V.F.7	Neuroscience Model: Details	304

II Contributions to the evaluation of conditional density models

306

VI	Fast and Scalable Score-Based Calibration Tests	307
VI.1	Introduction	308
VI.2	Background	311
VI.2.1	Calibration of Predictive Models	311
VI.2.2	Kernel Conditional Goodness-of-Fit Test	314
VI.3	Kernel Calibration-Conditional Stein Discrepancy	315
VI.4	Tractable Kernels for General Unnormalized Densities	317
VI.4.1	The Generalized Fisher Divergence (Kernel)	318
VI.4.2	The Kernelized Generalized Fisher Divergence (kernel) .	320
VI.5	Fast and scalable calibration tests	323
VI.6	Experiments	324

VI.7 Conclusion	327
Appendices	329
VI.A Conditional Goodness-of-Fit: General Operator-Valued Kernel	330
VI.B KCCSD as a special case of SKCE	333
VI.C Calibration implies expected coverage	335
VI.D Diffusion-Limit and Universality	336
VI.D.1 Fisher divergence as a diffusion limit	336
VI.D.2 Universality of the Exponentiated-GFD and Exponentiated-KGFD kernel	339
VI.E Background on Stein and Fisher divergences	341
VI.F Experimental Results	345
VI.F.1 Mean Gaussian Model	345
VI.F.2 Linear Gaussian Model	348
VI.F.3 Heteroscedastic Gaussian Model	348
VI.F.4 Quadratic Gaussian Model	349
VII Kernel-based Evaluation of Conditional Biological Sequence Models	350
VII.1 Introduction	351
VII.2 Problem Setting	353
VII.3 Conditional Goodness-of-Fit with ACMMD	354
VII.3.1 The Augmented Conditional MMD	354
VII.3.2 Testing Conditional Goodness-of-Fit with ACMMD	358
VII.4 Assessing Reliability with ACMMD	359
VII.5 Related Work	363
VII.6 Experiments	364
VII.6.1 A toy synthetic setting	365
VII.6.2 ACMMD Case Study: Inverse Folding Models	366
VII.7 Discussion	372
Appendices	374
VII.A Proof of Lemma VII.3.2	375

VII.B Asymptotic distribution of $\widehat{\text{ACMMD}}^2$	378
VII.B.1 Proof of Lemma VII.3.3	379
VII.B.2 Proof of Lemma VII.3.4	380
VII.C Type-I error control of the ACMMD test	381
VII.C.1 Quantile estimation and Decision Rule	381
VII.C.2 Wild-bootstrap and permutation-based approaches are equivalent in the ACMMD test	381
VII.C.3 Level of the ACMMD test	382
VII.D Proofs related to ACMMD–Rel	384
VII.D.1 Differences between the SKCE U-statistics and the ACMMD U-statistic	384
VII.D.2 Proof of Proposition VII.4.1	384
VII.D.3 Proofs regarding the impact of approximate kernels	386
VII.D.4 Additional Details for ACMMD and ACMMD–Rel in the synthetic example	389
VII.E Additional Experiments	400
VII.E.1 Additional Experiments for the semisynthetic Protein-MPNN data	400
VII.E.2 Additional Experiments for the structural superfamily evaluation	400
VII.F Known Kernels for protein sequences and structures	401
VIII Measuring data and model properties with <code>kdiscs</code>	405
VIII.1 Introduction	406
VIII.2 The use-case for <code>kdiscs</code> : measuring distributional properties	408
VIII.2.1 Estimating (Deviations from) Distributional Properties of data and models	408
VIII.2.2 Integrating estimators into hypothesis tests	410
VIII.3 <code>kdiscs</code>	411
VIII.3.1 Overview	411
VIII.3.2 Structure and Features	413

VIII.3.3 Additional features	420
VIII.4 Experiments	425
VIII.4.1 Synthetic Data Generation Mechanism	425
VIII.4.2 Experiments and Results	427
VIII.5 Conclusion	433
VIII.6 Acknowledgements	433
Appendices	434
VIII.A List of intermediate quantities by Statistic	435
VIII.A.1 For U-statistics	435
VIII.A.2 For kernels on distributions	438
VIII.A.3 Kernels and U-statistics Kernels on Gaussian distributions	440
VIII.B Additional Plots for the experiments	443
VIII.B.1 Benchmarking All Statistics	443
VIII.B.2 CPU vs. GPU runtime	443
VIII.B.3 Breakdown of total CPU vs. GPU runtime	446
VIII.B.4 Using Quadratic vs. Linear U-statistic	448
VIII.B.5 Caching Intermediate Quantities	449

List of Figures

IV.E.1 Overview of approximation results between different updates .	199
V.0.1 SBI methods on a Multimodal Simulator	218
V.1.1 NRE and NPE, Two Moons Simulator	222
V.4.1 Performance of UNLE, Synthetic Data	241
V.4.2 UNLE(DVI) conditional pairplots	242
V.4.3 UNLE, Pyloric Model	243
V.A.1 Graphical model of the triplet (X, θ, V)	247

V.B.1	Coverage Curves (Amortized)	252
V.B.2	Coverage Curves (Sequential)	252
V.B.3	Coverage Curves (UNLE, SLCP model)	253
V.B.4	Coverage Curves (NRE, SLCP model)	253
V.B.5	Coverage Curves (NPE, SLCP model)	253
V.B.6	Coverage Curves (NLE, SLCP model)	254
V.F.1	Posterior marginal pairplots, UNLE	300
V.F.2	Normalized Posterior Densities (UNLE, Two Moons)	302
V.F.3	UNLE Posterior Samples for Different Sampling Methods	303
V.F.4	Runtime Analysis	304
V.F.5	UNLE Pairwise Marginals (Pyloric Model)	305
VI.6.1	Rejection rates for KCCSD and SKCE, Gaussian kernel	324
VI.6.2	Rejection SKCE on LGM,HMC and QGM data	324
VI.E.1	Relationships between the Fisher divergence, the KL divergence, the MMD, and the KSD [156].	344
VI.F.1	False rejection rate of the KCCSD for MGM ($\delta = 0$)	345
VI.F.2	False rejection rate of the SKCE for MGM ($\delta = 0$)	345
VI.F.3	Rejection rate of the KCCSD for MGM ($\delta = 0.1$, $c = \mathbf{1}_d$)	346
VI.F.4	Rejection rate of the SKCE for MGM ($\delta = 0.1$, $c = \mathbf{1}_d$)	346
VI.F.5	Rejection rate of the KCCSD for MGM ($\delta = 0.1$, $c = e_1$)	346
VI.F.6	Rejection rate of the SKCE for MGM ($\delta = 0.1$, $c = e_1$)	347
VI.F.7	False rejection rate of the KCCSD for LGM ($\delta = 0$)	348
VI.F.8	False rejection rate of the SKCE for LGM ($\delta = 0$)	348
VI.F.9	Rejection rate of the KCCSD for HGM ($\delta = 1$)	348
VI.F.10	Rejection rate of the SKCE for HGM ($\delta = 1$)	349
VI.F.11	Rejection rate of the KCCSD for QGM ($\delta = 1$)	349
VI.F.12	Rejection rate of the SKCE for QGM ($\delta = 1$)	349
VII.6.1	ACCMD Estimates and Rejection Rates (synthetic)	366
VII.6.2	ACCMD Estimates and Rejection Rates (ProteinMPNN)	368

VII.6.3	ACCMD-Rel Estimates and Rejection Rates (ProteinMPNN)	369
VII.6.4	ACCMD(-Rel) Estimates (CATH)	371
VII.6.5	ACCMD(-Rel) Estimates (CATH, superfamilies)	372
VII.D.1	ACMMD Rejection Rate, Synthetic Example, Appendix	399
VII.E.1	ACMMD and ACMMD-Rel (ProteinMPNN, Appendix)	400
VII.E.2	ACMMD and ACMMD-Rel (CATH Superfamilies, Appendix)	401
VIII.3.1	Running a two-sample test using <code>kdiscs</code>	413
VIII.3.2	Main type structure of <code>kdiscs</code>	414
VIII.3.3	The <code>Statistic</code> type and its main method	414
VIII.3.4	Type declaration of the <code>BaseUStat</code> class.	416
VIII.3.5	Constructing and running an aggregated test using <code>kdiscs</code>	418
VIII.3.6	Running a SKCE test requiring numerical approximations.	426
VIII.4.1	MMD and KSD rejection rate	428
VIII.4.2	MMD and KCCSD runtime	429
VIII.4.3	SKCE Runtime Breakdown (CPU)	430
VIII.4.4	SKCE Runtime Breakdown (GPU)	430
VIII.4.5	Linear vs. Quadratic tests performance (MMD)	431
VIII.4.6	Composite Tests Runtime (MMD)	432
VIII.4.7	Rejection Rate of Exact vs. Approximate SKCE Tests	433
VIII.4.8	Rejection Rate of Exact vs. Approximate SKCE Tests (2)	433
VIII.B.1	Rejection Rate Across Scenarios	443
VIII.B.2	Runtime Across Scenarios	444
VIII.B.3	Runtime Across Scenarios (contd.)	445
VIII.B.4	Runtime Breakdown Across Scenarios	446
VIII.B.5	Runtime Breakdown Across Scenarios (contd.)	447
VIII.B.6	Linear-vs-Quadratic Time Tests across Scenarios	448
VIII.B.7	Linear-vs-Quadratic Time Tests across Scenarios (contd.)	449
VIII.B.8	Cumulative Runtime, Composite Tests (MMD)	450
VIII.B.9	Cumulative Runtime, Composite Tests (KSD)	451
VIII.B.10	Cumulative Runtime, Composite Tests (KCSD)	452

VIII.B.11	Cumulative Runtime, Composite Tests (KCCSD)	453
VIII.B.12	Cumulative Runtime, Composite Tests (SKCE)	454

List of Tables

V.2.1	Comparison of the main SBI methods in their modeling choice and their objective function.	225
VIII.2.1	Properties measurable using <code>kdiscs</code> with associated measures	411
VIII.3.1	U-statistics estimators for the metrics supported in <code>kdiscs</code>	417

Introduction

CHAPTER |

Executive Summary

I.1 Introduction

The last decade has witnessed the rise and spread of deep learning, a new statistical learning paradigm which uses artificial neural networks to model data. Advances in neural architectures, learning algorithms, data collection and hardware capacity has enabled deep learning to achieve transformative breakthroughs in computer vision and natural language, leading to the creation of tools now widely used by the public and in industry. More recently however, deep learning has been successfully applied to solve scientific problems in disciplines such as particle physics, neuroscience, chemistry and medicine. The impact of deep learning in these fields can be hard to overstate: Deepmind’s Nobel-prize-winning AlphaFold, which predicts the spatial structure of proteins from their sequence, was recently estimated to have produced 1 billion human-years of research output. By increasing the pace and breadth of scientific discoveries, the field of “Artificial Intelligence (AI) for science” has the potential to drive a new era of scientific progress.

The prospects for AI in science become apparent upon realizing that a central task in

scientific inquiry consists in forming predictions (or “inferences”) about unknown quantities of interest given some known information. In protein design for instance, one may wish to predict the structure of a protein given its amino-acid sequence, and vice-versa. In neuroscience, one may wish to infer the connectivity of a neural network given observed voltage traces. The nature of the input-output relationship is rarely known exactly; however, the available information often includes known input-output pairs which may be leveraged. Additionally, the available information is rarely sufficient to make exact predictions: to be most useful, algorithms should appropriately report the plausibility of their hypotheses. The latter offers a fertile ground for *deep, probabilistic* AI—a field which seeks to build flexible, uncertainty-aware predictors from data—to be employed within science.

From a more technical standpoint, such flexible and uncertainty-aware predictors often take the form of—parametric—conditional density models which seek to approximate the true conditional distribution of the target variable given the input variable. Working with such models typically involves three distinct steps: training, inference, and evaluation. In the training phase, the parameter of these models are optimized to make the model match the true conditional distribution (which we only have partial knowledge of through a finite amount of input output pairs). The inference (or prediction) phase consists in forming actionable summaries of the trained conditional density model given a specific input variable to understand the structure of the conditional distribution. Examples of such summaries include moments, quantiles, or pairwise marginal distributions, which can easily be visualized from samples. Finally, in the evaluation phase, an algorithm—typically using samples held-out from the training phase, for reasons explained later—is invoked to quantitatively estimate how accurate and reliable the model is.

Of course, there is not a single way to construct a parametric conditional density model, nor is there a single way, given a specific conditional density model, to train it, to perform inference with it, or to evaluate it. The co-design of these components should be made to provide the downstream user with the tools best adapted to the

constraints of their problem. In the following paragraph, I will briefly explain how certain lines of research (to which my work belongs) seek to address this challenge.

Model training is a statistical operation: during it, one finds a conditional distribution model which best fits a finite amount of observed samples. As such, this operation will be noisy, i.e. inexact; one key objective for methodologists, is thus to design methods, which despite their inexact nature, will remain as accurate as possible. On the other hand, training a model has a computational cost (in the form of energy and time), which is desirable to minimize. However, seldom can one optimize accuracy and computational cost completely separately. In many cases, one has to trade off computational efficiency and accuracy.

The most evident axis in which this trade-off emerges is the number of training samples: as that number grows, so does—as guaranteed by the theory of M-estimation—the accuracy of the model; on the other hand, the cost of training the model and collecting these samples will also grow. The trade-off between accuracy and training time is well-understood by theorists and practitioners alike; usually, and unless compute cost scales very badly the number of samples, training a model on as many samples as possible (especially for complex problem) is encouraged. Sample collection on the other hand, is much more problem-dependent. In Simulation-Based Inference, a conditional density model is trained using samples following some stochastic scientific model, which is sampled from using an auxiliary program, called the *simulator*. The cost of running this simulator varies widely based on the problem at hand. For fast simulators (as in simple benchmark problems, or very efficient simulators), the cost of sample collection can be negligible. In other cases, such as neuroscience, obtaining a single sample may require solving a complex stochastic differential equation. Here, and for a given training pipeline, obtaining enough samples to reach a certain accuracy may become prohibitive. Worst, in cases where samples do not come from a simulator, but from real-world data, obtaining more samples may require running expensive experiments. In the last two cases, the situation suggest exploring ways to improve model accuracy other than by collecting

more samples.

Aside from its training data, model training is affected (amongst others) by two other factors: the *model class*, e.g. the candidate models from which the trained model is selected, and the cost function. During the last decade, an emerging body of theoretical work has documented how these two factors can significantly affect both model accuracy and training time. In fact, one trend emerging from this body of work (including in my thesis) is that in such axes too, one often ends up trading-off accuracy and computational cost. However, the current literature on this topic still presents important gaps: it mostly focuses on non-conditional density models, and some popular training objectives have not been yet analyzed.

Evaluating a model consists in quantitatively assessing certain properties of the model that matter for the end user. Accuracy—perhaps the most important one—is related to the cost function evaluated at the trained model, often up to some unknown additive, model-independent constant (such as the entropy of the data distribution) which is hard to estimate. Thus, while the value of cost functions at the optimum can be used for relative comparisons between models, they are not good measures of absolute accuracy. Second, other model properties, such as reliability, cannot be inferred from the training loss, even up to a constant; instead, assessing them requires forming and estimating a separate metric. For these two reasons (and others), model evaluation has evolved into a field distinct from model training, devoted to constructing absolute measures of model performance that can be accurately estimated from samples. Here too, trade-offs between accuracy and computational cost are present; and it is important to enrich the set of available methods—often implicitly by walking along this trade-off. However, as model evaluation is a relatively new field, advances can also be made simply by designing metrics able to capture properties of interest, for data modalities or model classes not handled by existing methods.

I.2 Contributions

At a high level, the work described in this thesis consists in theoretical and methodological advances on all three fronts of the probabilistic prediction pipeline, i.e. training, inference and evaluation. In the first part of this thesis, I present several theoretical insights on the training of conditional density models, and apply these insights to construct new, sample-efficient Simulation-Based Inference methods. In the second part, I first provide a new absolute measure of reliability for density models that can be (i) robustly and efficiently estimated from samples, and (ii) used to construct hypothesis tests better behaved than the ones relying on previously available measures; second, I construct measures of model accuracy and reliability for conditional probability models of biological sequences, which can be used to evaluate certain state-of-the-art models in computational biology, including inverse-folding models.

Part 1: Theoretical insights in training conditional density models and applications to Simulation-Based Inference

Statistical Analysis of Neural Ratio Estimation In the first part of this chapter, we perform a statistical analysis of Neural Ratio Estimation (NRE), a popular conditional density estimation method used in Simulation-Based Inference. NRE’s model is trained using a variant of Noise-Contrastive Estimation (NCE), a well-known method for training unconditional density models, which proceeds by learning the density ratio between the data distribution and an auxiliary distribution with a known density. We analyze two variants of NRE: the standard one as introduced in [101], whose loss contains a quadratic number of terms, and a more computationally friendly variant obtained by subsampling the negative examples, which we call Incomplete NRE (INRE). We establish, under specific assumptions, their consistency and asymptotic normality, and provide an expression for their asymptotic variance. Additionally, we provide examples of *hard distributions* for NRE and INRE, e.g. distributions for which the (asymptotic) mean-squared error of the NRE estimator scales exponentially with dimension. The results of our analysis show that, even in moderate dimensions,

the performance of NRE can be suboptimal, and motivate the design of alternative training methods to perform Simulation-Based Inference.

Near Optimality of Contrastive Divergence Algorithms It is well known that a certain class of statistical estimators, called Maximum Likelihood Estimators (MLE), and applicable to conditional density estimation tasks, are *asymptotically optimal*: in the large sample regime, these estimators converge their true value at the fastest rate, or, in other words, they achieve the smallest asymptotic variance amongst all possible estimators, also called the *Crámer-Rao* lower bound. MLE has previously been used for conditional density estimation; however, its use was limited to (Conditional) Normalizing Flows, a class of models that act by transforming a simple distribution using an invertible map, parameterized by a neural network. While powerful, this class of models is known to be limited in expressiveness, which can ultimately decrease their sample efficiency. Energy-Based (also called Unnormalized) Models (EBMs) constitute a popular, alternative class of models which do not suffer from these issues. However, exact MLE for EBMs is usually intractable. Contrastive Divergence (CD) is an algorithm which performs approximate MLE for unconditional models. Given its proximity to MLE, one could hope that CD inherits the optimality properties of the former. However, despite its empirical success and it being known for more than two decades, little is known about its theoretical properties, and in particular the consistency, and statistical efficiency of its estimators. Prior work has rigorously established consistency of the CD estimators at an (asymptotic) rate which is slower than the standard parametric one, matched by many alternative methods, and leaving open the question as of whether CD could, in some scenarios, recover the optimality properties of MLE.

In the second chapter of this part, I present an article in which we significantly improved the existing theoretical guarantees for CD in several directions. We first consider an “online” version of CD, in which only one training sample is processed at each training step, and which was not analyzed by prior work. For this variant, we first show, under a reduced set of assumptions compared to prior work, that CD converges to the true parameter at the optimal “parametric” rate matched by many

other methods. Second, we show that by averaging the CD iterates obtained during training, one can form an estimator with a variance which asymptotically matches the Crámer-Rao Lower Bound, up to a multiplicative factor of at most 4. Second, we analyze the standard “batched” version of CD, where multiple samples are processed at each training step; in particular, this version generalizes the “full-batch” variant analyzed in prior work (where all data points are processed at each step). We show, in the full-batch case, that under a minor strengthening of an assumption used in prior work, CD can achieve a near-parametric rate. While this variant of CD does not achieve the optimal asymptotic variance, its finite-sample variance in low-to-moderate sample size regimes can be significantly lower than that of its online counterpart, suggesting a trade-off between the two variants: asymptotic optimality for the former, and better finite-sample performance for the latter. Overall, we answer positively to whether there exists practical, approximate MLE methods for EBMs with near-optimal sample efficiency; in particular, our results show that CD can be provably more efficient than other popular learning methods, such as Noise-Contrastive Estimation and Score Matching , which are used by many SBI methods.

Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference Motivated by the empirical success and near optimality of the contrastive divergence algorithm proved in the previous section, I then present, in the last section of this chapter, a work in which we design a new Simulation-Based Inference (SBI) method based on it. SBI tackles the problem of performing inference on the parameters of a scientific model, which can be sampled from using a simulator, but which are too complex for the conditional distribution of the simulations given the parameters to be known analytically. This limitation prevents the use of exact—frequentist or Bayesian—methods to estimate the most likely model parameters linked to a specific real world observation. SBI addresses this issue by learning a neural conditional density model of the simulator in question trained on synthetic data from the simulator; after which (approximate) inference—given a real observation—can then be performed as usual. Recent insights from the literature, as well as

from the statistical analysis of NRE (a popular SBI method) performed in the first chapter of this thesis, suggest that current SBI methods based on normalizing flows or contrastive objectives may suffer from a suboptimal sample efficiency. As a consequence, such methods may require a large number of simulations to perform accurate inference, resulting in long training times when the scientific model is expensive to sample from.

To address this issue, we proposed an SBI method which trains a Conditional Energy-Based Model surrogate using a variant of the contrastive divergence algorithm, adapted to the conditional setting. We show, in our theoretical analysis, that this algorithm can, like its unconditional counterpart, achieve near-optimal sample efficiency. Our analysis is a non-trivial extension of the conditional case, where we allow the spectral gaps of the Markov Kernels used by CD (a key quantity dictating the performance of the algorithm) to concentrate arbitrarily close to 1. For the inference part of our new method, we propose two variants: (i) an “exact” one which draws samples from the learned model using an auxiliary variable MCMC algorithm, and (ii) an approximate one, which learns an approximation of the unknown log-normalizing function, after which the posterior can be sampled from using standard MCMC. After demonstrating the competitiveness of our methods and the issues of existing methods on a set of benchmark problems, we use our method to perform accurate parameter estimation in a well-known biological neural network model given observed voltage traces using only a fraction of the simulations required by the previous best method, significantly reducing the computational burden.

Part 2: Advances in Kernel-Based Evaluation of Conditional Probability Models

In a second line of research, I design rigorous evaluation metrics for predictive probabilistic models. I focused on two performance properties: accuracy (how close the model is to the true distribution) and calibration, a property ensuring that the model is neither overconfident—nor underconfident—in its predictions. Our methods can in particular be used for surrogate SBI models: in fact, a primary motivation behind this work was to address the risks of using overconfident models in SBI,

which can conceal domain scientists from credible explanations of their data and impede on scientific progress. However, our methods are general and can be applied to evaluate conditional probability models in other areas of AI for science, as shown in the second part of this chapter for inverse-folding models.

Our metrics are grounded in kernel methods, an in particular maximum mean discrepancies (MMD). MMDs are kernel-based metrics between distributions, which, remarkably, can be estimated given only samples from these distributions. While MMDs offer a promising solution to the evaluation of predictive models, current MMD-based metrics for accuracy and calibration were either impractical, or unable to handle certain important applications.

Fast and Scalable Score-Based Kernel Calibration Tests Next, I present a paper in which we introduce a new kernel-based metric able to quantify the calibration of conditional probability models, alongside with a (1) consistent estimator and (2) a hypothesis test based on this estimator designed to detect statistically significant patterns of miscalibration given finitely many data points. Our pipeline has two key benefits: our metric is able to detect any pattern of miscalibration, and is significantly easier and cheaper to estimate than its alternatives. In particular, our metric can be estimated without having to sample from the conditional probability model, which can be expensive in the common case where this model is an intractable distribution such as an unnormalized posterior. To do so, we designed a new kernel on distributions which uses a generalization of the well-known Fisher Divergence, which is of independent interest.

Kernel-Based Evaluation of Conditional Biological Sequence Models A crucial assumption of the methods developed in existing conditional goodness-of-fit and calibration methods is that the modeled data is continuous. While this assumption encompasses a broad range of use cases, it does not cover certain important problems, such as the modeling of biological sequences. In fact, until our work, existing evaluation metrics for biological sequence models either did not take into account the uncertainty in the data, or were only *relative* measures of performance: these metrics could be used to compare models, but not to assess how close a model is

to solving the task it was trained on. To fill this gap, we developed a new pair of kernel-based metrics to evaluate both the accuracy and the calibration of biological sequence models. By expanding the theory of kernels on sequence spaces, we show that our metrics are absolute measures of accuracy and calibration. We showcased the use of these metrics by evaluating ProteinMPNN, a state-of-the-art model for the problem of inverse folding (inferring the sequence of a protein given its 3D structure). Our evaluation established that ProteinMPNN was still far from solving inverse folding. We additionally used our metrics to perform hyperparameter tuning on ProteinMPNN to maximize its performance.

A unified implementation of Kernel-Based Evaluation of Data and Models The two works presented in this second part build on a larger array of kernel-based methods to evaluate the *distributional* properties of data and models, allowing in particular to answer questions like “Are two sets of samples distributed equally?”, “Is there a relationship between these two variables?” or “Does this model fit the data?” in a quantifiable manner. Despite being popular tools used by many scientists, the software ecosystem for these methods is currently fragmented, with no library providing a comprehensive and user-friendly implementation of these methods. To address this issue, we introduce `kdiscs`, a Python package for measuring data and model properties with kernels. `kdiscs` implements estimators (and accompanying hypothesis tests) of kernel-based measures of most well-known distributional properties, including equality in distribution, independence, (conditional) goodness-of-fit and calibration. `kdiscs` is designed in a modular manner, and come with multiple layers, for both (scalable) statistical estimation using (in)complete U-statistics, single hypothesis testing, and test aggregation which are both extensible and interoperable. Finally, `kdiscs` is implemented in JAX, making use of its pytree model in order to support to a very generic class of data structures.

I.3 Structure of the Thesis

The five sections comprising the two main chapters of the thesis are based on the following works, three of which have been published, and three of which are currently in preparation, and soon to be submitted.

1. Part 1, Chapter 1

Pierre Glaser and Arthur Gretton. Statistical Analysis of Neural Ratio Estimation. In preparation, 2025

2. Part 1, Chapter 2

Pierre Glaser, Kevin Han Huang, and Arthur Gretton. Near-optimality of contrastive divergence algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Q74JVgKCP6>

3. Part 1, Chapter 3

Pierre Glaser, Michael Arbel, Arnaud Doucet, and Arthur Gretton. Maximum likelihood learning of energy-based models for simulation-based inference. A previous version of this work is available at <https://arxiv.org/abs/2210.147562025>, 2025

4. Part 2, Chapter 1

Pierre Glaser, David Widmann, Fredrik Lindsten, and Arthur Gretton. Fast and scalable score-based kernel calibration tests. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2023. URL <https://proceedings.mlr.press/v216/glaser23a.html>

5. Part 2, Chapter 2

Pierre Glaser, Steffanie Paul, Alissa M Hummer, Charlotte Deane, Debora Susan Marks, and Alan Nawzad Amin. Kernel-based evaluation of conditional biological sequence models. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=2dlmcTXfcY>

6. Part 2, Chapter 3

Pierre Glaser, Antonin Schrab, and Arthur Gretton. Measuring data and model properties with `kdiscs`. In preparation, 2025

I.4 Other Contributions

Works published over the course of this thesis that are not included are

- **Pierre Glaser**, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed KL gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021
- Zonghao Chen, Aratrika Mustafi, **Pierre Glaser**, Anna Korba, Arthur Gretton, and Bharath K Sriperumbudur. (De)-regularized maximum mean discrepancy gradient flow. Accepted (with minor revisions) at JMLR. Accessible at <https://arxiv.org/pdf/2409.14980.pdf>, 2024
- Tom George, **Pierre Glaser**, Kim Stachenfeld, Caswell Barry, and Claudia Clopath. Simpl: Scalable and hassle-free optimisation of neural representations from behaviour. In *The Thirteenth International Conference on Learning Representations*, 2025
- Hugh Dance, **Pierre Glaser**, Peter Orbanz, and Ryan Adams. Efficiently vectorized MCMC on modern accelerators. In *Forty-second International Conference on Machine Learning*, 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript prepared for publication but that has not yet been published (if already published please skip to item 3):

(a) What is the current title of the manuscript? Statistical Analysis of Neural Ratio Estimation

(b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?

No

If 'Yes' please give a link or doi:

(c) Where is the work intended to be published? Scimods, or COLT

(d) List the manuscript's authors in the intended authorship order: Pierre Glaser, Arthur Gretton

(e) Stage of publication: Draft complete

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

PG derived the results. AG advised.

3. In which chapter(s) of your thesis can this material be found?

Part 1, Chapter 1

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):

John Morris

Date:

30th of September 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published please skip to item 2):

- (a) What is the title of the manuscript?

Near-Optimality of Contrastive Divergence Algorithms

- (b) Please include a link to or doi for the work:

<https://openreview.net/pdf?id=Q74JVgKCP6>

- (c) Where was the work published?

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- (d) Who published the work?

Proceedings of Machine Learning Research

- (e) When was the work published?

2024

- (f) List the manuscript's authors in the order they appear on the publication:

Pierre Glaser, Kevin Han Huang, Arthur Gretton

- (g) Was the work peer reviewed?

Yes

- (h) Have you retained the copyright?

Yes

- (i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?

No. I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

PG conceived the project and the main theoretical framework. PG led the analysis of the online variant of CD. KHH led the analysis of the offline variant of CD, with some minor help from PG. PG wrote the paper, helped by KHH for the offline variant of CD. AG advised.

3. In which chapter(s) of your thesis can this material be found? Chapter 1

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):



Date:

30th of September 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript prepared for publication but that has not yet been published (if already published please skip to item 3):

(a) What is the current title of the manuscript? Maximum likelihood learning of energy-based models for simulation-based inference

(b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?

Yes

If 'Yes' please please give a link or doi: <https://arxiv.org/abs/2210.14756>

(c) Where is the work intended to be published? Journal of Machine Learning Research

(d) List the manuscript's authors in the intended authorship order: Pierre Glaser, Michael Arbel, Samo Hromadka, Arthur Gretton, Arnaud Doucet

(e) Stage of publication: Resubmission, new draft complete

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

MA initiated the project. PG derived the methodology, theory, and ran the experiments. MA, AG and AD helped with the write-up. SH helped with the experiments. AG and AD advised.

3. In which chapter(s) of your thesis can this material be found? Part I, Chapter 3
- e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):

John Peter

Date:

30th of September 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published please skip to item 2):

- (a) What is the title of the manuscript?

Fast and Scalable Score-Based Kernel Calibration Tests

- (b) Please include a link to or doi for the work:

<https://proceedings.mlr.press/v216/glaser23a/glaser23a.pdf>

- (c) Where was the work published?

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence

- (d) Who published the work?

Proceedings of Machine Learning Research

- (e) When was the work published?

2023

- (f) List the manuscript's authors in the order they appear on the publication:

Pierre Glaser, David Widmann, Fredrik Lindsten, Arthur Gretton

- (g) Was the work peer reviewed?

Yes

- (h) Have you retained the copyright?

Yes

- (i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?

No. I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

PG conceived the project and the methodology (the calibration metric and the score-based kernels). PG and DW collaborated on the theory, paper write-up and experiments. AG and FL advised.

3. In which chapter(s) of your thesis can this material be found? Part 2, Chapter 1

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):



Date:

30th of September 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published please skip to item 2):

- (a) What is the title of the manuscript?

Kernel-Based Evaluation of Conditional Biological Sequence Models

- (b) Please include a link to or doi for the work:

<https://proceedings.mlr.press/v235/glaser24a.html>

- (c) Where was the work published?

Proceedings of the 41 st International Conference on Machine Learning

- (d) Who published the work?

Proceedings of Machine Learning Research

- (e) When was the work published?

2023

- (f) List the manuscript's authors in the order they appear on the publication:

Pierre Glaser, Steffanie Paul, Alissa Hummer, Charlotte Deane, Debora Marks, Alan Amin

- (g) Was the work peer reviewed?

Yes

- (h) Have you retained the copyright?

Yes

- (i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)?

No. I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

PG conceived the project and the methodology, developed the theory (with help from AA) and experiments (with help from AA, SP and AH), and wrote the paper (with help from SP). CD and DM advised.

3. In which chapter(s) of your thesis can this material be found? Part 2, Chapter 2

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):



Date:

30th of September 2025

UCL Research Paper Declaration Form

Referencing the doctoral candidate's own published work(s)

1. For a research manuscript prepared for publication but that has not yet been published (if already published please skip to item 3):

(a) What is the current title of the manuscript? Measuring data and model properties with kdiscs

(b) Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?

No

If 'Yes' please give a link or doi:

(c) Where is the work intended to be published? JMLR OSS, or NeurIPS

(d) List the manuscript's authors in the intended authorship order: Pierre Glaser, Antonin Schrab, Arthur Gretton

(e) Stage of publication: Draft Complete

2. For multi-authored work please give a statement of contribution covering all authors (if single-author please skip to item 4):

PG wrote the paper and the code. AS helped with the composite testing code.
AG advised.

3. In which chapter(s) of your thesis can this material be found? Part 2, Chapter 3

e-Signatures confirming that the information above is accurate (this form should be co-signed by the supervisor/senior author unless this is not appropriate e.g. if the paper was a single-author work):

Candidate:



Date:

16th of July 2025

Supervisor/Senior Author signature (where appropriate):

John Morris

Date:

30th of September 2025

CHAPTER II

Background

This background section serves two main purposes:

- The first is setting up a unifying technical thread behind the contributions of this work by positioning conditional density estimation as one key statistical method behind two scientific tasks important in modern AI for science: probabilistic prediction and Simulation-Based Inference.
- The second is to provide a summary and a rationale of the main theoretical concepts manipulated in this thesis. These concepts are, to a large extent, assumed to be known and understood by the readers in the papers of this thesis, and this background section should place the unfamiliar reader in a position to understand the papers without resorting to (too many) external references. *This background section is thus not designed to be a re-hash of each paper’s background, introduction and related works sections, but rather a nice complement to them.*

In Section II.1, we introduce the two tasks of interest in this thesis, probabilistic prediction and Simulation-Based Inference, and show how both can be solved by estimating a conditional density from data. In Section II.2, we formalize the problem of (conditional) density estimation, and propose a set of considerations that should guide the design of such algorithms, namely computational and statistical efficiency. The last part of this section delves into the precise meaning of the latter term, as statistical efficiency plays a key role in the first part of this thesis’s contributions. Finally, Section II.3 introduces the kernel-based methodological framework that the contributions of the second part of this thesis relies on.

II.1 Bayesian Inference and Probabilistic Prediction through Conditional Density Estimation

In many applications, the task of predicting (or inferring) a certain quantity given some information about it is done *under uncertainty*: the available information is simply not enough to exactly pinpoint the quantity of interest. For instance, it is known [274, 223, 144] that many protein sequences can fold to a given 3-dimensional

structure. In such cases, a good prediction mechanism should report not one, but multiple possible values of the quantity of interest, along with their likelihoods. One way of describing (or modelling) such “one-to-many” systems is through the formalism of *probability distributions* and *random variables*. As we will quickly see, under this formalism, the best possible predictor becomes the *conditional probability distribution* of the target given the available information.

Next, we describe the two tasks considered in this thesis: *Probabilistic Prediction* and *Simulation-Based Inference*. As prediction and inference do not have a precise mathematical meaning [252], the differences between the two remain a subject of debate [114, 253, 113] in the statistics and machine learning community. With the definitions and notations we give next, we only seek to align with the use of these terms in the subfields of interest in this thesis, and to which our contributions belong.

II.1.1 (Probabilistic) Prediction

In the prediction setting, we denote by X and Y the random variables describing the observed information (also called the *input*) for the problem at hand, and Y the random variable representing the quantity of interest to predict (or the *output*). In that context, we denote $p_X(x \in \cdot)$ and $p_Y(y \in \cdot)$ the probability measures respectively on \mathcal{X} and \mathcal{Y} representing the (probability) distribution of the input and the output in the real world. Given some possible input value x , the *conditional probability* of Y given $X = x$, noted $p_{Y|X}(y \in \cdot | x)$, represents remaining set of possible values for Y , alongside with their respective plausibility, once we know that the input equals x . This conditional distribution is of course unknown; the approach to probabilistic prediction that we consider in this thesis will be to estimate it given a set of n input-output pairs $(X^{(1)}, Y^{(1)}), \dots, (X^{(n)}, Y^{(n)}) \sim p_{X,Y}((x,y) \in \cdot)$ collected from the real world.

II.1.2 (Bayesian) Inference

In the inference setting, the observed information X is *posited* to follow a certain stochastic scientific model, e.g. a probability distribution $p_{X|\Theta}(x \in \cdot | \theta)$ which depend on the model parameters θ . Assuming that the link from X to θ is one-to-

many, the goal of inference is then, given an observation X , to find the parameters that may have plausibly generated it: as in prediction, the goal can thus be formalized as computing the conditional distribution of θ given X , noted $p_{\Theta|X}(\theta \in \cdot | x)$. This setting significantly differs from the previous one:

- on the one hand, no real-world observations of θ , let alone of (θ, X) pairs, are available: thus, one cannot, *a priori*, solve this problem by fitting a model for $p_{\Theta|X}$.
- On the other hand, by assumption, the user already has “good” model for $p_{X|\Theta}$ at its disposal. This model is linked to the quantity of interest $p_{\Theta|X}$ through the celebrated Bayes’ rule:

$$p_{\Theta|X}(\theta \in \cdot | x) = p_{\Theta}(\theta \in \cdot) \frac{dp_{X|\Theta}}{dp_X}(x, \theta) \quad (\text{II.1})$$

where $\frac{dp_{X|\Theta}}{dp_X}(x, \theta)$ denotes the Radon-Nikodym derivative of $p_{X|\Theta}$ with respect to p_X . Yet, the marginal distribution of θ and the one on X , both present on the right-hand side of (II.1), are unknown, and prevent a direct computation of the posterior distribution.

In its simplest form, Bayesian inference refers to the set of methods that produce an approximation of $p_{\Theta|X}$ by leveraging Bayes’ rule while handling the unknown marginal distributions of θ and X in the following manner:

1. requiring the user to provide an estimate for marginal distribution p_{Θ} , which will henceforth be treated as the true one
2. the creation of algorithms able to approximately sample from $p_{\Theta|X}$ without the need to access the marginal p_X , such as Markov Chain Monte Carlo (MCMC) methods [178] or Variational Inference (VI, 131)

Simulation-Based Bayesian Inference For many modern scientific models, another point of complexity arises. The conditional distribution $p_{X|\Theta}$, assumed tractable above, may not be available in closed form, preventing a direct application of

Bayesian inference algorithms. Nevertheless, in such cases, it may be possible to sample from $p_{X|\Theta}$ given any value for θ . Bayesian Simulation-Based Inference [47] is a family of methods that produce an estimate of $p_{\Theta|X}$ using samples $\{(\theta_i, X_i)\}_{i=1}^n \sim \pi(\theta \in \cdot) p_{X|\Theta}(x \in \cdot | \theta)$ obtained by sampling parameters from a proposal distribution π , and sampling from the model $p_{X|\Theta}$ given these parameters. The benefit of introducing a proposal possibly differing from the prior, as explained later on, is to tailor the approximation of $p_{\Theta|X}(\theta \in \cdot | x)$ to a specific observation value x_o .

Over the last decade, various approaches have been introduced to solve this problem; one of them, pioneered by Wood [272], Papamakarios et al. [190] and further explored in this thesis [249], consists in using these samples to form an estimate $p_{\psi_n}(x \in \cdot | \theta)$ of the intractable likelihood $p_{X|\Theta}$ and then applying Bayes' rule (Equation (II.1)) to obtain estimate of the posterior given by distribution

$$p_{\Theta}(\theta \in \cdot) \frac{\frac{dp_{\psi}}{d\mu}(x | \theta)}{Z(\psi)} (\approx p_{\Theta|X}(\theta \in \cdot | x)), \quad Z(\psi) := \int_{\Theta} \frac{dp_{\psi}}{d\mu}(x_o | \theta) p_{\Theta}(d\theta) \quad (\text{II.2})$$

where we assumed that $p_{X|\Theta}, p_X, p_{\psi} \ll \mu$ for some known σ -finite measure μ , such as the Lebesgue measure or the counting measure for discrete spaces. $Z(\psi)$ is an unknown normalizing constant (independent of θ), making this posterior only known up to that constant; such unnormalized posterior can then be approximately sampled from using methods designed for that purpose, such as MCMC or Variational Inference.

II.1.3 Conditional Density Estimation in Inference and Prediction: Similarities and Differences

Above, we have introduced two important tasks in modern machine learning, probabilistic prediction and Simulation-Based Inference, that can both be tackled by estimating a conditional distribution (or density) from data. *Such conditional density estimators (CDE) constitute the key statistical object of interest of this thesis.* As a result, certain conditional density estimation algorithms developed within the SBI literature can apply to probabilistic prediction tasks, and vice versa [272, 190, 249].

Nonetheless, the CDE aspects of these two field do not overlap: indeed, in SBI, the marginal distribution p_{Θ} is known, while in probabilistic prediction, the marginal distribution p_Y is unknown. Consequently, SBI techniques that leverage p_{Θ} within their CDE cannot be directly applied to probabilistic prediction tasks. Neural Ratio Estimation, which we discuss in Chapter III, is an example of such a technique.

In the next section, we formalize and do a deeper dive into the well-established field of density estimation. The methodological, practical and theoretical insights developed in the unconditional case will serve as an important foundation to study conditional methods.

II.2 (Conditional) Density Estimation: Design Space and Principles

II.2.1 Setup and First Definitions

Conditional Density Estimation Following the notations used in the SBI section, we now introduce the formalism of conditional density estimation. Again assuming the random variables (X, θ) with their respective laws, (with $p_{X|\Theta}(x \in \cdot | \theta) \ll \mu$ for all θ , and for some known σ -finite measure μ) and an additional ‘‘proposal’’ distribution π (which can here be colluded with the prior) on Θ are given, conditional density estimation refers to the set of algorithms \mathcal{A} returning an approximation of $f^*(x, \theta) := \frac{dp(x \in \cdot | \theta)}{d\mu}(x | \theta)$, the density (e.g. Radon-Nikodym derivative) of $p_{X|\Theta}$ w.r.t μ , given samples

$$\mathcal{D}_n := (\theta^1, X^1), \dots, (\theta^n, X^n) \stackrel{\text{i.i.d.}}{\sim} \pi(\theta \in \cdot) p_{X|\Theta}(x \in \cdot | \theta).$$

We focus in this work on the *parametric* setting, where the possible approximations are given by the family $\{p_{\psi}(x \in \cdot | \theta = \diamond), \psi \in \Psi \subseteq \mathbb{R}^p\}$. Here, $p_{\bullet} : \psi \in \Psi \mapsto p_{\psi}$ is a *model* mapping ψ to some conditional distribution p_{ψ} from Θ to \mathcal{X} . In the following, we assume that such algorithms return a tuple (ψ_n, p_{ψ_n}) , where $\psi_n \in \Psi$ is a parameter estimate, and p_{ψ_n} is the corresponding conditional distribution.

Link with density estimation A conditional density estimation algorithm \mathcal{A} can be mapped into a density estimation algorithm \mathcal{A}_u , which, given the output (ψ_n, p_{ψ_n}) of \mathcal{A} , returns $(\psi_n, p_{\psi, \pi})$, where $p_{\psi, \pi}((x, \theta) \in \cdot) := \pi(\theta \in \cdot)p_{\psi_n}(x \in \cdot | \theta)$ is an estimator of the joint distribution $p_{X, \Theta_\pi}((x, \theta) \in \cdot) := \pi(\theta \in \cdot)p_{X|\Theta}(x \in \cdot | \theta)$. As such, some important properties of density estimation algorithm will also hold for conditional ones, as we will see in the next section. Nevertheless, the conditional density estimation algorithms considered in this thesis will have important algorithmic differences with their unconditional counterparts, and thus require their own analysis.

An important class of special case: well-specified problems We call a (conditional) density estimation problem is called *well-specified* if there exists a *unique* $\psi^* \in \Psi$ such that $p_{X|\Theta} = p_{\psi^*}$. In well-specified problems, it should be expected that the estimated model p_{ψ_n} approaches the true model p_{ψ^*} in the large sample limit.

II.2.2 Density Estimation via M-estimation

II.2.2.1 M-estimation

Definition Let $d(p_\psi, p_{X|\Theta})$ be a notion of distance between p_ψ and $p_{X|\Theta}$. The best approximation (according to d) of $p_{X|\Theta}$ within Ψ is then given by minimizing that distance over Ψ :

$$\psi^* := \arg \min_{\psi \in \Psi} d(p_\psi, p_{X|\Theta})$$

Of course, obtaining ψ^* by solving the problem above is impossible as the r.h.s (which includes the unknown $p_{X|\Theta}$) is not known. However, assuming access to an estimator of $d_n(\psi, \mathcal{D}_n)$ of it given n samples of $p_{X|\Theta}(x \in \cdot | \theta)$ (possibly, up to an additive constant independent of ψ), one could form an estimate ψ_n of ψ^* by minimizing the *estimate* of that distance over Ψ :

$$\psi_n := \arg \min_{\psi \in \Psi} d_n(\psi, \mathcal{D}_n) \approx \arg \min_{\psi \in \Psi} d(p_\psi, p_{X|\Theta}) = \psi^*$$

Estimators ψ_n obtained in the case of Equation II.3 are called *M-estimators* [115] (the “M” standing for “minimization”). They were originally introduced for the

special case of sample-average type estimates

$$d_n(\psi, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n \ell(X_i, \theta_i; \psi) \quad (\text{II.3})$$

before being generalized to arbitrary functions of the sample and the parameter.

Consistency Below, we restate a well-known result showing that, under appropriate conditions, this procedure is principled in two different senses.

Theorem II.2.1 ([257, p. 74],[256, Theorem 5.7]). *Assume that*

$$\sup_{\psi \in \Psi} |d_n(p_\psi, \mathcal{D}_n) + C - d(p_\psi, p_{X|\Theta})| \xrightarrow{\text{P}} 0$$

for some $C \in \mathbb{R}$. Then it holds that

$$d(p_{\psi_n}, p_{X|\Theta}) \xrightarrow{\text{P}} d(p_{\psi^*}, p_{X|\Theta})$$

If, moreover $\inf_{\|\psi - \psi^*\| \geq \varepsilon} d(p_\psi, p_{X|\Theta}) - d(p_{\psi^*}, p_{X|\Theta}) > 0$, for any $\varepsilon > 0$ then furthermore, we have $\psi_n \xrightarrow{\text{P}} \psi^*$.

Theorem II.2.1 shows that, under a *uniform law of large numbers* assumption, M-estimation converges to the best estimator in the sense of d . If, additionally, an *identifiability* condition holds, then additionally ψ_n converges to ψ^* . We defer to a few sections a more in-depth study of the properties of M-estimators. The proof of this result is very simple; see, e.g. [257, p. 74]. The “hard” part is to find suitable model classes and distances for which the uniform law of large numbers holds.

II.2.2.2 An important example: (Conditional) Maximum Likelihood Estimation

Perhaps the most famous density estimation algorithm is Maximum Likelihood Estimation (MLE), first formally introduced by Fisher in [68, 69]. Using the notations introduced above, the (conditional) maximum likelihood estimate $\psi_{n,\text{MLE}}$ of ψ^* is

given by:

$$\begin{aligned}\psi_{n,\text{MLE}} &:= \arg \max_{\psi \in \Psi} \frac{1}{n} \sum_{i=1}^n \log \frac{dp_{\psi_n}}{d\mu}(X_i | \theta_i) \\ &= \arg \max_{\psi \in \Psi} \frac{1}{n} \sum_{i=1}^n \log \frac{dp_{\psi,\pi}}{d(\mu \otimes \pi)}(X_i, \theta_i)\end{aligned}\quad (\text{II.4})$$

The second line shows that the conditional MLE with model p_ψ of $p_{X|\Theta}$ is equivalent to the MLE with model $q_{\psi,\pi}$ of the joint distribution p_{X,Θ_π} , bridging the gap between conditional and unconditional density estimation. There are many ways to justify the use of MLE. One of them is that, under integrability conditions, and assuming $p_{X|\Theta} \ll \mu$ for all θ , the average of its objective function equals, up to a normalizing constant, (minus) a notion of distance known as the Kullback Leibler (KL) divergence between the estimated model $p_{\psi,\pi}$ and the true model p_{X,Θ_π} :

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{dp_{\psi,\pi}}{d(\mu \otimes \pi)}(X_i, \theta_i) \right] \\ = -\text{KL}(p_{X,\Theta_\pi}, p_{\psi,\pi}) + \mathbb{E} \left[\log \frac{dp_{X,\Theta_\pi}}{d(\mu \otimes \pi)}(x, \theta) \right], \quad \text{KL}(p, q) := \int \log \frac{dp}{dq} dp.\end{aligned}$$

Importantly, we have that $\text{KL}(p, q) \geq 0$, and $\text{KL}(p, q) = 0$ if and only if $p = q$. MLE thus minimizes a loss whose expectation yields another one which, assuming well-specification, is optimized at the true distribution. One may thus hope that, under suitable conditions, for large n ,

$$\psi_{n,\text{MLE}} := \arg \max_{\psi \in \Psi} \frac{1}{n} \sum_{i=1}^n \log \frac{dp_{\psi_n}}{d\mu}(X_i | \theta_i) \approx \arg \max_{\psi \in \Psi} \text{KL}(p_{X,\Theta_\pi}, p_{\psi,\pi}) = \psi^*$$

At this point, MLE may seem rather arbitrary, and it is not clear, why MLE plays such an important role in Statistics and Machine Learning. However, as we will see soon after, MLE enjoys desirable statistical properties.

II.2.3 A cost-driven approach to designing Density Estimators

Before doing so, we first briefly formalize the cost-accuracy trade-off of density estimation algorithms first discussed in the introduction, and the light it sheds on their design:

- Let $a(n, \mathcal{A}) > 0$ be a measure of the expected accuracy of the estimated model ψ_n . For instance, $a(n, h_{\mathcal{A}}) = \mathbb{E} [d(p_{\psi_n}, p_{X|\Theta})]$, where $d(\bullet, \diamond)$ is a notion of divergence between distributions.
- Let $c(n, \mathcal{A})$ denote the expected cost (say, in seconds) of obtaining the result of algorithm \mathcal{A} on the dataset \mathcal{D}_n . As discussed in the introduction, this cost can be further broken down into two sub-costs:
 1. The cost of executing the algorithm \mathcal{A} on \mathcal{D}_n , noted $c_1(n, \mathcal{A})$.
 2. The cost of collecting the samples \mathcal{D}_n , noted $c_2(n)$.

For most “reasonable” algorithms, one should expect

- $\inf_{\mathcal{A}} a(n, \mathcal{A}) > 0$ for all n (one cannot solve a noisy problem exactly with finitely many samples), and $\lim_{n \rightarrow \infty} a(n, \mathcal{A}) = 0$, and that, in first approximation, $a(n, \mathcal{A})$ strictly decreases as n increases, and
- $\lim_{n \rightarrow \infty} c_1(n, \mathcal{A}) = \lim_{n \rightarrow \infty} c_2(n) = +\infty$ for all \mathcal{A} .

In such cases, one cannot minimize both computational cost and accuracy at the same time. A more reasonable quest is, given a target level of accuracy $\varepsilon > 0$, to find among algorithms with accuracy at least ε , the one with minimum cost. Let $n_{\varepsilon, \mathcal{A}}$ be the minimum number of samples required for \mathcal{A} to achieve an ε -accuracy. $n_{\varepsilon, \mathcal{A}}$ is known as the *sample complexity* [255, 25]. The cost of ε -accurate estimation is then given by:

$$c(n_{\varepsilon, \mathcal{A}}, \mathcal{A}) = c_1(n_{\varepsilon, \mathcal{A}}, \mathcal{A}) + c_2(n_{\varepsilon, \mathcal{A}}) \quad (\text{II.5})$$

From Equation II.5, we see that the cost of ε -accurate estimation is strongly dependent on the quantity $n_{\varepsilon, \mathcal{A}}$: in particular, a smaller $n_{\varepsilon, \mathcal{A}}$ implies a smaller sample collection cost c_2 , and possibly, (depending on the algorithm) a smaller computational cost c_1 . As a (theory-minded) first step towards this goal, one can seek, when designing a new algorithm, or analyzing existing ones, to obtain expressions, or at least satisfying approximations of their sample complexity.

II.2.4 Statistical efficiency of density estimators

II.2.4.1 From sample complexity to convergence bounds

As an equivalent quantity to $n_{\varepsilon, \mathcal{A}}$, one could instead look at the $\varepsilon_{n, \mathcal{A}}$, the accuracy achieved by \mathcal{A} on n samples. While different accuracy measures exist, a large body of statistics literature has chosen to focus on characterizing the distribution of $\psi_n - \psi^*$ the difference between the estimated parameter and the best one. We briefly discuss its pros and cons:

- On the one hand, this quantity is fundamental, in the sense that it determines the behavior of p_{ψ_n} , and thus of any divergence $d(p_{\psi_n}, p_{X|\Theta})$ of the resulting model to $p_{X|\Theta}$. In particular, from the distribution of $\psi_n - \psi^*$, one directly derive bounds for $\|\psi_n - \psi^*\|^2$, the mean squared error of the estimate ψ_n to the true parameter.
- On the other hand, for this quantity to make sense, one needs to assume uniqueness [256, 73, 229] of the best parameter ψ^* , which does not hold when using modern model classes such as neural networks.

Notwithstanding the limitations of this quantity, many analyses of modern Machine Learning techniques [92, 136, 151, 37] still focus on the distribution of $\psi_n - \psi^*$ for simpler models where uniqueness, as the results can still provide useful insights into the behavior of more complex models trained in the same manner. In this thesis, we will follow this approach, through theoretical analysis of uniquely well specified models. Ultimately, I believe that refinements of this analysis should be used to handle non-unique minimizers.

II.2.4.2 Rate of convergence of M-estimators, asymptotic normality

In Theorem II.2.1, we have shown that under appropriate conditions, M-estimators were consistent. Under additional regularity conditions, it is possible to derive their speed of convergence.

Theorem II.2.2 ([256, Theorem 5.34]). *Assume that*

- d_n is of the form of Equation II.3

- the conditions of Theorem II.2.1
- d_n takes the form of Equation II.3, where $\ell(X, \theta; \psi)$ is thrice continuously differentiable w.r.t ψ , and such that $\mathbb{E}[\ell(X_i, \theta_i; \psi)]$ can be differentiated w.r.t ψ thrice under the integral sign, and that $H_\psi d$ is non-singular.

Then it holds that

$$\sqrt{n}(\psi_n - \psi^*) \xrightarrow{d} \mathcal{N}(0, H^{-1}GH^{-1})$$

where $H := \mathbb{E}[H_\psi \ell(X_i, \theta_i; \psi)^{-1} |_{\psi=\psi^*}]$ and $G = \text{Cov}[\nabla_\psi \ell(X_i, \theta_i; \psi) |_{\psi=\psi^*}]$.

Asymptotic Variance of MLE In the special case of MLE, and assuming well-specification, things happen to cancel out nicely, as, noting $p_\psi^\mu(x|\theta) := \frac{dp_\psi}{d\mu}(x|\theta)$ for brevity,

$$\begin{aligned} H_{\text{MLE}} &:= \iint H_\psi \log p_{\psi^*}^\mu(x|\theta) p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) \\ &= \iint \left(\frac{H_\psi p_{\psi^*}^\mu(x|\theta)}{p_{\psi^*}^\mu(x|\theta)} (x|\theta) - \frac{\nabla_\psi p_{\psi^*}^\mu(\nabla_\psi p_{\psi^*}^\mu)^\top(x|\theta)}{p_{\psi^*}^\mu(x|\theta)^2} \right) p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) \\ &= \int H_\psi p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) \\ &\quad - \int \nabla_\psi \log p_{\psi^*}^\mu(x|\theta) \nabla_\psi \log p_{\psi^*}^\mu(x|\theta)^\top p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) \\ &= G_{\text{MLE}} \end{aligned}$$

where the final equality follows from the fact that

$$\int H_\psi p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) = H_\psi \int p_{\psi^*}^\mu(x|\theta) \mu(dx) \pi(d\theta) = H_\psi 1 = 0$$

Noting $\mathcal{I}(\psi^*) = \mathbb{E}[(\nabla_\psi \log p_\psi^\mu(\nabla_\psi \log p_\psi^\mu)^\top)(X_i|\theta_i)] \Big|_{\psi=\psi^*} := G_{\text{MLE}}$, also called the *Fisher Information Matrix* of the data at ψ^* , we thus have:

$$\sqrt{n}(\psi_{n,\text{MLE}} - \psi^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\psi^*)^{-1})$$

II.2.5 Limits on the efficiency of statistical estimators

Above, we have showed that M-estimators converge at a $1/\sqrt{n}$ rate to the best parameter. Moreover, we obtained an expression for the multiplicative term of that rate, e.g. the variance of the asymptotic distribution of $\sqrt{n}(\psi_n - \psi^*)$. Consequently, all M-estimators converge at the same rate. Additionally, this results implies that asymptotically, the best estimator among some fixed subset is the one with the smaller asymptotic variance.

In the quest to find efficient estimators, a natural question becomes “Does there exist a “best” estimator, e.g. an estimator with an asymptotic variance lower than *any other*?”. As discussed in the previous section, statistical estimation is a noisy problem: one should not expect to estimate the best parameter ψ^* exactly using a finite amount of samples. Even more strongly, in most scenarios, there should exist a quantity (dependent on n and p^*) lower-bounding the efficiency of *all* “reasonable” statistical estimators \mathcal{A} .

II.2.5.1 Efficiency of Unbiased Estimators

The Cramér-Rao inequality constitutes a fundamental result in that direction. It was first conjectured by Fisher [68, 69], and independently established by Cramér [45] and Rao [205]. It provides a lower bound on the finite-sample variance of any estimator and suggest that MLE’s is close to being an “optimal” estimator. This lower bound is guaranteed to be non-vacuous when the estimator is unbiased—i.e. verifies $\mathbb{E}\psi_n = \psi^*$. The theory of biased estimators is more subtle, and we defer its discussion to the next section. We restate the result, which is influential to this thesis, below. In the following, we formalize the dependence of ψ_n on the dataset by writing $\psi_n := \psi_n(Z_1, \dots, Z_n)$, where $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} p_{\psi^*}$. For two positive definite matrices A, B , we write $A \succeq B$ (resp. $A \preceq B$) if $A - B$ is positive semi-definite (resp. negative semi-definite).

Theorem II.2.3 (Cramér-Rao Inequality [45, 205, 271, 118]). *Assume that the following conditions holds:*

- $p_\psi^\mu(z)$ is differentiable w.r.t ψ for all $\psi \in \text{int}(\Psi)$ and almost all z .

- $\mathbb{E} \left(\frac{\partial \log p_{\psi^*}^\mu(Z_i)}{\partial \psi_i} \right)^2 < +\infty$ for $i \in 1, \dots, p$
- The function $\mathbb{E} \psi_n(Z_1, \dots, Z_n)$ is differentiable w.r.t ψ under the expectation sign.
- $\mathbb{E} [\psi_n] = \psi^*$
- The matrix $\mathcal{I}(\psi) := \mathbb{E} [\nabla_\psi \log p_\psi^\mu(Z_1) \nabla_\psi \log p_\psi^\mu(Z_1)^\top]$ is non-singular for all $\psi \in \Psi$.

Then it holds that

$$\mathbb{E} [\|\psi_n - \psi^*\|^2] \succeq \frac{\mathcal{I}(\psi^*)^{-1}}{n} \quad (\text{II.6})$$

Consequently, if additionally, $\sqrt{n}(\psi_n - \psi^*) \xrightarrow{d} U$ for some random variable U , then we have $\text{Cov } U \succeq \frac{\mathcal{I}(\psi^*)^{-1}}{n}$.

This result sheds some light on multiple phenomena:

- First, it shows that the $1/\sqrt{n}$ rate of convergence of M-estimators is optimal in the sense that no unbiased estimator can converge faster than that rate.
- Second, the prefactor of lower bound is the inverse Fisher Information Matrix at ψ^* , $\mathcal{I}(\psi^*)^{-1}$, is precisely asymptotic variance of the MLE!

Importantly, nothing guarantees that the *finite sample* variance of MLE matches the Cramér-Rao lower bound. In fact, looking at proofs of Theorem II.2.3, we see that the conditions imposed on the estimator to make the Cramér-Rao lower bound tight are stringent, and may not be verified by practical estimators, whose variance is likely to contain additional terms. However, this result shows that the asymptotic variance of MLE is the best attainable one among unbiased estimators admitting a limit distribution. From this perspective, MLE constitutes a serious candidate when aiming to build low-cost, accurate estimators.

II.2.5.2 The efficiency of arbitrary estimators

The analysis performed above contains is restricted to unbiased estimators converging to some limiting distribution in the large sample limit. Similar conclusions follow

for the case of arbitrary estimators. However, the mathematics behind it become a lot more subtle. While not core to this thesis, we provide a brief summary of these results; we refer the interested reader to [261, Section 2] for an informal review of the history of the field, and to [256] for a more formal treatment of the topic. First, while the Cramér-Rao inequality admits analogues applying to biased estimators, the resulting lower bound becomes estimator-dependent, and is not even guaranteed to be strictly positive. The reason is that one can construct learning algorithms that are “excellent” at a given point, but terrible for most other points. As an example, take for constant estimator $\psi_n := \psi_0$. This estimator achieves exactly 0 error at $\psi^* = \psi_0$ regardless of the number of samples. It has 0 variance at all other points, but a non-zero MSE $\|\psi_0 - \psi^*\|^2$ that does not vanish as n increases (it is called inconsistent). It thus presents an accuracy ceiling that cannot be overcome by increasing the number of samples. The estimator above is quite crude: it is perfect at a point, but inconsistent at any other. A more subtle instance was constructed by Hodges [107]. His estimator, while biased, is consistent at all points, and “superefficient” (e.g. has a smaller asymptotic variance than MLE) at the point $\psi^* = 0$. Nevertheless, the asymptotic optimality of MLE remains largely unaffected, owing to two important results:

- For biased estimators admitting a limiting distribution, it was shown [1, 145, 256], superefficiency is extremely rare: it can only be achieved on a Lebesgue null set.
- For general estimators, it was shown that at each point, the worst-case asymptotic discrepancy of an estimator in a vanishing neighborhood of each point is worse than the one of MLE [95, 256].

II.2.5.3 Conclusion and limitations: From statistical complexity to computational complexity

The theoretical results stated above rigorously established that Maximum Likelihood Estimation is asymptotically the most statistically efficient estimator. It suggests that estimators deviating from the MLE may have failure modes that should be

investigated, and that, from a statistical perspective, “MLE-inspired” estimators may be desirable. However, the analysis performed omits several computational considerations that can play an important role in practice:

- First, the analysis is asymptotic, and does not provide any finite-sample optimality guarantees for MLE, or any other estimator. Much more recently, several works managed to provide finite-sample guarantees for certain classes of estimators [229, 185]. Proving them required almost complete overhaul of proof-strategies compared to their asymptotic counterparts [229]. Moreover, the asymptotic-to-finite sample translation program is far from complete: for instance, these works focused for the most part on obtaining error upper-bounds, and left open optimality questions as the ones addressed above. We expect that translating optimality results from the asymptotic regime to the finite-sample one will be highly non-trivial.
- Second, the convergence guarantees of M-estimators assume that minimization can be performed exactly. This assumption can be violated in practice, to varying degrees. First, exact optimization in finite time rarely occurs for standard optimization algorithms such as stochastic gradient descent, in particular for non-convex losses. Second, for certain class of models, such algorithms cannot even be used as is, due to the presence of intractable quantities preventing an unbiased estimation of the gradient [105].

Overall, a complete and satisfying analysis of density estimators should be *computational*: it should (i) provide finite-sample statements (ii) account for optimization errors, and (iii) provide an end-to-end cost analysis of the algorithm. Nonetheless, obtaining generic results of this type may still be decades away.

II.3 Comparing Distributions with Kernels

In this section, we provide a brief overview of a paradigm used for comparing distributions from samples, based on Reproducing Kernel Hilbert Spaces, which is heavily relied upon in the second part of this thesis. As we will show then, this

paradigm allows designing *absolute* measures of model performance and reliability which can be estimated from samples, and statistical tests to evaluate whether a model fits its training data.

II.3.1 Motivation: constructing estimable distances between distributions

Let $(\mathcal{Z}, \mathcal{L})$ be some measurable space, and consider two probability distributions μ and ν on $(\mathcal{Z}, \mathcal{L})$. We aim to construct a notion of distance between them. Moreover, we want this distance to admit estimators given samples from μ and ν : indeed, our motivation from creating such distances is to use it compare models $p_{\psi_n}^\mu$ from the target distribution p_Z , the latter being only accessible through its training data. Consequently, the KL divergence discussed above (as well as any f -divergence [4]) cannot be naively used, as estimating them requires knowledge of dp_Z/dp_{ψ_n} .

To construct such a distance, let us start from the simple realization that:

$$\begin{aligned}\mu = \nu &\iff \int f d\mu = \int f d\nu \quad \forall f \in \mathbb{R}^{\mathcal{Z}} \text{ measurable} \\ &\iff \int f d\mu - \int f d\nu = 0, \quad \forall f \in \mathbb{R}^{\mathcal{Z}} \text{ measurable}\end{aligned}$$

The converse implication directly follows by taking $f(\bullet) = \mathbf{1}_A(\bullet)$ for any set $A \in \mathcal{L}$. This equality can be used to construct a first notion of distance between μ and ν , given by trying to maximize the difference between the two integrals over all measurable functions f :

$$d(\mu, \nu) := \sup_{f \in \mathbb{R}^{\mathcal{Z}} \text{ measurable}} \left\{ \int f d\mu - \int f d\nu \right\} \in \mathbb{R}_+ \cup \{+\infty\} := \overline{\mathbb{R}}$$

For which, by construction $d(\mu, \nu) = 0 \iff \mu = \nu$, and one can check that $d(\mu, \nu) = d(\nu, \mu) \geq 0$ for all μ, ν . The benefit of this “proto-distance” is that it only involves μ and ν through their integrals: it thus admits a natural estimator

given samples $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} \mu$ and $Z'_1, \dots, Z'_n \stackrel{\text{i.i.d.}}{\sim} \nu$:

$$\widehat{d}(\mu, \nu) := \sup_{f \in \mathbb{R}^Z \text{ measurable}} \left\{ \frac{1}{n} \sum f(Z_i) - \frac{1}{n} \sum_{i=1}^n f(Z'_i) \right\}$$

On the other hand, we actually have $d(\mu, \nu) = +\infty$ for all $\mu \neq \nu$: consequently, this distance cannot be used to assess the relative closeness of say two measures μ_1 and μ_2 to some target measure ν , as both will be flagged as equally far away from it.

From a technical perspective, the reason for the latter limitations is that the space of measurable functions is very large, allowing the supremum to be very large. To overcome this limitation one can modify the definition of d by restricting its search set to some subset of measurable functions $\mathcal{F} \subset \mathbb{R}^Z$:

$$d_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left\{ \int f d\mu - \int f d\nu \right\} \in \mathbb{R}_+ \cup \{+\infty\} := \overline{\mathbb{R}} \quad (\text{II.7})$$

with associated estimator

$$\widehat{d}_{\mathcal{F}}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum f(Z_i) - \frac{1}{n} \sum_{i=1}^n f(Z'_i) \right\} \quad (\text{II.8})$$

\mathcal{F} should be chosen (i) small enough to ensure that $d(\mu, \nu) < +\infty$ in most cases, (ii) large enough to distinguish any two measures, and (iii) the estimator $\widehat{d}_{\mathcal{F}}(\mu, \nu)$ can be computed, e.g. that the supremum of Equation II.8 can be found. The class of distances introduced in Equation II.7 is known as *Integral Probability Metrics* (IPM) [172]. It gives rise to many interesting distances studied in Probability and Statistics, sometimes already known under different formulas.

Ensuring simultaneously (ii) and (iii) looks like a complex challenge. Most well-known optimization algorithms apply to finite-dimensional problems, while, on the other hand, finite-dimensional function spaces seem not large enough to distinguish all measures.

II.3.2 Reproducing Kernel Hilbert Spaces The Maximum Mean Discrepancy

To construct a distance satisfying all three points, we will borrow powerful function spaces discovered by functional analysts and heavily applied in statistics, called Reproducing Kernel Hilbert Spaces (RKHS). To arrive at it, we will revisit the famous “kernel trick” applied to the problem of comparing distributions. After that, we move on to define RKHS, their associated pseudo-distances called Maximum Mean Discrepancies (MMD), and conditions to ensure they are distances.

II.3.2.1 The kernel trick

When $\mathcal{Z} = \mathbb{R}^d$, perhaps the most non-trivial computationally friendly set of functions is the Euclidean space of linear functions from \mathcal{Z} to \mathbb{R} , $\mathcal{L}(\mathcal{Z}, \mathbb{R}) = \{z \mapsto \langle a, z \rangle, a \in \mathbb{R}^d\}$, endowed with the inner product $\langle a, a' \rangle$. In that setting, the optimization problem of d and \hat{d} becomes *homogenous*: to distinguish between two measures μ and ν with this space, one can restrict the search to the unit ball. By the Cauchy-Schwarz inequality, this implies that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n f(Z_i) - \frac{1}{n} \sum_{i=1}^n f(Z'_i) &= a^\top \left(\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n Z'_i \right) \\ &\leq \|a\| \left\| \frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n Z'_i \right\| \end{aligned}$$

with equality if and only for

$$a^* = \frac{\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n Z'_i}{\left\| \frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n Z'_i \right\|}$$

from which we have

$$\begin{aligned} \hat{d}_{\mathcal{L}(\mathcal{Z}, \mathbb{R})}(\mu, \nu) &= \left\| \frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n Z'_i \right\| \\ &= \left(\sum_{i,j=1}^n \langle Z_i, Z_j \rangle + \sum_{i,j=1}^n \langle Z'_i, Z'_j \rangle - 2 \sum_{i,j=1}^n \langle Z_i, Z'_j \rangle \right)^{1/2} \end{aligned}$$

On the positive side, this estimator admits a closed-form expression, fulfilling condition (iii) above. On the negative side, it only applies to subsets of \mathbb{R}^d , and is only able to distinguish measures with different means. To address the two limitations simultaneously, one could apply the same procedure on a transformed version of the data $\{\phi(Z_i)\}_{i=1}^n, \{\phi(Z'_i)\}_{i=1}^n$, where $\phi : \mathcal{Z} \rightarrow \mathbb{R}^{d'}$. That procedure yields the estimator

$$\begin{aligned} & \widehat{d}_{\mathcal{L}(\mathbb{R}^{d'}, \mathbb{R}) \circ \phi}(\mu, \nu) \\ &= \left(\sum_{i,j=1}^n \langle \phi(Z_i), \phi(Z_j) \rangle + \sum_{i,j=1}^n \langle \phi(Z'_i), \phi(Z'_j) \rangle - 2 \sum_{i,j=1}^n \langle \phi(Z_i), \phi(Z'_j) \rangle \right)^{1/2} \end{aligned} \quad (\text{II.9})$$

Setting $\phi(z) := (z_1, \dots, z_d, z_1^2, z_1 z_2, \dots, z_d^2)$, the resulting distance will now distinguish measures with different means and covariances. However, the resulting distance is still not able to distinguish measures with different higher-order moments. Ultimately, we may need to consider a feature map ϕ taking values in an *infinite-dimensional* vector space admitting an inner product structure, e.g. a Hilbert space \mathcal{H} .

Importantly, the only thing needed from ϕ to evaluate Equation II.9 is the inner product $\langle \phi(Z_i), \phi(Z_j) \rangle$. The successful bet of the kernel trick is to approach the problem from the opposite direction: could we construct functions $k(z, z')$ that implicitly define an inner product $k(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}_0}$ for some unknown, but possibly infinite dimensional Hilbert space \mathcal{H}_0 and feature map $\phi : \mathcal{Z} \rightarrow \mathcal{H}_0$?

II.3.2.2 Reproducing Kernel Hilbert Spaces

The positivity property of the inner product suggests the following definition and terminology such k :

Definition II.3.1 (kernel [233, Definition 4.15], [20, Definition 2]). Let \mathcal{Z} be some set. A function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ is called a positive definite function (or *kernel*) if, for all $n \in \mathbb{N}$, $z_1, \dots, z_n \in \mathcal{Z}$, and $a_1, \dots, a_n \in \mathbb{R}$, it holds that

$$\sum_{i,j=1}^n a_i a_j k(z_i, z_j) \geq 0$$

It turns out [233, Theorem 4.16] that indeed, for any kernel, there exists a Hilbert space \mathcal{H}_0 and feature map $\phi : \mathcal{Z} \rightarrow \mathcal{H}_0$ such that $k(z, z') = \langle \phi(z), \phi(z') \rangle$. The pair (ϕ, \mathcal{H}_0) is not unique (one can, for instance, add zeros to the feature map and resulting Hilbert space). However, the feature map is only the mean to an end of creating a flexible Hilbert space \mathcal{H} of functions verifying:

$$f(z) = \langle w, \phi(z) \rangle, \quad \forall z \in \mathcal{Z}, w \in \mathcal{H}_0$$

Note that (i) by setting $w = \phi(z)$ for some $z \in \mathcal{Z}$, we have that the function k_z defined as $k_z(z') = k(z, z')$ satisfies $k_z \in \mathcal{H}$, and (ii) this space is independent of the choice \mathcal{H}_0 and ϕ associated with k (as the l.h.s does not depend on it). Combining this fact with an application of the Riesz representation theorem (see [233, p.119-121]) makes the pair $(\mathcal{H}, z \mapsto k_z)$ a canonical choice of feature map and Hilbert space associated with k , allowing to define \mathcal{H} in a self-contained manner, and resulting in the following result:

Theorem II.3.2 ([233, Theorem 4.21] [14, p. 347]). *Let k be a kernel. Then there exists a unique Hilbert space \mathcal{H} , called Reproducing Kernel Hilbert Space (RKHS) such that*

- $k_z \in \mathcal{H}$ for all $z \in \mathcal{Z}$
- for all $f \in \mathcal{H}$, it holds that $f(z) = \langle f, k_z \rangle_{\mathcal{H}}$ for all $z \in \mathcal{Z}$

Moreover, each RKHS has a unique kernel k verifying the properties above.

II.3.2.3 Maximum Mean Discrepancies

We are now ready to define a flexible family of distance between distributions [90]:

Definition II.3.3 (Maximum Mean Discrepancy; MMD, Definition 2 of Gretton et al. [90]). Let \mathcal{H} be a RKHS with kernel k . The Maximum Mean Discrepancy $\text{MMD}(\mu, \nu)$ between two distribution $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ is defined as:

$$\text{MMD}(\mu, \nu) := d_{\mathcal{B}_{\mathcal{H}}(0_{\mathcal{H}}, 1)}(\mu, \nu) = \sup_{f \in \mathcal{B}_{\mathcal{H}}(0_{\mathcal{H}}, 1)} \left\{ \int f d\mu - \int f d\nu \right\} \quad (\text{II.10})$$

By using the properties of RKHS and repeating steps of our derivations from Section II.3.2.1, a natural estimator candidate for $\text{MMD}(\mu, \nu)$ Given samples $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} \mu$ and $Z'_1, \dots, Z'_n \stackrel{\text{i.i.d.}}{\sim} \nu$ is then given by

$$\begin{aligned}\widehat{d}_{\mathcal{L}(\mathcal{Z}, \mathbb{R})}(\mu, \nu) &= \left\| \frac{1}{n} \sum_{i=1}^n k_{Z_i} - \frac{1}{n} \sum_{i=1}^n k_{Z'_i} \right\|_{\mathcal{H}} \\ &= \frac{1}{n} \left(\sum_{i,j=1}^n k(Z_i, Z_j) + \sum_{i,j=1}^n k(Z'_i, Z'_j) - 2 \sum_{i,j=1}^n k(Z_i, Z'_j) \right)^{1/2}\end{aligned}$$

This estimator will be consistent when $\mathbb{E} \sqrt{k(Z_1, Z_1)}, \mathbb{E} \sqrt{k(Z'_1, Z'_1)} < +\infty$; see e.g. Theorem 8 of [90].

Universality So far, we have come up with a family of distances between distribution that is easy to estimate from samples. For the kernel to yield a distance between distributions, we have posited that its associated feature map needed to be infinite-dimensional. The full story is a bit more subtle; below, we provide a sufficient condition on the kernel to yield a distance.

Definition II.3.4 (Section 2 of 232). A kernel k is said to be

- *universal* if \mathcal{Z} is a compact Haussdorf space and \mathcal{H} is dense in $\mathcal{C}(Z)$, the set of continuous functions on \mathcal{Z} endowed with the supremum norm.
- *c_0 -universal* if \mathcal{Z} is a locally compact Haussdorf space, $k_z \in \mathcal{C}_0(\mathcal{Z})$ for all z (e.g. k is “ c_0 ”), and \mathcal{H} is dense in $\mathcal{C}_0(Z)$, the set of continuous functions on \mathcal{Z} vanishing at infinity endowed with the supremum norm.

These properties will be used in the second part of this thesis to construct absolute measures of accuracy and reliability of conditional density estimators.

II.3.3 Vector-Valued RKHS

We finish this section with a brief primer on vector-valued RKHS, which are used in the second part of this thesis.

From RKHS to vector-valued RKHS In certain cases, it may be convenient to define spaces similar to RKHS, but returning values belonging to a Hilbert space \mathcal{U} . The reproducing property of RKHS

$$f(z) = \langle f, k_z \rangle \quad (\text{II.11})$$

ontologically breaks, since now, we have $f(z) \in \mathcal{U}$ while $\langle f, k_z \rangle \in \mathbb{R}$. It is however possible to re-express Equation II.11 by resorting to *adjoints*, a more general construct than inner products. Indeed, for $f \in \mathcal{H}$, we have

$$\langle f, k_z \rangle = k_z^* f \quad (\text{II.12})$$

where k_z^* is the adjoint of k_z , a linear form from \mathcal{H} to \mathbb{R} . Generalizing Equation II.12 to \mathcal{U} -valued functions f implies that the kernel K (written in uppercase to distinguish it from the scalar-valued kernels) should now be such that $K_z^* \in \mathcal{L}(\mathcal{H}, \mathcal{U})$, and thus $K_z \in \mathcal{L}(\mathcal{U}, \mathcal{H})$. With that in mind, we are ready to propose the following definitions:

Definition II.3.5 (Operator-valued kernel [33, Section 2.2]). Let \mathcal{Z} be a set, \mathcal{U} be a separable Hilbert space. A function $K : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{L}(U)$ is called a (\mathcal{U} -)operator-valued kernel if, for all $z_1, \dots, z_n \in \mathcal{Z}, u_1, \dots, u_n \in \mathcal{U}$

$$\sum_{i,j=1}^n \langle u_i, K(z_i, z_j) u_j \rangle_{\mathcal{U}} \geq 0$$

Definition II.3.6 (Vector-Valued RKHS [33, Section 2.2]). Let \mathcal{Z} be a set, \mathcal{U} be a separable Hilbert space, and K be a \mathcal{U} -operator-valued kernel. Define $K_z : \mathcal{U} \rightarrow \mathcal{F}(\mathcal{Z}, \mathcal{U})$ such that $(K_z u)(z') = K(z, z') u$. A Hilbert space $\mathcal{H} \subseteq \mathcal{F}(\mathcal{Z}, \mathcal{U})$ is called a vector-valued RKHS with kernel K if:

- for all $z \in \mathcal{Z}, K_z \in \mathcal{L}(\mathcal{U}, \mathcal{H})$
- for all $f \in \mathcal{H}$, we have $f(z) = K_z^* f$

Universality of vector-valued RKHS The universality properties of scalar-valued kernels can be extended almost verbatim to vector-valued RKHS. In particular, we

say that, for a second-countable locally compact Hausdorff space \mathcal{Z} , a \mathcal{U} -operator-valued kernel K is c_0 -universal [33, Theorem 1] if $K_z \in \mathcal{C}_0(\mathcal{Z}, \mathcal{U})$ for all $z \in \mathcal{Z}$ (e.g. K is “ c_0 ”) and \mathcal{H} is dense in $\mathcal{C}_0(\mathcal{Z}, \mathcal{U})$, the set of continuous functions on \mathcal{Z} vanishing at infinity endowed with the supremum norm.

Part I

Contributions to the training of conditional density models

CHAPTER III

A Statistical Analysis for NRE

This Chapter is based on the following work:

Pierre Glaser and Arthur Gretton. Statistical Analysis of Neural Ratio Estimation.
In preparation, 2025

Abstract

Neural Ratio Estimation (NRE) is a popular method for performing simulation-based posterior estimation given a known prior and samples from a joint distribution. For such a task, it can be crucial for the posterior estimation task to be *statistically efficient*, e.g. to require as few simulations as possible to obtain an accurate posterior estimate. However, currently, little is known about its statistical efficiency. In this work, we bridge this gap by performing an asymptotic statistical analysis of NRE. We analyze two variants of NRE: the standard one, whose loss contains a quadratic number of terms, and a more computationally friendly variant obtained by subsampling the negative examples. We show that these two variants are consistent, asymptotically normal, and we provide an expression for their asymptotic variance. Next, we provide examples of *hard distributions* for NRE, e.g. joint distributions for which the asymptotic variance of NRE scales exponentially with dimension. This shows that NRE can be significantly less efficient than other posterior estimation methods like Maximum Likelihood Estimation, the *theoretically asymptotically optimal* posterior estimator. In doing so, we provide a rigorous framework for statistically analyzing contrastive methods—including NCE, improving on existing results.

III.1 Introduction and Motivation

Statistical Problem: Conditional Density Estimation Say we observe some samples $\{X_i, \theta_i\}_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \pi(\theta \in \bullet) p_{\mathcal{X}|\Theta}(x \in \cdot | \theta = \diamond) := p_{\mathcal{X},\Theta}$ where π is known, and we wish to learn a model of the conditional distribution $p_{\mathcal{X}|\Theta}$.

Motivation: Simulation-Based Inference The setup above is central to Simulation-Based Inference (SBI, 47). In SBI, one wishes to infer the (unknown) parameter θ_0 of a stochastic scientific model (such as an SDE) given some observed variable X_0 . If the conditional distribution $p_{\mathcal{X}|\Theta}$ induced by the model were known, one could, given a prior π on θ , perform (Bayesian) inference such as MCMC on the posterior using Bayes rule:

$$p_{\Theta|\mathcal{X}}(\theta \in \bullet | x) = \pi(\theta \in \bullet) \frac{dp_{\mathcal{X}|\Theta}}{dp_{\mathcal{X}}}(x, \theta). \quad (\text{III.1})$$

However, scientific models are often too complex for $p(X|\theta)$ to be known analytically. If, nonetheless, one can sample from the model at any parameter value θ (e.g. draw some $X|\theta \sim p_{\mathcal{X}|\Theta}(x \in \cdot | \theta)$), one can use these samples to learn a density model of the posterior, on which inference could then be performed, hence the relevance of statistical problem discussed above.

Conditional Density Estimation Method of interest: Neural Ratio Estimation

There are many ways to estimate the conditional distribution $p_{\Theta|\mathcal{X}}$ [190, 101, 88, 188] from samples $\{X_i, \theta_i\}_{i=1}^N$. Neural Ratio Estimation (NRE, 101), the approach of interest in this work, which crucially assumes that π is known, is a popular posterior estimation method in SBI which consists in learning a model of the (log-)ratio:

$$g_{\psi_{\text{NRE},N}}(x, \theta) \approx \log \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta).$$

Here, $p_{\mathcal{X}}$ is the marginal distribution of X , $p_{\mathcal{X}} \otimes \pi$ is the product of the X and θ marginals, and g_\bullet is some parametric ratio model with parameter space $\psi \in \Psi \subseteq \mathbb{R}^d$. From this learned ratio, and further assuming that $p_{\mathcal{X}} \ll c$ for some measure c (e.g.

the Lebesgue measure) on \mathcal{X} , NRE returns a posterior estimate is given by

$$p_{\psi_{NRE,N}}(\theta \in \bullet | x) := \exp(g_{\psi_{NRE,N}}(x, \theta)) \times \pi(\theta \in \bullet). \quad (\text{III.2})$$

Note that if $g_\psi = \log \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta)$, then by Equation III.1 we have $p_\psi = p_{\Theta|\mathcal{X}}$. To learn g_ψ , one can train it to discriminate between samples from $p_{\mathcal{X},\Theta}$ and $p_{\mathcal{X}} \otimes \pi$ via logistic regression: indeed, samples $\{X_i, \theta_i\}_{i=1}^N$ from $p(x, \theta)$ are available by assumption, while to obtain samples from $p_{\mathcal{X}} \otimes \pi$, it suffices to create pairs $\{X_i, \theta_j\}_{1 \leq i \neq j \leq N}$, as (X_i, θ_i) is independent of (X_j, θ_j) for $i \neq j$.

Purpose of this work: Analyzing the statistical properties of NRE In this work, we are interested in quantifying the statistical performance of NRE estimators. In particular, assuming that there exists a ψ^* such that $g_{\psi^*} = \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \otimes \pi)}$, we wish to establish properties of (squared) error $\|\psi_{NRE,N} - \psi^*\|^2$ of $\psi_{NRE,N}$ to ψ^* , which is a random quantity. Characterizing properties of some parametric estimator ψ_N is a common in the machine learning and statistics literature. Desirable properties of interest include

1. (consistency) $\psi_N \rightarrow \psi^*$ (in probability, or almost surely)
2. (convergence rate) $\|\psi_N - \psi^*\|^2 = O_p(N^{-1})$, and
3. (asymptotic normality) $\sqrt{N}(\psi_N - \psi^*) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ .

The first two points, describing (1) the consistency, and (2) the convergence rate of ψ_N to ψ^* as $N \rightarrow \infty$, are—once established—statements independent of estimation technique. These two points alone are usually not enough to compare parametric estimators, as many of them are both consistent, and converge at the $O_p(N^{-1})$ rate, the so-called *parametric rate*. The third point quantifies the multiplicative factor in front of the convergence rate constitutes, in a sense, the finest problem-dependent quantifier of the performance of an estimator. Contrary to the first two points, computing the asymptotic variances Σ_1 and Σ_2 of two different estimators ψ_N^1 and ψ_N^2 , offers a way to compare their performance: namely, a smaller asymptotic variance means achieving, for values of N such that asymptotic regime holds accurately,

a smaller mean squared error. It is also known that the best possible asymptotic variance of parametric estimators is given by the Cramér-Rao Lower Bound (CRLB, 256) and that this variance is achieved by the Maximum Likelihood Estimator (MLE) [256, 98]. Over the last decades, these ingredients have been leveraged to describe the performance of many computationally attractive alternatives to MLE, such as Score Matching [117, 136] and Noise Contrastive Estimation [92, 163, 151, 37].

Contributions In this work, we perform an asymptotic statistical analysis of NRE. We analyze two variants of NRE: the standard one as introduced in [101], whose loss contains a quadratic number of terms, and a more computationally friendly variant obtained by subsampling the negative examples, which we call Incomplete NRE (INRE). We establish, under specific assumptions, their consistency and asymptotic normality, and provide an expression for their asymptotic variance. Additionally, we provide examples of *hard distributions* for NRE and INRE, e.g. distributions for which the (asymptotic) mean-squared error of the NRE estimator scales exponentially with dimension. This shows that NRE can be significantly less efficient than other posterior estimation methods like Conditional Maximum Likelihood Estimation. Our statistical analysis allows to analyze both NRE and the more well-known Noise Contrastive Estimation (NCE, 92), a density estimation method for unnormalized models, recovering rigorously certain results already claimed in the literature.

III.2 Related Work

Existing analyses of NRE Closest to our work is the one of Ma and Collins [163], which analyzed the properties of NRE and NPE style estimators (assuming an empirically estimated prior) in the context of Natural Language Processing. Their goal was to *upper bound* their asymptotic variances, while our goal is to provide lower bounds for them. Aside from technical limitations of their results (discrete sample spaces, compact parameter space, bounded score functions), their analysis is asymptotic, and restricted to an estimator minimizing a loss equal in expectation to ours, but containing only n terms instead n^2 present in ours. This difference in number of terms yields estimators with drastically different asymptotic variances, as

shown in our work. Finally, the results of 163 contain technical issues: in particular, their appeal to the standard variant of the Central Limit Theorem [163, Equation 21] is not justified, as the loss that they consider is not an average of i.i.d terms, and thus must account for dependence across terms.

Sample complexity of Logistic Regression An important component of NRE is the density ratio estimation step, performed using Logistic Regression. This problem has a long history in the machine learning literature; two accuracy measures are typically considered: (1) the distance of the estimated parameter ψ_N to the true parameter ψ^* , of interest in our work, and (2) the excess classification risk, which is the difference between the classification risk of the estimated classifier and that of the optimal classifier, traditionally studied using the theory of Empirical Risk Minimization (ERM, 27, 258, 277). The latter is relevant when classification is the end goal of the learning procedure; however, in NRE and NCE, classification is only a surrogate task to the original goal of density estimation, making the ERM analyses not useful to quantify the performance of NRE and NCE. Results regarding the first measure (distance to the true parameter) were obtained originally treated using either the asymptotic theory of M -estimators [256]. More recently, finite sample bounds in a variety of settings were obtained; see, e.g. [112, 185, 64, 36, 139]. Analyses [139, 36] typically break down the distance to the true parameter into two sub-quantities (i) the difference between the direction of the estimated parameter and that of the true parameter, $\left\| \frac{\psi^*}{\|\psi^*\|} - \frac{\psi_N}{\|\psi_N\|} \right\|$, and (ii) the difference between the norms of the estimated and true parameters, $\|\psi^*\| - \|\psi_N\|$. When the two distributions to classify are almost linearly separated, parameter direction can be estimated accurately, but the parameter norm is much harder to estimate [139], with constants in the bounds depending exponentially on parameters of the problem [36, Section 3.1, Lemma 28]. While in this regime, parameter direction is enough to achieve a good classification accuracy, it is not enough to achieve a good density estimation accuracy, which requires accurate estimation of the parameter norm as well. We discuss more thoroughly the state of the art in this area in Appendix III.E.

Logistic Regression in the presence of non i.i.d data Additionally, from a technical standpoint, there are a few differences between the standard setup of logistic regression studied in this works, and the setup of NRE. The major difference with our setup, is that the loss does not write as a sample average of independent loss functions, but instead of a U-statistics, making existing analyses of logistic regression, typically assuming i.i.d data, stale.

III.3 Background

III.3.1 Neural Ratio Estimation

We briefly elaborate on the description of NRE provided in the introduction. NRE's estimator is given by

$$\begin{aligned}\psi_{N,\text{NRE}} &:= \arg \min_{\psi \in \Psi} J_{\text{NRE}}(\psi) \\ J_{\text{NRE}}(\psi) &:= \frac{1}{N} \sum_{i=1}^N \log(h(g_\psi(X_i, \theta_i))) + \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} \log(1 - h(g_\psi(X_i, \theta_j)))\end{aligned}\tag{III.3}$$

where $h(y) = \frac{1}{1+\exp(-y)}$ is the sigmoid function, from which the posterior estimate is given by Equation III.2. In the following we denote, for all $\psi \in \Psi$, the population version of $J_{\text{NRE}}(\psi)$ by

$$\bar{J}_{\text{NRE}}(\psi) := \mathbb{E}[J_{\text{NRE}}(\psi)]\tag{III.4}$$

The learning procedure of NRE is justified by the following lemma.

Lemma III.3.1. *Assume that $p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1) = p_{\psi^*}(\theta = \bullet | X_1)$ almost surely for some $\psi^* \in \Psi$, where p_{ψ^*} is given in Equation III.2. Then it holds that $\psi' \in \arg \min_{\psi \in \Psi} \bar{J}_{\text{NRE}}(\psi)$ if and only if $p_{\psi'}(\theta = \bullet | X_1) = p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1)$ almost surely.*

We provide a simple proof (in the style of 256, Lemma 5.35) of Lemma III.3.1 in Section III.C.1 for convenience.

Incomplete Neural Ratio Estimation Importantly, $J_{\text{NRE}}(\psi)$ contains a quadratic $\mathcal{O}(N^2)$ number of terms, and thus model evaluations. This suggests that an near-exact minimization of $J_{\text{NRE}}(\psi)$ may be computationally prohibitive for large N . As a first step towards understanding the behavior of inexactly-minimized NRE for large N , we consider the estimator resulting from minimizing a subsampled version of $J_{\text{NRE}}(\psi)$, given by:

$$\begin{aligned}\psi_{N,\text{INRE}} &:= \arg \min_{\psi \in \Psi} J_{\text{INRE}}(\psi) \\ J_{\text{INRE}}(\psi) &:= \frac{1}{N} \sum_{i=1}^N \log(h(g_\psi(X_i, \theta_i))) + \frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} \log(1 - h(g_\psi(X_i, \theta_j))) \quad (\text{III.5})\end{aligned}$$

where elements of \mathcal{D} all belong to $\mathcal{D}_c := \{(i, j), i \neq j, i \in [N], j \in [N]\}$, the set of all pairs (i, j) drawn without replacement from $[N] \times [N]$. Setting $\mathcal{D} = \mathcal{D}_c$ recovers the original NRE estimator, while using \mathcal{D} such that $|\mathcal{D}| = \mathcal{O}(N)$ yields an estimator with a objective function that can be computed in linear time. In this work, we will consider the case where \mathcal{D} is formed of $r(N)$ tuples (i, j) drawn from \mathcal{D}_c (i) uniformly at random, (ii) with replacement, and (iii) such that $\lim_{N \rightarrow \infty} \frac{r}{N} \rightarrow \alpha$. We leave for future work the investigation of other sampling schemes.

III.3.2 Noise Contrastive Estimation

III.3.2.1 Method

NRE shares important similarities with Noise Contrastive Estimation (NCE, 92) a well-known (unconditional) density estimation. Given N i.i.d samples $\{Z_i\}_{i=1}^N$ from some distribution $\mu(z \in \cdot)$, M i.i.d samples $\{\tilde{Z}_i\}_{i=1}^M$ sampled from some distribution $\gamma(z \in \cdot)$, and a parametric model of the form $p_\psi(z \in \cdot) := \gamma(z \in \cdot) \times e^{g_\psi(z)}$ (for some model g_\bullet) and some parameter space Ψ , Noise Contrastive Estimation (NCE, 92) finds an estimate $p_{\hat{\psi}}$ of μ , by performing logistic regression on the samples $\{Z_i\}_{i=1}^N$ and $\{\tilde{Z}_i\}_{i=1}^M$, e.g. by maximizing

$$J_{\text{NCE}}(\psi) := \frac{1}{N} \left(\sum_{i=1}^N \log \left(h \left(\log \left(\frac{1}{v} \frac{dp_\psi}{d\gamma}(Z_i) \right) \right) \right) \right) +$$

$$\sum_{i=1}^M \log \left(1 - h \left(\log \left(\frac{1}{v} \frac{dp_\psi}{d\gamma}(\tilde{Z}_i) \right) \right) \right) \quad (\text{III.6})$$

where we assumed that $M = vN$ for some $v > 0$.

III.3.2.2 Known properties of NCE

NCE is well-known in the statistical estimation literature, and its properties have been studied in particular in Gutmann and Hyvärinen [92] and Lee et al. [151].

Consistency and asymptotic normality Gutmann and Hyvärinen [92] claimed the consistency and asymptotic normality of the NCE estimators; in particular, they claimed, under standard assumptions, and assuming that there exists a ψ^* such that $p_{\psi^*} = \mu$, that $\sqrt{N}(\psi_{\text{NCE},N} - \psi^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_{\text{NCE}})$, where

$$\mathcal{V}_{\text{NCE}} = \mathcal{I}_v^{-1} - \left(1 + \frac{1}{v} \right) \mathcal{I}_v^{-1} \mathbb{E} [P_v(Z_1) \nabla_\psi g_{\psi^*}(Z_1)] \mathbb{E} [P_v(Z_1) \nabla_\psi g_{\psi^*}^\top(Z_1)] \mathcal{I}_v^{-1}, \quad (\text{III.7})$$

where $P_v(z) := \frac{1}{1 + \frac{1}{v} \frac{du}{dy}(z)}$ and $\mathcal{I}_v = \mathbb{E} [P_v(Z_1) \nabla_\psi g_{\psi^*}(Z_1) \nabla_\psi g_{\psi^*}(Z_1)^\top]$. We flesh out this expression briefly. Note first that \mathcal{V}_{NCE} writes as the difference of two positive semi-definite matrices; the first one, \mathcal{I}_v^{-1} , is the Fisher information matrix of the logistic regression problem solved by NCE. The second one comes from the fact that the number of samples from the noise distribution is exactly set to be a multiple of the number of samples from the data distribution; this in turns induces a negative correlation between the gradient of the log likelihood of the noise samples and the one of the data samples, leading to a reduction of the asymptotic variance of the estimator. This dependence is weak however, and in general, we expect that \mathcal{V}_{NCE} mostly behaves like \mathcal{I}_v^{-1} .

Asymptotic efficiency of NCE More recently, Lee et al. [151] provided examples of distributions and models for which NCE with a Gaussian noise distribution behaves (asymptotically) exponentially badly with dimension: in particular, for such distributions and models, they showed that

$$\lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NCE},N} - \psi^*\|^2] = \exp(\Omega(d)).$$

This result was showed for *product* distributions, although we believe that these arguments could be extended to distributions with weak, but non-zero dependence structures like Markov chains.

Differences between NCE and NRE At a high level, NRE shares some similarities with NCE: it seeks to estimate a (conditional) density by solving a classification problem. However, there are some notables differences: (i) the target is a conditional density and not a density, (ii) the noise distribution is set to be the product of marginals, and (iii) the loss function is not a sample average of independent losses anymore. For this reason, we expect the properties of NRE, in particular its statistical efficiency, to possibly differ from NCE.

III.4 Consistency and Asymptotic Efficiency of NRE

In this section, we establish the consistency and asymptotic normality of (I)NRE. To do so, we construct a framework able to handle both NCE and NRE in a unified manner. We make use of this framework to (i) establish the consistency and asymptotic normality of NCE in a rigorous manner, complementing the previous analysis of Gutmann and Hyvärinen [92], and (ii) establish the consistency and asymptotic normality of NRE.

III.4.1 Consistency and Asymptotic normality of NRE

Define $\mathcal{I}_{\text{NRE},1} := \mathbb{E} \left[\nabla_{\psi} (\log \frac{dp_{\psi}}{d\pi}(x, \theta) \nabla_{\psi} \log \frac{dp_{\psi}}{d\pi}(x, \theta) (X_1, \theta_1)^{\top} \text{sp}'(g_{\psi}(X_1, \theta_1)) \right]$, $\text{sp}(z) := \log(1+z)$ is the softplus function, which can be understood as an expected conditional equivalent of \mathcal{I}_v for $v = 1$. We make the following assumptions:

Assumption A1. *For $\check{Z} \in (X_1, \theta_1), (X_1, \theta_2)$, the function $\Psi \ni \psi \mapsto \log g_{\psi}(\check{Z})$ is almost surely thrice differentiable. Moreover, it holds, for all $\psi \in \Psi, k \in [3]$, $\|\nabla_{\psi}^{(k)} \log g_{\psi}(\check{Z})\| \leq l_k(\check{Z})$ a.s. for some l_k such that $\mathbb{E}[l_k(\check{Z})^{3-k+1}] < +\infty$.*

Assumption A2. *There exists a ψ^* such that $p_{\psi^*}(\theta \in \bullet | X_1) = p_{\Theta|X}(\theta \in \bullet | X_1)$ almost surely*

Assumption A3. *The matrix $\mathcal{I}_{\text{NRE},1}$ is full rank.*

Assumption A4. $\nabla_\psi J_N$ admits at most one root, for all $N \in \mathbb{N}$, where the loss function J_N is to be specified.

[A1](#) is a standard domination assumption which will ensure that $\bar{J}_{\text{NRE}}(\psi)$ (on which, as we will see, NRE relies) admits derivatives obtained by differentiating under the integral sign. Indeed, as shown in Appendix [III.D.1](#), under [A1](#), \bar{J}_{NRE} is thrice differentiable. [A2](#) is a well-specification assumption, which stipulates that there exists a model p_{ψ^*} which perfectly equals the data distribution. [A3](#) is an assumption ensuring that the Hessian of the objective at the optimum is full rank. While [A2](#) is hard to verify in practice, cases when assumptions [A1](#) and [A4](#) arise include exponential family-type function classes of the form

$$p_\psi(\theta \in \cdot | x) = \pi(\theta \in \bullet) e^{\langle \phi(x, \theta), \psi_{-1} \rangle - \psi_{-1}}$$

which includes the exponential family densities with sufficient statistics $\phi(z)$ and sufficient parameter ψ_{-1} , by setting $\psi_{-1} = \log \int e^{\langle \phi(x, \theta), \psi_{-1} \rangle} \pi(d\theta) < +\infty$, which under Assumption [A2](#) will be independent of x at ψ^* .

To account for the fact that J_{NRE} may not have a minimizer, we consider the NRE estimator $\psi_{\text{NRE},N}$ given by

$$\psi_{\text{NRE},N} := \begin{cases} \nabla_\psi J_{\text{NRE}}^{-1}(\{0_d\}) & \text{if } (\nabla_\psi J_{\text{NRE}})^{-1}(\{0_d\}) \neq \emptyset \\ \psi_0 & \text{otherwise} \end{cases} \quad (\text{III.8})$$

and similarly $\psi_{\text{INRE},N}$.

III.4.1.1 Consistency of (I)NRE

We are now ready to state our first result, establishing the consistency of NRE.

Proposition III.4.1. *Assume [A1](#), [A3](#), [A2](#), and that J_{NRE} satisfies [A4](#). Then it holds that*

- (i) $\lim_{N \rightarrow \infty} \mathbb{P} \left[\{(\nabla_\psi J_{\text{NRE}})^{-1}(\{0\}) \neq \emptyset\} \right] = 1$
- (ii) $\psi_{\text{NRE},N} \xrightarrow{P} \psi^*$,

and similarly for $\psi_{\text{INRE},N}$.

The proof relies on more general results established in Appendix III.A, which can be specialized to both NRE and NCE-style estimators. This allows, us, in passing, to establish in Appendix III.B the consistency of NCE in a rigorous manner, and without assuming the uniform convergence of the empirical NCE objective to its population counterpart, complementing the previous analysis of Gutmann and Hyvärinen [92].

III.4.1.2 Asymptotic variance of the (I)NRE estimator

Similarly, we now state the asymptotic normality and variance of the (I)NRE estimator. These statements are proved in Appendix III.C; as for Proposition III.4.1, they rely on more general results established in Appendix III.A, which can be specialized to both NRE and NCE-style estimators, and thanks to which we also rigorously establish in Appendix III.B.2 the asymptotic normality and variance of NCE.

Proposition III.4.2. *Under the setup of Proposition III.4.1, it holds that*

$$\sqrt{N}(\psi_{\text{NRE},N} - \psi^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\text{NRE},1}^{-1} \Lambda_{\text{NRE}} \mathcal{I}_{\text{NRE},1}^{-1})$$

where:

$$\Lambda_{\text{NRE}} = \mathcal{I}_{\text{NRE},2} - \left(\mathbb{E} \left[(B(X_1) - B(\theta_1)) (B(X_1) - B(\theta_1))^{\top} \right] + BB^{\top} \right) \quad (\text{III.9})$$

and we defined

$$\begin{aligned} \mathcal{I}_{\text{NRE},1} &:= \mathbb{E} \left[P_{\text{NRE}}(X_1, \theta_1) \times \nabla_{\psi} g_{\psi^*}(\theta_1, X_1) \nabla_{\psi} g_{\psi^*}(\theta_1, X_1)^{\top} \right] \\ \mathcal{I}_{\text{NRE},2} &:= \mathbb{E} \left[P_{\text{NRE}}(X_1, \theta_1)^2 \times \nabla_{\psi} g_{\psi^*}(\theta_1, X_1) \nabla_{\psi} g_{\psi^*}(\theta_1, X_1)^{\top} \right] \\ P_{\text{NRE}}(x, \theta) &:= \frac{1}{1 + \frac{d p_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta)}, \quad B(X_1, \theta_1) := P(X_1, \theta_1) \nabla_{\psi} g_{\psi^*}(X_1, \theta_1), \\ B(X_1) &:= \mathbb{E}[B(X_1, \theta_1) | X_1], \quad B(\theta_1) := \mathbb{E}[B(X_1, \theta_1) | \theta_1] \\ B &:= \mathbb{E}[B(X_1, \theta_1)]. \end{aligned}$$

Moreover, the INRE estimator $\psi_{\text{INRE},N}$ satisfies

$$\begin{aligned} \sqrt{N}(\psi_{\text{INRE},N} - \psi^*) &\xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{\text{NRE},1}^{-1} \Lambda_{\text{INRE}} \mathcal{I}_{\text{NRE},1}^{-1}) \\ \Lambda_{\text{INRE}} &= (1 - \alpha^{-1}) \mathcal{I}_{\text{NRE},2} + \alpha^{-1} \mathcal{I}_{\text{NRE},1} - \mathbb{E} \left[(B(X_1) - B(\theta_1))(B(X_1) - B(\theta_1))^{\top} \right] \\ &\quad - (1 + \alpha^{-1}) BB^{\top}. \end{aligned} \tag{III.10}$$

III.4.1.3 Discussion

We briefly discuss the main properties of NRE and INRE's asymptotic variances, which dictate, asymptotically, how statistically efficient these estimators are. First we note that \mathcal{V}_{NRE} , as \mathcal{V}_{NCE} , can be decomposed as the difference of two positive semi-definite matrices. The second one, as for NCE, comes from correlations between the terms of the gradient. However, in addition to the correlations present in NCE, NRE contains additional one arising from the fact that the noise distribution samples and the data distribution samples are not independent anymore. We now discuss the first one, given by $\mathcal{I}_{\text{NRE},1}^{-1} \mathcal{I}_{\text{NRE},2} \mathcal{I}_{\text{NRE},1}^{-1}$, which we expect to be the dominant one. First, we note that by treating NRE as a $\mathcal{I}_{\text{NRE},1}$ matches the \mathcal{I}_{NCE} one would have obtained by performing NCE between N samples from the data distribution and N other independent samples from the product of marginals, using the same model as NRE. However, the presence of a U-statistics in NRE's loss induces a variance decrease, as we have:

$$\begin{aligned} \mathcal{I}_{\text{NRE},2} &= \int P_{\text{NRE}}^2 \nabla_{\psi} g_{\psi^*} \nabla_{\psi} g_{\psi^*}^{\top} d p_{\mathcal{X}, \Theta} \\ &\stackrel{(a)}{\preceq} \int P_{\text{NRE}} \nabla_{\psi} g_{\psi^*} \nabla_{\psi} g_{\psi^*}^{\top} d p_{\mathcal{X}, \Theta} = \mathcal{I}_{\text{NRE},1} \end{aligned} \tag{III.11}$$

where (a) follows since $P_{\text{NRE}} < 1$, implying

$$\mathcal{I}_{\text{NRE},1}^{-1} \mathcal{I}_{\text{NRE},2} \mathcal{I}_{\text{NRE},1}^{-1} \preceq \mathcal{I}_{\text{NRE},1}^{-1}. \tag{III.12}$$

This difference is non-negligible: indeed, the r.h.s of Equation III.12 can be expected to roughly scale with $\mathbb{E}[P_{\text{NRE}}(X_1, \theta_1)]^{-1}$. However, as we will see, for very different noise and data distributions, $P_{\text{NRE}}(X_1, \theta_1)$ will be on average very small, which is what [151] leveraged to construct hard distributions for NCE. On the other hand, the l.h.s of Equation III.12 can be understood as “homogenous” in $P_{\text{NRE}}(X_1, \theta_1)$, as $\mathcal{I}_{\text{NRE},2}$ contains a $P_{\text{NRE}}(X_1, \theta_1)^2$ term, each $\mathcal{I}_{\text{NRE},1}^{-1}$ two $P_{\text{NRE}}(X_1, \theta_1)^{-1}$. Thus, in the case of NRE, possible variance blow ups may be more attributed to a more subtle “Jensen Gap”, e.g. to situations when

$$\frac{\mathbb{E}[P_{\text{NRE}}(X_1, \theta_1)^2]}{\mathbb{E}[P_{\text{NRE}}(X_1, \theta_1)]^2} = e^{\Omega(d)}.$$

Finally, we note that INRE’s variance “sits in between” NRE’s and NCE’s one. In particular, decomposing $\mathcal{V}_{\text{INRE}}$ as the difference of positive definite matrices as above, we see that the first one is a convex combination of $\mathcal{I}_{\text{NRE},1}$ and $\mathcal{I}_{\text{NRE},2}$, with weights $(1 - \alpha^{-1})$ and α^{-1} respectively, where we recall that $\alpha := \lim_{N \rightarrow \infty} \frac{r}{N}$. A large α (e.g. using INRE with a large number of noise samples) implies a variance close to NRE’s one, while a small α (e.g. using INRE with a small number of noise samples) implies a variance close to NCE’s one. We thus expect INRE to suffer from the possibly drastic “inhomogeneity” issues of NCE, provided that the negative terms in its variance can be controlled.

III.5 Constructing hard distributions for NRE

Next, we construct hard conditional distributions for both NRE and INRE, in the sense that the asymptotic variance of the NRE estimator will be much larger than the Cramér-Rao lower bound. The idea consists in creating a distribution resulting in $\nabla_\psi J_{\text{NRE}}(\psi^*)$ having a variance which is not small enough to offset the flatness of the NRE objective at the optimum. Our results can be summarized as follows:

Theorem III.5.1. *Under the setup of Theorem III.4.1, there exists distributions, such that under mild assumptions on the model,*

$$(i) \quad \lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E}[\|\psi_{\text{INRE},N} - \psi^*\|^2] = \exp(\Omega(d)), \text{ and}$$

$$(ii) \lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NRE},N} - \psi^*\|^2] = \exp(\Omega(d)).$$

The remainder of this section is dedicated to the proof of Theorem III.5.1.

III.5.1 Lower bounds on the asymptotic efficiency of NCE

As a first step, we consider the simpler setup of NCE, and seek to obtain lower bounds as a function of $\text{TV}(\gamma, \nu)$, the total variation between γ and μ . Our approach shares similarity with the one of Lee et al. [151]; however, unlike Lee et al. [151], our results hold for non-Gaussian distributions, any model satisfying certain integrability assumptions, and for all d . These changes are made possible by our handling the negative definite term in NCE's asymptotic variance, which does not resort to anti-concentration inequalities as in Lee et al. [151]. In this section, we assume $\nu = 1$. These bounds will be directly used to prove the lower bounds on INRE stated above.

Theorem III.5.2. *Under the setup of Theorem III.B.2, the NCE estimator $\psi_{\text{NCE},N}$ satisfies*

$$\begin{aligned} (i) \quad & \|\mathcal{I}_1^{-1}\| \geq (1 - \text{TV}(\gamma, \mu))^{-1/2} (\mathbb{E} [\|g_{\psi^*}(Z_1)\|^4])^{-1/2} \\ (ii) \quad & \mathcal{I}_1^{-1} \mathbb{E} [P_1(Z_1) \nabla_{\psi} g_{\psi^*}(Z_1)] \mathbb{E} [P_1(Z_1) \nabla_{\psi} g_{\psi^*}^\top(Z_1)] \mathcal{I}_1^{-1} \preceq \mathbb{E} [P_1(Z_1)] \mathcal{I}_1^{-1} \end{aligned}$$

Implying

$$\lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NCE},N} - \psi^*\|^2] \geq \frac{\text{TV}(\mu, \gamma)}{(1 - \text{TV}(\mu, \gamma))^{1/2} \mathbb{E} [\|\nabla_{\psi} g_{\psi^*}(Z_1)\|^4]^{1/2}}.$$

The proof can be found in Appendix III.B.2. In particular, if the denominator $(1 - \text{TV}(\mu, \gamma))^{1/2} \times \mathbb{E} [\|g_{\psi^*}(Z_1)\|^4]^{1/2}$ scales poorly with dimension, so will the variance of the NCE estimator. For instance, when μ and γ are product distributions, $\mu = \mu_1^{\otimes d}$, $\gamma = \gamma_1^{\otimes d}$, it holds

$$1 - \text{TV}(\mu, \gamma) \leq \frac{1}{2} (1 - D_H(\mu, \gamma)) = \frac{1}{2} (1 - D_H(\mu_1, \gamma_1))^d$$

where $D_H(\mu, \nu)$ is the Hellinger distance between μ and ν . Consequently, we have, for models such that $\mathbb{E} [\|\nabla_{\psi} g_{\psi^*}(Z_1)\|^4] = O(\text{Poly}(d))$, that:

$$\lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NCE},N} - \psi^*\|^2] = \exp(\Omega(d)).$$

We now leverage this result to prove the statement of Theorem III.5.1 for INRE.

III.5.2 Proof of the lower bounds on the asymptotic efficiency of INRE

As discussed above, the reason for the inefficiency of NCE is more drastic than the one of NRE. Thus, while INRE interpolates between NRE and the one of NCE, our argument will mostly rely on the inefficiency of NCE. In particular, by Theorem III.5.2, we directly have that

$$BB^\top \preceq (1 - \text{TV}(p_{\mathcal{X}, \Theta}, p_{\mathcal{X}} \otimes \pi)) \mathcal{I}_{\text{NRE}, 1}$$

and

$$\begin{aligned} & \alpha \left\| \mathcal{I}_{\text{NRE}, 1}^{-1} \right\| \\ & \geq \alpha (1 - \text{TV}(p_{\mathcal{X}, \Theta}, p_{\mathcal{X}} \otimes \pi))^{-1/2} \mathbb{E} [\|\nabla_\psi g_{\psi^*}((X_1, \theta_1))\|^4]^{-1/2}, \end{aligned} \quad (\text{III.13})$$

with the latter exponentially blowing up with dimensions for product distributions; in particular, it is not necessary to try to lower bound $\mathcal{I}_{\text{NRE}, 2}$ for INRE to obtain a blow up. It remains to control the additional negative terms present in the asymptotic variance of INRE, namely

$$\mathbb{E} \left[(B(X_1) - B(\theta_1)) (B(X_1) - B(\theta_1))^\top \right].$$

We do so in the lemma below.

Lemma III.5.3. *Assume that $p_{\mathcal{X}|\Theta} = p_{1,\mathcal{X}|\Theta} \otimes \dots \otimes p_{d,\mathcal{X}|\Theta}$ is a product distribution, with $\text{KL}(p_{1,\mathcal{X}|\Theta}(\bullet|\theta), p_{1,\mathcal{X}}) > \beta$ for some $\beta > 0$, and we noted $p_{1,\mathcal{X}}$ the marginal distribution of the first coordinate of X_1 . Then it holds that:*

$$\text{Tr} \left(\mathbb{E} \left[(B(X_1) - B(\theta_1)) (B(X_1) - B(\theta_1))^\top \right] \right) = \exp(-\Omega(d)) \mathcal{I}_{\text{NRE}, 1}.$$

The proof (see Appendix III.C.4) is simple, and heavily relies on the uniform lower bound on the KL divergence between the conditional and marginal distributions. We

can now conclude the proof of Theorem III.5.1 for the INRE case:

Proof of Theorem III.5.1 for INRE. Assuming distributions of the setup of Lemma III.5.3, putting everything together, we obtain, that

$$\begin{aligned} \mathcal{I}_{\text{NRE},1}^{-1} \Lambda_{\text{INRE}} \mathcal{I}_{\text{NRE},1}^{-1} &\succeq (\alpha - e^{-\Omega(d)}) \mathcal{I}_{\text{NRE},1}^{-1}. \\ \implies \lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NRE},N} - \psi^*\|^2] &= \text{Tr}(\mathcal{I}_{\text{NRE},1}^{-1} \Lambda_{\text{INRE}} \mathcal{I}_{\text{NRE},1}^{-1}) \\ &\geq (\alpha - e^{-\Omega(d)}) \|\mathcal{I}_{\text{NRE},1}^{-1}\| \\ &= \exp(\Omega(d)), \end{aligned}$$

where we assumed that $\mathbb{E}[\|\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\|^4] = O(\text{Poly}(d))$. \square

III.5.3 Proof of Theorem III.5.1 for NRE

In this section, we now extend the arguments above to the case of NRE. Whilst relying on the same family of population objectives, the NRE has a lower asymptotic variance than the one of NCE. To obtain lower bounds for the NRE estimator, we will need a more refined set of tools. In particular, while the lower bounds for NCE required only product distributions to yield an exponentially bad accuracy, our NRE lower bound requires a specific distribution to yield the same behavior. Let us note

$$\mathcal{V}_{\text{NRE}} = \Sigma_1 - \Sigma_2$$

where

$$\begin{aligned} \Sigma_1 &:= \mathcal{I}_{\text{NRE},1}^{-1} \mathcal{I}_{\text{NRE},2} \mathcal{I}_{\text{NRE},1}^{-1} \\ \Sigma_2 &:= \mathcal{I}_{\text{NRE},1}^{-1} \left(\mathbb{E} \left[(B(X_1) - B(\theta_1)) (B(X_1) - B(\theta_1))^{\top} \right] + BB^{\top} \right) \mathcal{I}_{\text{NRE},1}^{-1} \end{aligned}$$

As for NCE, we separately provide a lower bound for the second moment term Σ_1 , and an upper bound for the squared (conditional) first moment term Σ_2 . In this section, we consider NCE models of the form:

$$g_{\psi}(x, \theta) = g_{0,\psi_{-1}}(x, \theta) + \psi_{-1} \quad (\text{III.14})$$

which were the ones studied in Gutmann and Hyvärinen [92].

III.5.3.1 The hard distribution

Let us note $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_1$. Let $L > 1$ be a positive integer, μ be an arbitrary measure on \mathcal{X}_1 , and let p_1, \dots, p_L L probability densities (w.r.t μ) such that $p_i(x)p_j(x) = 0$ for all $x \in \mathcal{X}$, and $1 \leq i \neq j \leq L$. Finally, let $\eta \in (0, 1)$. We will consider a pair of random variables θ_1, X_1 distributed according to

$$\begin{cases} \theta_{11}, \dots, \theta_{1d} & \stackrel{\text{i.i.d}}{\sim} \mathcal{U}([L]) \\ X_{1i} | \theta_{1i} = i & \sim ((1 - \eta)p_i(x_i) + \frac{\eta}{L} \sum_{j=1}^L p_j(x_i))\mu(dx_i) \end{cases} \quad (\text{III.15})$$

Here, X_{1i}, θ_{1i} stand for the i -th components of X_1, θ_1 , respectively. The associated marginal distribution of X_{1i} is then $p_{\mathcal{X}}(dx_i) = (\frac{1}{L} \sum_{j=1}^L p_j(x))\mu(dx_i)$, and we have:

$$\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}(x_i, \theta = j) \quad (= f(x, j)) = \eta + (1 - \eta)L \times p_j(x).$$

We now proceed to lower bound the asymptotic variance of the NRE estimator for this distribution.

III.5.3.2 Lower bound on Σ_1

We first provide a generic lower bound on Σ_1 depending on $\mathcal{I}_{\text{NRE},1}$ and $\mathcal{I}_{\text{NRE},2}$. Indeed,

$$\begin{aligned} & \text{Tr}(\mathcal{I}_{\text{NRE},1}^{-1} \mathcal{I}_{\text{NRE},2} \mathcal{I}_{\text{NRE},1}^{-1}) \\ & \stackrel{(a.1)}{=} \text{Tr}(\mathcal{I}_{\text{NRE},2} \mathcal{I}_{\text{NRE},1}^{-2}) \stackrel{(a.2)}{=} \text{Tr}(\mathcal{I}_{\text{NRE},2}^{1/2} \mathcal{I}_{\text{NRE},1}^{-2} \mathcal{I}_{\text{NRE},2}^{1/2}) \\ & \stackrel{(b)}{\geq} \frac{1}{\lambda_{\max}(\mathcal{I}_{\text{NRE},1})^2} \text{Tr}(\mathcal{I}_{\text{NRE},2}) \end{aligned} \quad (\text{III.16})$$

where, in (a.1) and (a.2), we used the cyclicity of the trace, and in (b) we used that $\mathcal{I}_{\text{NRE},1}^{-1} \succeq \frac{I_d}{\lambda_{\max}(\mathcal{I}_{\text{NRE},1})}$, that $A \succeq B$ implies $CAC \succeq CBC$ for A, B, C p.s.d, and that $A \preceq B$ implies $\text{Tr}(A) \leq \text{Tr}(B)$, for A, B p.s.d. We now proceed by upper bounding $\lambda_{\max}(\mathcal{I}_{\text{NRE},1})$ and lower bounding $\text{Tr}(\mathcal{I}_{\text{NRE},2})$.

Step 1: Upper bounding $\mathcal{I}_{\text{NRE},1}$ Applying Hölder's inequality instead of the Cauchy-Schwarz inequality, one obtains

$$\begin{aligned}
 & \lambda_{\max}(\mathcal{I}_{\text{NRE},1}) \\
 & \leq \text{Tr}(\mathcal{I}_{\text{NRE},1}) = \int \frac{1}{1 + \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \times \pi)}(x, \theta)} \|\nabla_{\psi} g_{\psi^*}(x, \theta)\|^2 p_{\mathcal{X},\Theta}(dx, d\theta) \\
 & \leq \left(\int \left(\frac{1}{1 + \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \times \pi)}(x, \theta)} \right)^v p_{\mathcal{X},\Theta}(dx, d\theta) \right)^{\frac{1}{v}} \\
 & \quad \times \left(\int \|\nabla_{\psi} g_{\psi^*}(x, \theta)\|^{\frac{2v}{v-1}} p_{\mathcal{X},\Theta}(dx, d\theta) \right)^{\frac{v-1}{v}} \\
 & \stackrel{(a)}{\leq} \left(\int \frac{1}{1 + \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \times \pi)}(x, \theta)} \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \times \pi)}(x, \theta) (p_{\mathcal{X}} \times \pi)(dx, d\theta) \right)^{\frac{1}{v}} \quad (\text{III.17}) \\
 & \quad \times \left(\int \|\nabla_{\psi} g_{\psi^*}(x, \theta)\|^{\frac{2v}{v-1}} p_{\mathcal{X},\Theta}(dx, d\theta) \right)^{\frac{v-1}{v}} \\
 & \stackrel{(b)}{\leq} \left(\int \min \left(\frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X},\Theta}}(x, \theta), 1 \right) (p_{\mathcal{X}} \times \pi)(dx, d\theta) \right)^{\frac{1}{v}} \\
 & \quad \times \left(\int \|\nabla_{\psi} g_{\psi^*}(x, \theta)\|^{\frac{2v}{v-1}} p_{\mathcal{X},\Theta}(dx, d\theta) \right)^{\frac{v-1}{v}} \\
 & = (1 - \text{TV}_v((p_{\mathcal{X}} \times \pi), p_{\mathcal{X},\Theta}))^{1/v} \left(\mathbb{E} \left[\|\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\|^{\frac{2v}{v-1}} \right] \right)^{\frac{v-1}{v}}
 \end{aligned}$$

where in (a), we relied on the fact that $v > 1$, and in (b), we used the fact that $\frac{r}{1+r} \leq \min(1, r)$ for all $r > 0$.

Step 2: Lower bounding $\mathcal{I}_{\text{NRE},2}$ Lower-bounding $\mathcal{I}_{\text{NRE},2}$ is slightly more delicate, and makes heavy use of the form of the hard distribution considered above. We obtain the following lemma:

Lemma III.5.4. *For the hard distribution considered above, and under the setup of Theorem III.4.2, it holds that, for any $\epsilon > 0$*

$$\text{Tr}(\mathcal{I}_{\text{NRE},2}) \geq \frac{1}{4} \left(\frac{1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X},\Theta})}{2} \right)^{1+\epsilon}.$$

Proof. Our starting point is a decomposition of $(1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta}))$ as :

$$\begin{aligned}
 1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta}) &= \int \min\left(\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}, 1\right) (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &= \int \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \geq 1\right\}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &\quad + \int \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \leq 1\right\}} \frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &= \int \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \geq 1\right\}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) + \int \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \leq 1\right\}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &= (p_{\mathcal{X}} \times \pi) \left[\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 1 \right] + p_{\mathcal{X}, \Theta} \left[\frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} > 1 \right]
 \end{aligned}$$

where we noted $(p_{\mathcal{X}} \times \pi) \left[\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 1 \right] := \mathbb{P} \left[\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}(X_1, \theta_2) > 1 \right]$, and similarly for the second term. Since models of the form of Equation III.14 verify

$$\|\nabla_{\psi} g_{\psi^*}(x, \theta)\| = \|(\nabla_{\psi_{-1}} g_{0, \psi_{-1}}(x, \theta), 1)\| \geq 1$$

for such models, we have:

$$\begin{aligned}
 \text{Tr}(\mathcal{I}_{\text{NRE}, 2}) &= \int \frac{1}{\left(1 + \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}\right)^2} \|\nabla_{\psi} g_{\psi^*}(x, \theta)\|^2 p_{\mathcal{X}, \Theta}(dx, d\theta) \\
 &\geq \int \frac{1}{\left(1 + \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}\right)^2} p_{\mathcal{X}, \Theta}(dx, d\theta) \\
 &\geq \int \frac{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}}{\left(1 + \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}\right)^2} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &\geq \frac{1}{4} \int \min\left(\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}, 1\right) \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \leq 1\right\}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \\
 &\geq \frac{1}{4} \left(\int \min\left(\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}, 1\right) (p_{\mathcal{X}} \times \pi)(dx, d\theta) \right. \\
 &\quad \left. - \int \min\left(\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}, 1\right) \mathbb{I}_{\left\{\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} \geq 1\right\}} (p_{\mathcal{X}} \times \pi)(dx, d\theta) \right)
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{4} \left((1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta})) - (p_{\mathcal{X}} \times \pi) \left[\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 1 \right] \right) \\
&= \frac{1}{4} p_{\mathcal{X}, \Theta} \left[\frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} > 1 \right]
\end{aligned}$$

One way to ensure that the lower bound provided in Equation III.16 is to relate the two tail probabilities $(p_{\mathcal{X}} \times \pi) \left[\frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 1 \right]$, $p_{\mathcal{X}, \Theta} \left[\frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} > 1 \right]$, which will allow us to link the last line of the derivations above to $(1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta}))$. In the case of product distributions $(p_{\mathcal{X}} \times \pi) = p_{1, \mathcal{X}} \times \pi_1 \otimes \dots \otimes p_{1, \mathcal{X}} \times \pi_1$, $p_{\mathcal{X}, \Theta} = p_{1, \mathcal{X}, \Theta} \otimes \dots \otimes p_{1, \mathcal{X}, \Theta}$, these two tail probabilities can be rewritten as

$$\begin{aligned}
p_{\mathcal{X}, \Theta} \left[\log \frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} > 0 \right] &= p_{\mathcal{X}, \Theta} \left[\frac{1}{n} \sum_{i=1}^n \log \frac{dp_{1, \mathcal{X}} \times \pi_1}{dp_{1, \mathcal{X}, \Theta}}(C_i) > 0 \right] \\
&\stackrel{\text{i.i.d.}}{\sim} C_1, \dots, C_d \sim p_{1, \mathcal{X}, \Theta}
\end{aligned}$$

and

$$\begin{aligned}
(p_{\mathcal{X}} \times \pi) \left[\log \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 0 \right] &= (p_{\mathcal{X}} \times \pi) \left[\frac{1}{n} \sum_{i=1}^n \log \frac{dp_{1, \mathcal{X}, \Theta}}{dp_{1, \mathcal{X}} \times \pi_1}(D_i) > 0 \right], \\
&\stackrel{\text{i.i.d.}}{\sim} D_1, \dots, D_d \sim p_{1, \mathcal{X}} \times \pi_1
\end{aligned}$$

which suggests resorting to (anti-) concentration results to study such quantities. Generic results guarantee exponential concentration of the log likelihood ratio $\log \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}$, but not at a rate sufficient to be negligible to the first term $(1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta}))$. However, with our specific hard distribution, we will be able to obtain the right concentration results.

Let us consider the random variables:

$$\begin{aligned}
A &:= \frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}}(X_1, \theta_1) = \begin{cases} \frac{1}{(1-\eta)L+\eta} & \text{w.p. } 1 - \frac{(L-1)\eta}{L} \\ \frac{1}{\eta} & \text{w.p. } \frac{(L-1)\eta}{L} \end{cases} \\
B &:= \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}(X_1, \theta_2) = \begin{cases} (1-\eta)L+\eta & \text{w.p. } \frac{1}{L} \\ \eta & \text{w.p. } \frac{L-1}{L} \end{cases}
\end{aligned}$$

as well as $C = \log A, D = \log B$. It holds that $\mathbb{E}[A] = \mathbb{E}[B] = 1$. Noting

- $C_{\min} = -\log((1-\eta)L + \eta) < 0$
- $C_{\max} = -\log \eta > 0$
- $D_{\min} = -C_{\max} < 0$
- $D_{\max} = -C_{\min} > 0$

we have:

$$C = C_{\min} + U_1 \times (C_{\max} - C_{\min})$$

$$D = D_{\min} + U_2 \times (D_{\max} - D_{\min})$$

where $U_1 \sim \mathcal{B}(p_C := \frac{(L-1)\eta}{L})$, $U_2 \sim \mathcal{B}(p_D := \frac{1}{L})$. Let $C_i, i = 1, \dots, d$ i.i.d copies of C , and $D_i, i = 1, \dots, d$ i.i.d copies of D . It holds that:

$$\begin{aligned} \log \frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} &= \sum_{i=1}^d C_i = dC_{\min} + (C_{\max} - C_{\min}) \sum_{i=1}^d U_{1i} \\ &= dC_{\min} + (C_{\max} - C_{\min})V_1 \end{aligned}$$

where $V_1 \sim \mathcal{B}(d, \frac{(L-1)\eta}{L})$, and similarly for $\log \frac{d p_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)}$. Thus, it holds that:

$$\begin{aligned} \log \frac{d(p_{\mathcal{X}} \times \pi)}{dp_{\mathcal{X}, \Theta}} > 0 &\iff V_1 \geq \frac{-dC_{\min}}{C_{\max} - C_{\min}} := d \times k_C(\eta, L) \\ \log \frac{d p_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 0 &\iff V_2 \geq \frac{-dD_{\min}}{D_{\max} - D_{\min}} := \frac{dC_{\max}}{C_{\max} - C_{\min}} = d(1 - k_C(\eta, L)) \end{aligned}$$

For $u, v \in [0, 1]$, define $kl(u, v) = u \log u/v + (1-u) \log(1-u)/(1-v)$, which is infinitely differentiable on $(0, 1) \times (0, 1)$. To obtain bounds, we need to distinguish the cases (I) : $p_C < k_C$, $\neg(I)$: $p_C > k_C$, (II) : $p_D < 1 - k_C$, $\neg(II)$: $p_D > 1 - k_C$. Note that

$$p_D > 1 - k_C \implies p_C = \frac{L-1}{L}\eta < 1 - p_D < k_C$$

implying $\neg(II) \wedge \neg(I)$ can be ruled out.

Case 1: $(I) \wedge (II)$ Using Chernoff bounds for Binomial variables, it holds that:

$$(ii) := (p_{\mathcal{X}} \times \pi) \left[\log \frac{dp_{\mathcal{X}, \Theta}}{dp_{\mathcal{X}} \times \pi} > 0 \right] \leq \exp \left(-d \times kl(1 - k_C(\eta, L), \frac{1}{L}) \right) = p_D(\eta, L, d)$$

while using anti-concentration bounds for Binomial variables, we have:

$$(i) := p_{\mathcal{X}, \Theta} \left[\log \frac{dp_{\mathcal{X}} \times \pi}{dp_{\mathcal{X}, \Theta}} > 0 \right] \geq \frac{1}{\sqrt{2d}} \exp \left(-d \times kl(k_C(\eta, L), \frac{(L-1)\eta}{L}) \right) = p_C(\eta, L, d)$$

Note that

$$\begin{aligned} (b) &:= kl \left(1 - k_C, \frac{1}{L} \right) = (1 - k_C) \log \frac{1 - k_C}{\frac{1}{L}} + k_C \log \frac{k_C}{\frac{L-1}{L}} \\ (a) &:= kl \left(k_C, \frac{(L-1)\eta}{L} \right) = k_C \log \frac{k_C}{\frac{(L-1)\eta}{L}} + (1 - k_C) \log \frac{1 - k_C}{1 - \frac{(L-1)\eta}{L}} \end{aligned}$$

implying:

$$(a) - (b) = k_C \log \frac{1}{\eta} + (1 - k_C) \log \frac{1}{(1 - \eta)L + \eta} = 0$$

Case 2: $\neg(I) \wedge (II)$ In that case, using a lower tail anti-concentration bound for Binomial variables, we have:

$$\begin{aligned} (i) &:= p_{\mathcal{X}, \Theta} \left[\log \frac{dp_{\mathcal{X}} \times \pi}{dp_{\mathcal{X}, \Theta}} > 0 \right] \\ &\geq \frac{1}{\sqrt{2d}} \exp \left(-d \times kl(1 - k_C(\eta, L), 1 - \frac{(L-1)\eta}{L}) \right) \\ &= p_C(\eta, L, d) \end{aligned}$$

again, noting:

$$(a)' := 1 - k_C \times \log \frac{1 - k_C}{1 - \frac{(L-1)\eta}{L}} + k_C \log \frac{k_C}{\frac{(L-1)\eta}{L}}$$

we have:

$$(a') - (b) = k_C \log \frac{1}{\eta} + (1 - k_C) \log \frac{\frac{1}{L}}{1 - \frac{(L-1)\eta}{L}} = 0$$

Case 3: $\neg(I) \wedge (II)$ Using a lower tail Chernoff bound for (ii), we have

$$(ii) \leq \exp\left(-d \times kl(k_C, \frac{L-1}{L})\right) = p_D(\eta, L, d)$$

Noting

$$(b)' = k_C \log \frac{k_C}{\frac{L-1}{L}} + (1 - k_C) \log \frac{1 - k_C}{\frac{1}{L}}$$

we have:

$$(a) - (b') = k_C \log \frac{1}{\eta} + (1 - k_C) \log \frac{1}{(1 - \eta)L + \eta} = 0$$

Conclusion regarding the lower bound on $\mathcal{I}_{NRE,2}$ In all three cases, the lower (resp. upper) bounds on (i) (resp. (ii)) have the same exponent, and we have:

$$p_C \geq \frac{1}{\sqrt{2d}} \times p_D$$

We thus have, for any $\varepsilon > 0$ and large enough $d \equiv d(\varepsilon)$, that

$$p_C \geq p_D^{1+\varepsilon}$$

Implying

$$\begin{aligned} p_C &= p_D + p_C - p_D \\ &= (1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta})) - p_D \\ &\geq (1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta})) - p_C^{\frac{1}{1+\varepsilon}} \\ \implies 2p_C^{\frac{1}{1+\varepsilon}} &\geq p_C + p_C^{\frac{1}{1+\varepsilon}} \geq (1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta})) \end{aligned}$$

Implying

$$\begin{aligned} \text{Tr}(\mathcal{I}_{NRE,2}) &\geq \frac{1}{4} \left(1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta}) - (p_{\mathcal{X}} \times \pi) \left[\log \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \times \pi)} > 0 \right] \right) \\ &= \frac{p_C}{4} \\ &\geq \frac{1}{4} \left(\frac{1 - \text{TV}((p_{\mathcal{X}} \times \pi), p_{\mathcal{X}, \Theta})}{2} \right)^{1+\varepsilon} \end{aligned}$$

□

Final lower bound on Σ_1 Plugging the last result into Equation III.16, we obtain

$$\begin{aligned} \text{Tr}(\Sigma_1) &= \text{Tr}\left(\mathcal{I}_{\text{NRE},1}^{-1}\mathcal{I}_{\text{NRE},2}\mathcal{I}_{\text{NRE},1}^{-1}\right) \\ &\geq \frac{1}{4(1-\text{TV}_v((p_{\mathcal{X}} \times \pi), p_{\mathcal{X},\Theta}))^{\frac{2-v(1+\epsilon)}{v}} \left(\mathbb{E}\left[\|\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\|^{\frac{2v}{v-1}}\right]\right)^{\frac{2(v-1)}{v}}} \end{aligned} \quad (\text{III.18})$$

which will explode with d for $v < \frac{2}{1+\epsilon}$, as long as $\mathbb{E}\left[\|\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\|^{\frac{2v}{v-1}}\right] = \mathcal{O}(\text{Poly}(d))$. It remains to upper bound Σ_2 , the (average conditional) squared first moment.

III.5.3.3 Upper bounding Σ_2

First, using the same reasoning that in the beginning of the proof of Theorem III.5.2, but conditionally on θ_1 , we have:

$$\begin{aligned} &B(\theta_1)B(\theta_1)^\top \\ &= \mathbb{E}\left[P_1(X_1, \theta_1)\nabla_{\psi} g_{\psi^*}(X_1, \theta_1) \mid \theta_1\right] \mathbb{E}\left[P_1(X_1, \theta_1)\nabla_{\psi} g_{\psi^*}^\top(X_1, \theta_1) \mid \theta_1\right] \\ &\leq \mathbb{E}\left[P_1(X_1, \theta_1) \mid \theta_1\right] \times \mathbb{E}\left[P_1(X_1, \theta_1)\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\nabla_{\psi} g_{\psi^*}^\top(X_1, \theta_1) \mid \theta_1\right] \end{aligned}$$

Note first that, for all $x, i \in \mathcal{X}_1 \times [L]$,

$$\frac{dp_{1,\mathcal{X},\Theta}}{dp_{1,\mathcal{X}} \times \pi_1}(x, i) = \begin{cases} (1-\eta)L + \eta & \text{if } x \in \text{supp}(p_i) \\ \eta & \text{otherwise.} \end{cases}$$

Consequently, since, given $\theta = i, x \in \text{supp}(p_i)$ w.p $1 - \frac{(L-1)\eta}{L}$,

$$\frac{dp_{1,\mathcal{X},\Theta}}{dp_{1,\mathcal{X}} \times \pi_1}(X_{11}, \theta_{11}) \mid \theta_{11} = i \sim \begin{cases} (1-\eta)L + \eta & \text{w.p. } 1 - \frac{(L-1)\eta}{L} \\ \eta & \text{w.p. } \frac{(L-1)\eta}{L} \end{cases}$$

Consequently, $\mathbb{E}\left[\frac{dp_{\mathcal{X},\Theta}}{dp_{\mathcal{X}} \times \pi}(X_1, \theta_1) \mid \theta_1\right] = \mathbb{E}\left[\frac{dp_{1,\mathcal{X},\Theta}}{dp_{1,\mathcal{X}} \times \pi_1}(X_{11}, \theta_{11}) \times \dots \times \frac{dp_{1,\mathcal{X},\Theta}}{dp_{1,\mathcal{X}} \times \pi_1}(X_{1d}, \theta_{1d}) \mid \theta_1\right]$ is independent of θ_1 , and we have

$$\mathbb{E}[P_1(X_1, \theta_1) \mid \theta_1] = \mathbb{E}[P_1(X_1, \theta_1)].$$

We thus have

$$\begin{aligned}
& \mathbb{E} \left[B(\theta_1) B(\theta_1)^\top \right] \\
& \preceq \mathbb{E} \left[\mathbb{E} [P_1(X_1, \theta_1)] \times \mathbb{E} \left[P_1(X_1, \theta_1) \nabla_\psi g_{\psi^*}(X_1, \theta_1) \nabla_\psi g_{\psi^*}^\top(X_1, \theta_1)^\top \mid \theta_1 \right] \right] \\
& \preceq \mathbb{E} [P_1(X_1, \theta_1)] \times \mathbb{E} \left[P_1(X_1, \theta_1) \nabla_\psi g_{\psi^*}(X_1, \theta_1) \nabla_\psi g_{\psi^*}^\top(X_1, \theta_1)^\top \right] \\
& = \mathbb{E} [P_1(X_1, \theta_1)] \times \mathcal{I}_{\text{NRE},1}
\end{aligned}$$

Similarly, given $x \in \mathcal{X}$, and i_x the unique integer in $[L]$ such that $x \in \text{supp}(p_{i_x})$, $p(\theta = i_x \mid x) = 1 - \frac{(L-1)\eta}{L}$, and $p(\theta = j) = \frac{1}{L}$, for all $j \in [L] \setminus \{i_x\}$, we have:

$$\frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \times \pi)}(X_1, \theta_1) \mid X_1 = x \sim \begin{cases} (1-\eta)L + \eta & \text{w.p. } 1 - \frac{(L-1)\eta}{L} \\ \eta & \text{w.p. } \frac{(L-1)\eta}{L} \end{cases}$$

Implying $\mathbb{E} [(B(X_1)B(X_1)^\top)] \preceq \mathbb{E} [P_1(X_1, \theta_1)] \times \mathcal{I}_{\text{NRE},1}$. Finally, $\mathbb{E} [BB^\top]$ is exactly the term of the first part of the proof of Theorem III.5.2. Consequently, we have

$$\begin{aligned}
\Sigma_2 &= \mathcal{I}_{\text{NRE},1}^{-1} \left(\mathbb{E} \left[(B(X_1) - B(\theta_1)) (B(X_1) - B(\theta_1))^\top \right] + BB^\top \right) \mathcal{I}_{\text{NRE},1}^{-1} \\
&\preceq 5 \times \mathbb{E} [P_1(X_1, \theta_1)] \times \mathcal{I}_{\text{NRE},1}^{-1} \\
&\implies \text{Tr}(\Sigma_2) \leq 5d
\end{aligned} \tag{III.19}$$

Since $\text{Tr}(\mathcal{I}_{\text{NRE},1})^{-1} \leq d\lambda_{\max}(\mathcal{I}_{\text{NRE},1}) = \frac{d}{\lambda_{\min}(\mathcal{I}_{\text{NRE},1})} \leq \frac{d}{\mathbb{E}[P(X_1, \theta_1)]}$, since $\mathcal{I}_{\text{NRE},1} \geq \mathbb{E}[P(X_1, \theta_1)]$.

III.5.3.4 Putting everything together

Combining Equations III.18 and III.19, we obtain:

$$\begin{aligned}
& \lim_{N \rightarrow \infty} N \times \mathbb{E} [\|\psi_{\text{NRE},N} - \psi^*\|^2] \\
& \geq \frac{1}{4 \left(1 - \text{TV}_v((p_{\mathcal{X}} \times \pi), p_{\mathcal{X},\Theta}) \right)^{\frac{2-v(1+\epsilon)}{v}} \left(\mathbb{E} \left[\left\| \nabla_\psi g_{\psi^*}(X_1, \theta_1) \right\| \frac{2v}{v-1} \right] \right)^{\frac{2(v-1)}{v}}} - 5d
\end{aligned} \tag{III.20}$$

and the result follows by taking $v < \frac{2}{1+\varepsilon}$, as long as $\mathbb{E} \left[\|\nabla_{\psi} g_{\psi^*}(X_1, \theta_1)\|^{2v/(v-1)} \right] = \mathcal{O}(\text{Poly}(d))$. \square

III.6 Discussion: towards finite sample bounds

In the previous sections, we provided examples of hard distributions, for both INRE and NRE, where the asymptotic variance of the estimators scales exponentially with dimension. The requirements hard NRE distributions had to satisfy were relatively mild, the main one being that it had to write as a product distribution, and the other one imposing a uniform lower bound on the KL between the conditional distributions and the marginals. We expect that similar bounds can be obtained if independence is slightly relaxed, by considering, e.g. time series, and we also expect that other, more sophisticated techniques to handle the negative correlations could relax the uniform lower bound requirement on the KL. This suggests that INRE may struggle in high dimensions in a variety of settings. Constructing hard distributions for NRE was more delicate. We believe that using a similar technique, similar results may be obtained for more standard distributions, such as Gaussians. Nevertheless, while showing that NRE presents shortcomings in high dimensions, our results suggest that from an asymptotic perspective, NRE is more robust to certain pathologies than INRE, owing to the variance reduction it enjoys from using a U-statistics in its gradient estimate. Yet, INRE, whose objective can be computed in linear time, scales better than NRE, whose quadratic-time objective makes it impractical when the number of samples is large.

That being said, our analysis is asymptotic. While asymptotic analyses are common in the analysis of machine learning techniques [92, 37, 151, 136, 247] as they are relatively simple to obtain, allow for both lower and upper bounds, and can align well with finite-sample behavior, they target a quantity (the limit of the rescaled mean-squared error) that is a priori meaningless in practice. Such analyses should thus be taken with caution, as they do not account for certain finite-sample phenomena. In particular, the finite-sample variance of the empirical NRE and NCE losses, obtained using [109], contains a term, which vanishes faster than $\frac{1}{N}$ for NRE, but at the $\frac{1}{N}$

speed for NCE, and which is the reason behind the bad behavior of NCE. While this term is indeed insignificant for NRE asymptotically as $N \rightarrow \infty$, for finite N , and given its exponential scaling with dimension, this term may have a significant detrimental impact of the finite-sample performance of NRE, which may, in certain regimes, suffer from the more general pitfalls of NCE.

More satisfying results would consist in obtaining finite-sample bounds, e.g. bounds that hold for all settings of d, n . To achieve this, one could either rely on general finite sample analyses of M-estimators [185, 229], or on more specific analyses of classification [36, 139, 100].

Appendix

III.A Proofs of General Results

In this section, we establish the consistency and asymptotic normality of NRE. To do so, we construct a framework able to handle both NCE and NRE in a unified manner. We make use of this framework to (i) establish the consistency and asymptotic normality of NCE in a rigorous manner, complementing the previous analysis of 92, and (ii) establish the consistency and asymptotic normality of NRE.

III.A.1 Additional Notations for NCE

NRE shares important similarities with Noise Contrastive Estimation (NCE, 92) a well-known (unconditional) density estimation. Given N i.i.d $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ samples $\{Z_i\}_{i=1}^N$ from some distribution with density (w.r.t the Lebesgue measure) $\mu(\cdot)$, a parametric model of the form $p_\psi(z) := p_{\psi,-1}^0(z) \times e^{-\psi-1}$, and some parameter space Ψ , and M i.i.d samples $\{\tilde{Z}_i\}_{i=1}^M$ sampled from some distribution with known density $\gamma(\cdot)$, Noise Contrastive Estimation (NCE, 92) finds an estimate $p_{\hat{\psi}}$ of $\mu(\cdot)$, by performing logistic regression on the samples $\{Z_i\}_{i=1}^N$ and $\{\tilde{Z}_i\}_{i=1}^M$, with logit model

$\log \frac{1}{v} \frac{dp_\psi}{d\gamma}$, e.g. by maximizing

$$\begin{aligned} J_{\text{NCE}}(\psi) &:= \frac{1}{N} \left(\sum_{i=1}^N \log \left(h \left(\log \frac{1}{v} \frac{dp_\psi}{d\gamma}(Z_i) \right) \right) + \sum_{i=1}^M \log \left(1 - h \left(\log \frac{1}{v} \frac{dp_\psi}{d\gamma}(\tilde{Z}_i) \right) \right) \right) \\ &\propto \frac{-1}{N+M} \sum_{(Y,Z) \in \{(1,Z_i)\}_{i=1}^N \cup \{(-1,Z_i)\}_{i=1}^M} \log(1 + e^{-Y \log \frac{1}{v} \frac{dp_\psi}{d\gamma}(Z)}) \end{aligned} \quad (\text{III.21})$$

where $h(y) = \frac{1}{1+\exp(-y)}$ is the sigmoid function, and we assumed that $M = vN$ for some $v > 0$. In this setting, each point (Y, Z) is distributed marginally¹ according to

$$Y = \begin{cases} 1 & \text{with probability } \frac{1}{1+v} \\ -1 & \text{with probability } \frac{v}{1+v} \end{cases} \quad Z|Y \sim \begin{cases} \mu & \text{if } Y = 1 \\ \gamma & \text{if } Y = -1 \end{cases}$$

yielding a distribution for $Y|Z$ with logits $\log \frac{1}{v} \frac{d\mu}{d\gamma}$. Loosely speaking, as $N \rightarrow \infty$, and if we further assume that there is some ψ^* such that $\mu = p_{\psi^*}$, the logit model $g_\psi := \frac{1}{v} \frac{dp_\psi}{d\gamma}$ will approach the true logits, implying that p_ψ will approach the target μ . In particular, we have $\log \int p_{\psi^*}^0(z) dz = \psi_{-1}$, e.g. ψ_{-1} recovers the log-normalizing constant of the unnormalized $p_{\psi^*}^0$. NCE is well-known in the statistical estimation literature, the properties of its estimator—namely, its consistency, asymptotic normality, and asymptotic variance—have been studied in Gutmann and Hyvärinen [92].

In the following we denote, for all $\psi \in \Psi$,

$$\bar{J}_{\text{NCE}}(\psi) := \mathbb{E}[J_{\text{NCE}}(\psi)] = -(1+v) \times \mathbb{E} \left[\log \left(1 + e^{-Y g_\psi(Z)} \right) \right] \quad (\text{III.22})$$

For concrete instances of μ, v, γ and p_\bullet , which will arise in later sections, we will make the dependence on \bar{J}_{NCE} explicit through the notations $\bar{J}_{\text{NCE}}(\psi) = \bar{J}_{\text{NCE},\mu,v,p_\bullet}(\psi)$, $g_\psi = g_{\psi,v,\gamma,p_\bullet}$. $J_{\text{NCE}}(\psi) = \bar{J}_{\text{NCE},\mu,v,p_\bullet}(\psi)$, $g_\psi = g_{\psi,v,\gamma}$, and

III.A.2 General framework

Here, we lay out the general framework we will use to analyze both NCE and NRE. The framework follows the notations of the NCE section, assuming a pair of random

¹This can be proved using exchangeability of the data at hand.

variables $(Z_1, \tilde{Z}_1) \in \mathcal{Z}^2$ and a model p_\bullet with parameter space Ψ . The variables (Z_2, \tilde{Z}_1) will be set to be $((X_1, \theta_1), (X_1, \theta_2))$ to treat the case of NRE within this framework. Moreover, to handle NCE and NRE in a unified manner, a population loss function \bar{J} , and a sequence of random loss functions J_N , set to be the NCE and NRE empirical losses.

Assumption AA1. *For $\check{Z} \in Z_1, \tilde{Z}_1$, the function $\Psi \ni \psi \mapsto \log p_\psi(\check{Z})$ is a.s. thrice differentiable. Moreover, it holds, for all $\psi \in \Psi, k \in [3]$, $\|\nabla_\psi^{(k)} \log \frac{dp_{\psi^*}}{d\gamma}(\check{Z})\| \leq l_k(\check{Z})$ a.s. for some l_k such that $\mathbb{E}[l_k(\check{Z})^{3-k+1}] < +\infty$.*

Assumption AA2. *There exists a unique $\psi^* \in \text{int}(\Psi)$ such that $p_{\psi^*} = \mu$.*

Assumption AA3. *The matrix $\mathbb{E} \left[\nabla_\psi \log \frac{dp_{\psi^*}}{d\gamma} p_{\psi^*}(Z_1) \otimes \nabla_\psi \log \frac{dp_{\psi^*}}{d\gamma} (Z_1) \mathbf{s} p'(Z_1) \right]$ is full rank.*

Assumption AA4. *$\nabla_\psi J_N$ admits at most one root, for all $N \in \mathbb{N}$.*

We consider the following estimator, given by:

$$\psi_N = \begin{cases} \nabla_\psi J_N^{-1}(\{0_d\}) & \text{if } (\nabla_\psi J_N)^{-1}(\{0_d\}) \neq \emptyset \\ \psi_0 & \text{otherwise} \end{cases} \quad (\text{III.23})$$

In words, our estimator consists in the unique solution to the equation $\nabla_\psi J_N(\psi) = 0$ if it exists, and otherwise, defaults back to some fixed value ψ_0 .

We now follow the standard asymptotic analysis program outlined in the introduction by deriving generic consistency and asymptotic normality results for M-estimators of losses converging to \bar{J}_{NCE} (including NRE and NCE). This result will be used to prove the consistency and asymptotic normality of both NCE and NRE as special cases. The first theorem is a general consistency result, using well-known arguments in the asymptotic statistics literature, but which does not assume that the loss writes as a sample averages.

Theorem III.A.1. Assume that the triplet $(Z_1, \tilde{Z}_1, p_{\bullet})$ satisfies assumption AA1, AA2, AA3, Let $J_N(\psi)$ be a function such that (i) $J_N(\psi)$ is thrice differentiable, (ii) $\nabla_{\psi}^{(k)} J_N(\psi) \xrightarrow{p} \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}}$ for $k \in \{0, 1, 2\}$, (iii) AA4, and (iv) $\max_{\psi \in \overline{\mathcal{V}(0, \psi^*)}} \|\nabla_{\psi}^{(3)} J_N(\psi)\| = O_p(1)$, for some neighborhood $\mathcal{V}(\psi^*)$ around ψ^* . Then it holds that

- $\lim_{N \rightarrow \infty} \mathbb{P} [\{(\nabla_{\psi} J_N^{-1}(\{0\})) \neq \emptyset\}] = 1$
- $\psi_N \xrightarrow{p} \psi^*$.

Proof of Theorem III.A.1. Proving the theorem is a direct adaptation of known results [256, Theorem 5.42] As, by assumption, \bar{J}_{NRE} is thrice differentiable, its second derivative is continuous, and bounded in a bounded neighborhood of ψ^* . Consequently, as $\nabla_{\psi}^{(2)} \bar{J}_{\text{NCE}}(\psi^*)$ is full-rank, there exists a sub-neighborhood \mathcal{V} where $\nabla_{\psi}^{(2)} \bar{J}_{\text{NCE}}(\psi)$ is full rank. By the inverse function theorem, there exists a $\delta > 0$ such that $\nabla_{\psi} \bar{J}_{\text{NCE}}(\psi)$ is a homeomorphism from some open neighborhood $\overline{G_{\delta}} \subset \mathcal{V}$ to $\overline{\mathcal{B}(0_d, \delta)}$. Moreover, $\overline{\mathcal{B}(0, \delta)}$, we have

$$\nabla_{\psi} (\nabla_{\psi} \bar{J}_{\text{NCE}})^{-1}(h) = \nabla_{\psi}^{(2)} (\nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}(h))^{-1}.$$

Since $\nabla_{\psi}^{(2)} \bar{J}_{\text{NCE}}$ is continuous and full rank and has bounded norm on \mathcal{V} , so is its inverse. Denoting $\sigma_{\max,1} = \max_{h \in \overline{\mathcal{B}(0, \delta)}} \|\nabla_{\psi} (\nabla_{\psi} \bar{J}_{\text{NCE}})^{-1}(h)\| \in (0, \infty)$ and $\sigma_{\max,2} = \max_{\psi \in G_{\delta}} \|\nabla_{\psi}^{(2)} \bar{J}_{\text{NCE}}(\psi)\| \in (0, \infty)$, by Lemma III.D.4, for all $h, h' \in \overline{\mathcal{B}(0, \delta)}$

$$\text{diam}(G_{\delta}) = \max_{\psi, \psi' \in G_{\delta}} \|\psi - \psi'\| = \max_{h, h'} \|\nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}(h) - \nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}(h')\| \leq 2\sigma_{\max,1} \delta$$

where the second equality holds since $\nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}$ is a homeomorphism from $\overline{G_{\delta}}$ to $\overline{\mathcal{B}(0, \delta)}$. Similarly,

$$\begin{aligned} 2\delta &= \text{diam}(\overline{\mathcal{B}(0, \delta)}) \\ &= \max_{h, h' \in \mathcal{B}(0, \delta)} \|\psi - \psi'\| \\ &= \max_{\psi, \psi' \in G_{\delta}} \|\nabla_{\psi} \bar{J}_{\text{NCE}}(\psi) - \nabla_{\psi} \bar{J}_{\text{NCE}}(\psi')\| \leq \sigma_{\max,2} \times \text{diam}(G_{\delta}) \end{aligned}$$

Implying

$$\frac{2}{\sigma_{\max,2}} \delta \leq \text{diam}(G_\delta) \leq 2\sigma_{\max,1} \delta$$

Now, a Taylor expansion of J_N and \bar{J}_{NCE} around ψ^* yields:

$$\begin{aligned} & \nabla_\psi J_N(\psi) - \nabla_\psi \bar{J}_{\text{NCE}}(\psi) \\ &= \nabla_\psi J_N(\psi^*) - \nabla_\psi \bar{J}_{\text{NCE}}(\psi^*) + \left(\nabla_\psi^{(2)} J_N(\psi^*) - \nabla_\psi^{(2)} \bar{J}_{\text{NCE}}(\psi^*) \right) (\psi - \psi^*) \\ &+ \left\langle \psi - \psi^*, (\nabla_\psi^{(3)} J_N(\tilde{\psi}_1) - \nabla_\psi^{(3)} \bar{J}_{\text{NCE}}(\tilde{\psi}_2))(\psi - \psi^*) \right\rangle \end{aligned}$$

where $\tilde{\psi}_1, \tilde{\psi}_2$ are points in the line segment between ψ and ψ^* . Note that by assumption, we have

$$\left\langle \psi^* - \psi, \left(\nabla_\psi^{(3)} J_{\text{NCE}}(\tilde{\psi}_1) - \nabla_\psi^{(3)} \bar{J}_{\text{NCE}}(\tilde{\psi}_2) \right), \psi^* - \psi \right\rangle = \|\psi - \psi^*\|^2 O_p(1).$$

Similarly, by assumption, for $k \in \{1, 2, 3\}$, we have

$$\nabla_\psi^{(k)} J_N(\psi^*) = \nabla_\psi^{(k)} \bar{J}_{\text{NCE}}(\psi^*) + o_p(1)$$

Using this fact, we obtain

$$\sup_{\psi \in \overline{G_\delta}} \|\nabla_\psi J_{\text{NCE}}(\psi) - \nabla_\psi \bar{J}_{\text{NCE}}(\psi)\| = o_p(1) + \delta o_p(1) + \delta^2 O_p(1), \quad (\text{III.24})$$

where we used the fact that $\frac{2}{\sigma_{\max,2}} \leq \|\psi - \psi^*\| \leq 2\sigma_{\max,1} \delta$ for all $\psi \in G_\delta$. Let $K_{\delta,N}$ the event such that the r.h.s is bounded by δ , which approaches 1 as $N \rightarrow \infty$ for $\delta < 1$. Let

$$A : h \in \mathcal{B}(0, \delta) \longmapsto h - \nabla_\psi J_N \circ \nabla_\psi \bar{J}_{\text{NCE}}^{-1}(h)$$

On $K_{\delta,N}$, it holds that

$$\|A(h)\| = \|h - \nabla_\psi J_N \circ \nabla_\psi \bar{J}_{\text{NCE}}^{-1}(h)\| = \|(\nabla_\psi \bar{J}_{\text{NCE}} - \nabla_\psi J_N) \circ \underbrace{\nabla_\psi \bar{J}_{\text{NCE}}^{-1}(h)}_{\in G_\delta}\| \leq \delta$$

where the last inequality used Equation III.24. Thus, the map $\overline{\mathcal{B}(0, \delta)}$ into itself, and

is continuous. Thus, by the Brouwer Fixed Point theorem, it thus admits a fixed point h_n , and we have $\nabla_{\psi} J_N(\nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}(h_N)) = 0$, e.g. $(\nabla_{\psi} \bar{J}_{\text{NCE}}^{-1}(h_N)) \in \nabla_{\psi} J_N(\{0\})$. We thus have that

$$1 \leq \mathbb{P} [\nabla_{\psi} J_N(\{0\}) \neq \emptyset] \geq \mathbb{P} [K_{\delta, N}] \xrightarrow[N \rightarrow \infty]{} 1,$$

and thus, by the sandwich theorem, we have that:

$$\lim_{N \rightarrow \infty} \mathbb{P} [\nabla_{\psi} J_N(\{0\}) \neq \emptyset] = 1.$$

To prove the last statement, we have that:

$$\begin{aligned} & \mathbb{P} [\|\psi_N - \psi^*\| < \delta] \\ &= \mathbb{P} [\|\psi_N - \psi^*\| < \delta | K_n] P[K_n] + \mathbb{P} [\|\psi_N - \psi^*\| < \delta | K_N^c] P[K_N^c] \\ &\geq P[K_N] + 1_{\|\psi_0 - \psi^*\| < \delta} P[K_N^c] \rightarrow 1 \end{aligned}$$

Implying by the sandwich theorem that $\psi_N \xrightarrow{P} \psi^*$ as $N \rightarrow \infty$. \square

Next, we formally establish asymptotic normality result for estimator of the form of Equation III.23, and provide a formula for its asymptotic covariance matrix. This result will be used to prove the asymptotic normality of both NCE and NRE as special cases.

Theorem III.A.2. *Under the setup of Theorem III.A.1, and assuming further that $\sqrt{N} \times \nabla_{\psi} J_N(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda)$ for some $\Lambda \in \mathbb{R}^{d \times d}$, then we have that*

$$\sqrt{N} (\psi^* - \psi_N) \xrightarrow{d} \mathcal{N}(0, \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1} \Lambda \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1}).$$

Proof of Theorem III.A.2. The proof is again a direct modification application of [256, Theorem 5.41]. Let E_N be the event when J_{NCE} admits a root, e.g. $E_N = \{(\nabla_{\psi} J_{\text{NCE}})^{-1}(\{0_d\}) \neq \emptyset\}$. It holds that:

$$\begin{aligned} & H_{\psi} J_{\text{NCE}}(\psi^*) (\psi^* - \psi_N) \\ &= H_{\psi} J_{\text{NCE}}(\psi^*) (\psi^* - \psi_N) 1_{E_N} + H_{\psi} J_{\text{NCE}}(\psi^*) (\psi^* - \psi_N) 1_{E_N^c} \end{aligned} \tag{III.25}$$

Let us note $\tilde{\psi}_N := \nabla_{\psi} J_{\text{NCE}}^{-1}(\{0_d\})$. On E_N , it holds that

$$\begin{aligned} 0 &\stackrel{(a)}{=} \nabla J_{\text{NCE}}(\psi_N) \\ &\stackrel{(b)}{=} \nabla_{\psi} J_{\text{NCE}}(\psi^*) + \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)(\tilde{\psi}_N - \psi^*) \\ &\quad + \langle \tilde{\psi}_N - \psi^*, \nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_1)(\tilde{\psi}_N - \psi^*) \rangle \end{aligned}$$

where ψ_1 belongs to the line segment between $\tilde{\psi}_N$ and ψ^* , (a) holds by definition of ψ_N and E_N , and (b) uses a Taylor expansion of ∇J_{NCE} around ψ^* . Multiplying out by \sqrt{N} and isolating the Hessian term, we have, on E_N ,

$$\begin{aligned} &\sqrt{N} \times \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)(\psi^* - \tilde{\psi}_N) 1_{E_N} \\ &= 1_{E_N} \times \sqrt{N} \nabla_{\psi} J_{\text{NCE}}(\psi^*) + 1_{E_N} \sqrt{N} \langle \tilde{\psi}_N - \psi^*, \nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_1)(\tilde{\psi}_N - \psi^*) \rangle \end{aligned}$$

Now, we have

$$\begin{aligned} 1_{E_N} \langle \tilde{\psi}_N - \psi^*, \nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_1)(\tilde{\psi}_N - \psi^*) \rangle &\leq 1_{E_N} \|\tilde{\psi}_N - \psi^*\|^2 \|\nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_1)\| \\ &\leq \|\tilde{\psi}_N - \psi^*\|^2 O_P(1) \end{aligned}$$

where the last inequality holds by assumption. Since, J_N satisfies the assumptions of Theorem III.A.1, Theorem III.A.1, holds, and we have that $\psi_N \xrightarrow{P} \psi^*$. Consequently, by definition, $\|\psi_N - \psi^*\| = o_p(1)$, and thus, we have:

$$1_{E_N} \langle \tilde{\psi}_N - \psi^*, \nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_1)(\tilde{\psi}_N - \psi^*) \rangle = a_N \times \|\psi_N - \psi^*\|$$

where $a_N = o_p(1)$. On the other hand, since $\mathbb{P}(E_N^c) \xrightarrow[N \rightarrow \infty]{} 0$,

$$\begin{aligned} &\nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)(\psi^* - \tilde{\psi}_N) 1_{E_N^c} = \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)(\psi^* - \psi_0) \times o_p(1) \\ &\implies \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)(\psi^* - \tilde{\psi}_N) 1_{E_N^c} := b_N \xrightarrow{P} 0_d. \end{aligned}$$

Putting everything together, we thus have,

$$(\nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*) + a_N) \times \sqrt{N} 1_{E_N} (\psi^* - \tilde{\psi}_N) = 1_{E_N} \sqrt{N} \nabla J_{\text{NCE}}(\psi^*) + b_N$$

Since, by assumption, we have

$$\sqrt{N} \nabla_{\psi} J_{\text{NCE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

as well as $1_{E_N} \xrightarrow{p} 1$ and $b_N \xrightarrow{p} 0$ by Theorem III.A.1, by Slutsky's theorem, it holds that

$$1_{E_N} \sqrt{N} \times \nabla_{\psi} J_{\text{NCE}}(\psi^*) + b_N \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

By Lemma III.D.3, this implies that

$$\sqrt{N} (\psi^* - \psi_n) \xrightarrow{d} \mathcal{N}(0, \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1} \Lambda \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1}).$$

□

III.B Consistency and Asymptotic Efficiency of NCE

Before moving forward, we use the results of the previous section to provide general consistency and efficiency results for NCE which extends beyond the guarantees of [92] in the following ways: first, they do not require uniform convergence of the loss J_{NCE} to \bar{J}_{NCE} over Ψ , which typically places restrictive assumptions over the model class \mathcal{P}_{Ψ} . Second, they deal with the problem of the existence of solutions to the NCE problem, which were shown to hold only for the population loss in [92], but are not guaranteed for finite N . Third, they rigorously establish the asymptotic distribution of the resulting NCE estimator.

We consider the following NCE estimator, of the form of Equation III.23, but specialized to J_{NCE} , e.g.

$$\psi_{\text{NCE},N} := \begin{cases} \nabla_{\psi} J_{\text{NCE}}^{-1}(\{0_d\}) & \text{if } (\nabla_{\psi} J_{\text{NCE}})^{-1}(\{0_d\}) \neq \emptyset \\ \psi_0 & \text{otherwise} \end{cases} \quad (\text{III.26})$$

Using Theorem III.A.1, we now establish consistency of NCE.

Corollary III.B.1. *Under the setup of Theorem III.A.1, specialized to $J_N := J_{\text{NCE}}$, it*

holds that

- $\lim_{N \rightarrow \infty} \mathbb{P} \left[\{(\nabla_{\psi} J_{\text{NCE}})^{-1}(\{0\}) \neq \emptyset\} \right] = 1$
- $\psi_{\text{NCE},N} \xrightarrow{p} \psi^*$.

Proof of Corollary III.B.1. We show that J_{NCE} verifies the four assumptions necessary to apply the theorem. By Lemma III.D.1, \bar{J}_{NCE} and J_{NCE} are thrice differentiable, and $\mathbb{E} \left[\nabla_{\psi}^{(k)} J_{\text{NCE}}(\psi) \right] = \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}}$ for $k \in \{0, 1, 2, 3\}$. Thus, we have, for $k \in \{0, 1, 2, 3\}$,

$$\begin{aligned} \nabla_{\psi}^{(k)} J_{\text{NCE}}(\psi^*) &= \frac{N}{N+M} \times \frac{1}{N} \left(\sum_{i=1}^N \nabla_{\psi}^{(k)} f_{\psi^*}(1, Z_i) + \sum_{i=1}^M \nabla_{\psi}^{(k)} f_{\psi^*}(-1, \tilde{Z}_i) \right) \\ &= \frac{1}{N(1+v)} \sum_{i=1}^N \left(\nabla_{\psi}^{(k)} f_{\psi^*}(1, Z_i) + \sum_{j=0}^{v-1} \nabla_{\psi}^{(k)} f_{\psi^*}(-1, \tilde{Z}_{(i-1) \times v+j}) \right) \\ &= \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}}(\psi^*) + o_p(1) \end{aligned}$$

where the last line used the law of large numbers, and the independence of Z_i with \tilde{Z}_j, Z_j for $i \neq j$. Thus, condition (ii) is satisfied. (iii) is satisfied by assumption. Finally, as AA1 holds, using the notations of Equation III.50, it holds that

$$\|\nabla_{\psi}^{(3)} \bar{J}_{\text{NCE}}(\tilde{\psi}_1)\| \leq (1+v) \times \mathbb{E}[a_3(Z)] < +\infty .$$

Moreover, we have

$$\begin{aligned} \|\nabla_{\psi}^{(3)} J_{\text{NCE}}(\tilde{\psi}_2)\| &\leq \frac{1}{N} \left(\sum_{i=1}^N a_3(1, Z_i) + \sum_{i=1}^M a_3(-1, \tilde{Z}_i) \right) \\ &= \frac{1}{N} \sum_{i=1}^N \left(a_3(1, Z_i) + \sum_{j=0}^{v-1} a_3(-1, \tilde{Z}_{(i-1) \times v+j}) \right) \\ &= (1+v) \times \mathbb{E}[a_3(Y, Z)] + o_p(1) \end{aligned} \tag{III.27}$$

where the last line used the law of large numbers, and the independence of Z_i with \tilde{Z}_j, Z_j for $i \neq j$. Thus, we indeed have (iv) $\sup_{\psi_1, \psi_2 \in V(\psi^*)} \|J_{\text{NCE}}(\psi_1) - \bar{J}_{\text{NCE}}(\psi_2)\| = O_p(1)$, for any neighborhood $V(\psi^*)$ around ψ^* . The result follows by applying Theorem III.A.1. \square

Next, we use the framework of Section III.A.2 to formally establish the asymptotic normality of the NCE estimator, and provide a formula for its asymptotic covariance matrix.

Corollary III.B.2. *Under the setup of Corollary III.B.1, it holds that*

$$\sqrt{N}(\psi_{\text{NCE},N} - \psi^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}_{\text{NCE}}),$$

where

$$\mathcal{V}_{\text{NCE}} = \mathcal{I}_v^{-1} - \left(1 + \frac{1}{v}\right) \mathcal{I}_v^{-1} \mathbb{E}[P_v(Z_1) \nabla_\psi g_{\psi^*}(Z_1)] \mathbb{E}[P_v(Z_1) \nabla_\psi g_{\psi^*}^\top(Z_1)] \mathcal{I}_v^{-1}. \quad (\text{III.28})$$

Here, $P_v : z \rightarrow \frac{1}{1 + \frac{1}{v} \times \frac{du}{dy}(z)}$ and \mathcal{I}_v is given by

$$\mathcal{I}_v = \mathbb{E}[P_v(Z_1) \nabla_\psi g_{\psi^*}(Z_1) g(Z_1)^\top].$$

To prove this corollary, we start with the following lemma, which characterizes the asymptotic distribution of the gradient of the NCE objective function at the optimum ψ^* .

Lemma III.B.3. *Under AA1, AA2, it holds that:*

$$\sqrt{N} \times \nabla_\psi J_{\text{NCE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0_d, \Lambda)$$

where

$$\begin{aligned} \Lambda &= \mathbb{E}[\nabla_\psi g_{\psi^*}(Z_1) \nabla_\psi g_{\psi^*}(Z_1)^\top \text{sp}'(-g_{\psi^*}(Z_1))] \\ &\quad - \left(1 + \frac{1}{v}\right) \mathbb{E}[\nabla_\psi g_{\psi^*}(Z_1) \text{sp}'(g_{\psi^*}(Z_1))] \mathbb{E}[\nabla_\psi g_{\psi^*}(Z_1) \text{sp}'(g_{\psi^*}(Z_1))]^\top \end{aligned} \quad (\text{III.29})$$

Proof. We assume $v \in \mathbb{N}^*$ for simplicity. Recalling the formula for $\nabla_\psi J_{\text{NCE}}(\psi)$ given in Equation III.51, we have

$$\sqrt{N} \times \nabla_\psi J_{\text{NCE}}(\psi^*) = [I_d, I_d] \times$$

$$\underbrace{\frac{1}{\sqrt{N}} \times \sum_{i=1}^N \left[\nabla_{\psi} g_{\psi}(Z_i) \text{sp}'(-g_{\psi}(Z_i)), \sum_{j=1}^v \nabla_{\psi} g_{\psi}(\tilde{Z}_{v(i-1)+j}) \text{sp}'(g_{\psi}(\tilde{Z}_{v(i-1)+j})) \right]}_{Z_n} \quad (\text{III.30})$$

By the Multivariate Central Limit Theorem, it holds that:

$$Z_n \xrightarrow{d} \mathcal{N}(0_{2d}, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} \text{Cov} [\nabla_{\psi} g_{\psi^*}(Z_i) \text{sp}'(-g_{\psi}(Z_i))] & 0_{d \times d} \\ 0_{d \times d} & v \times \text{Cov} [\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi}(\tilde{Z}_1))] \end{bmatrix}$$

Consequently, by the continuous mapping theorem, it holds that

$$\begin{aligned} & \sqrt{N} \times \nabla_{\psi} J_{\text{NCE}}(\psi^*) \\ & \xrightarrow{d} \mathcal{N}(0_{2d}, [I_d, I_d] \Sigma [I_d, I_d]^{\top}) \\ & = \mathcal{N}(0_{2d}, \text{Cov} [\nabla_{\psi} g_{\psi^*}(Z_i) \text{sp}'(-g_{\psi}(Z_i))] + v \times \text{Cov} [\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))]) \end{aligned}$$

We can make the value of the latter covariance more explicit by noting that

$$\begin{aligned} & \text{Cov} [\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(-g_{\psi^*}(Z_1))] \\ & = \underbrace{\mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(-g_{\psi^*}(Z_1))) (\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(-g_{\psi^*}(Z_1)))^{\top} \right]}_{(i)} \\ & - \underbrace{\mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(-g_{\psi^*}(Z_1))) \right] \mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(-g_{\psi^*}(Z_1))) \right]^{\top}}_{(ii)} \quad (\text{III.31}) \end{aligned}$$

and

$$\text{Cov} [\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))]$$

$$\begin{aligned}
&= \underbrace{\mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))) (\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1)))^\top \right]}_{(iii)} \\
&\quad - \underbrace{\mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))) \right] \mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))) \right]^\top}_{(iv)}. \quad (\text{III.32})
\end{aligned}$$

By Equation III.53 of Lemma III.D.2, we have that:

$$(i) + v \times (iii) = \mathbb{E} \left[\nabla_{\psi} g_{\psi^*}(Z_1) \nabla_{\psi} g_{\psi^*}(Z_1)^\top \text{sp}'(-g_{\psi^*}(Z_1)) \right] = \mathcal{I}_v(\psi^*).$$

On the other hand, by Lemma III.D.2, we have that:

$$\mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))) \right] = -\frac{1}{v} \mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))) \right]$$

Implying

$$\begin{aligned}
(ii) + v \times (iv) &= \left(1 + \frac{1}{v}\right) \mathbb{E} \left[\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(g_{\psi^*}(Z_1)) \right] \mathbb{E} \left[\nabla_{\psi} g_{\psi^*}(Z_1) \text{sp}'(g_{\psi^*}(Z_1)) \right]^\top \\
&= \left(1 + \frac{1}{v}\right) \mathbb{E}[P_v] \mathbb{E}[P_v]^\top
\end{aligned}$$

Combining the two results concludes the proof. \square

Proof of Corollary III.B.2. J_{NCE} was already shown to verify the conditions of Theorem III.A.1. Moreover, we have, by Lemma III.B.3, that

$$\sqrt{N} \nabla_{\psi} J_{\text{NCE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda)$$

where Λ is given in Equation III.29. Consequently, assuming that $\nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)$ is invertible, Theorem III.A.2 holds, and we have:

$$\sqrt{N} (\psi^* - \psi_{\text{NCE},N}) \xrightarrow{d} \mathcal{N}(0, \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1} \Lambda \nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)^{-1}).$$

We conclude the proof of Corollary III.B.2 by expliciting the value of $\nabla_{\psi}^{(2)} J_{\text{NCE}}(\psi^*)$.

We have, by III.D.1, that

$$\begin{aligned}
\nabla_{\psi}^2 J_{\text{NCE}}(\psi^*) &= (1 + v) \times \mathbb{E} \left[\nabla_{\psi}^2 g_{\psi^*}(Z) \times Y \times \text{sp}'(-Y g_{\psi^*}(Z)) \right. \\
&\quad \left. - (\nabla_{\psi} g_{\psi^*}(Z) \otimes \nabla_{\psi} g_{\psi^*}(Z)) \times \text{sp}''(-Y g_{\psi^*}(Z)) \right] . \\
&= (1 + v) \times \mathbb{E} \left[\nabla_{\psi}^2 g_{\psi^*}(Z) \times Y \times \text{sp}'(-Y g_{\psi^*}(Z)) \right] \\
&\quad - \mathbb{E} \left[(\nabla_{\psi} g_{\psi^*}(Z) \otimes \nabla_{\psi} g_{\psi^*}(Z)) \times \text{sp}''(-Y g_{\psi^*}(Z)) \right] .
\end{aligned}$$

The first term is 0 by Lemma III.D.2. Using Lemma III.D.5, we have that:

$$\begin{aligned}
h(-g_{\psi^*}(z))(1 - h(-g_{\psi^*}(z))) &= h(g_{\psi^*}(z))(1 - h(g_{\psi^*}(z))) \\
&= \frac{1}{(1 + \frac{1}{v} \frac{d\mu}{d\gamma}) \times (1 + v \frac{d\gamma}{d\mu})} \quad (\text{III.33})
\end{aligned}$$

implying

$$\begin{aligned}
&(1 + v) \times \mathbb{E} \left[(\nabla_{\psi} g_{\psi}(Z) \otimes \nabla_{\psi} g_{\psi^*}(Z)) \times h(-Y g_{\psi}^*(Z))(1 - h(-Y g_{\psi^*}(Z))) \right] \\
&= \left(\int (\nabla_{\psi} g_{\psi^*}(z) \otimes \nabla_{\psi} g_{\psi^*}(z)) \frac{1}{(1 + \frac{1}{v} \frac{d\mu}{d\gamma}) \times (1 + v \frac{d\gamma}{d\mu})} \right. \\
&\quad \left. \times (1 + v \frac{d\gamma}{d\mu}(z)) \mu(dz) \right) \\
&= \int (\nabla_{\psi} g_{\psi^*}(z) \otimes \nabla_{\psi} g_{\psi^*}(z)) \frac{1}{(1 + \frac{1}{v} \frac{d\mu}{d\gamma}(z))} \mu(dz)
\end{aligned}$$

which is invertible by Assumption AA3. The result follows. \square

III.B.1 Proof of Lemma III.B.4

Lemma III.B.4. *Assume that there exists a ψ^* such that $p_{\psi^*}(Z_1) = \mu(Z_1)$ almost surely. Then it holds that $\psi^* \in \arg \min_{\psi \in \Psi} \bar{J}_{\text{NCE}}(\psi)$.*

Proof of Lemma III.B.4. Note first that $0 \leq J_{\text{NCE}}(\psi^*) = -\mathbb{E}[H[Y|Z]] < \log 2 < +\infty$,

where, $H(Y|Z)$ is the conditional entropy of Y given Z . Thus, it holds that

$$\begin{aligned} & \bar{J}_{\text{NCE}}(\psi) - \bar{J}_{\text{NCE}}(\psi^*) \\ &= \mathbb{E} [\text{KL}(p_{\psi^*}(Y = \cdot|Z)||p_\psi(Y = \cdot|Z))] \\ &\leq \mathbb{E} \left[p_{\psi^*}(Y = 1|Z) \times \left(\sqrt{p_\psi(Y = 1|Z)} - \sqrt{p_\psi(Y = 1|Z)} \right)^2 \right. \\ &\quad \left. + (1 - p_{\psi^*}(Y = 1|Z)) \times \left(\sqrt{1 - p_\psi(Y = 1|Z)} - \sqrt{1 - p_\psi(Y = 1|Z)} \right)^2 \right] \end{aligned}$$

with equality if and only if $p_\psi(Y = 1|Z)(= h(\log p_\psi(Z)/\gamma(Z))) = p_{\psi^*}(Y = 1|Z) = h(\log(\frac{p(Z)}{\gamma(Z)}))$ almost surely. \square

III.B.2 Proof of Theorem III.5.2

Proof of Theorem III.5.2. First, note that

$$\begin{aligned} & \left\langle \psi, \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}(Z_1)] \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}^\top(Z_1)], \psi \right\rangle \\ &= (\mathbb{E} [P_1(Z_1) \langle \nabla_\psi g_{\psi^*}(Z_1), \psi \rangle])^2 \\ &= \left(\mathbb{E} \left[\sqrt{P_1(Z_1)} \sqrt{P_1(Z_1)} \langle \nabla_\psi g_{\psi^*}(Z_1), \psi \rangle \right] \right)^2 \\ &\leq \mathbb{E} [P_1(Z_1)] \mathbb{E} \left[P_1(Z_1) \langle \nabla_\psi g_{\psi^*}(Z_1), \psi \rangle^2 \right] \\ &= \left\langle \psi, \mathbb{E} [P_1(Z_1)] \mathbb{E} \left[P_1(Z_1) \nabla_\psi g_{\psi^*}(Z_1) \nabla_\psi g_{\psi^*}(Z_1)^\top \right], \psi \right\rangle. \end{aligned}$$

Thus,

$$\begin{aligned} & \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}(Z_1)] \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}^\top(Z_1)] \preceq \mathbb{E} [P_1(Z_1)] \mathcal{I}_1 \\ \implies & \mathcal{I}_1^{-1} \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}(Z_1)] \mathbb{E} [P_1(Z_1) \nabla_\psi g_{\psi^*}^\top(Z_1)] \mathcal{I}_1^{-1} \preceq \mathbb{E} [P_1(Z_1)] \mathcal{I}_1^{-1} \end{aligned}$$

Since

$$\begin{aligned} \mathbb{E} [P_1(Z_1)] &= \int \frac{1}{1 + \frac{d\mu}{d\gamma}(z)} \frac{d\mu}{d\gamma}(z) \gamma(dz) \\ &\stackrel{(a)}{\leq} \int \min \left(\frac{d\mu}{d\gamma}(z), 1 \right) \gamma(dz) \\ &= 1 - \text{TV}(\mu, \gamma), \end{aligned}$$

where (a) followed from the fact that $r/(1+r) \leq \min(r, 1)$ for any $r \geq 0$, we thus have

$$\mathcal{V}_{\text{NCE}} \succeq \text{TV}(\mu, \gamma) \times \mathcal{I}_1^{-1}.$$

Second, let us note that:

$$\begin{aligned} & \lambda_{\min}(\mathcal{I}_1) \\ & \leq \text{Tr}(\mathcal{I}_1) \\ & = \int \frac{1}{1 + \frac{d\mu}{d\gamma}(z)} \|\nabla_\psi g_{\psi^*}(z)\|^2 \mu(z) dz \\ & \leq \left(\int \frac{1}{\left(1 + \frac{1}{v} \frac{d\mu}{d\gamma}(z)\right)^2} \mu(z) dz \right)^{1/2} \left(\int \|\nabla_\psi g_{\psi^*}(z)\|^4 \mu(z) dz \right)^{1/2} \quad (\text{III.34}) \\ & \leq \left(\int \min\left(\frac{d\mu}{d\gamma}, 1\right)^2 dz \right)^{1/2} \left(\int \|\nabla_\psi g_{\psi^*}(z)\|^4 \mu(z) dz \right)^{1/2} \\ & = (1 - \text{TV}(\gamma, \mu))^{1/2} \mathbb{E} [\|g_{\psi^*}(Z_1)\|^4] \end{aligned}$$

Thus,

$$\begin{aligned} \lim_{N \rightarrow \infty} \sqrt{N} \times \mathbb{E} [\|\psi_{\text{NCE}, N} - \psi^*\|^2] &= \text{Tr}(\mathcal{V}_{\text{NCE}}) \\ &\geq \|\mathcal{V}_{\text{NCE}}\| = \text{TV}(\mu, \gamma) \times \|\mathcal{I}_1^{-1}\| = \frac{\text{TV}(\mu, \gamma)}{\lambda_{\min}(\mathcal{I}_1)} \\ &\geq \frac{\text{TV}(\mu, \gamma)}{(1 - \text{TV}(\mu, \gamma))^{1/2} \mathbb{E} [\|\nabla_\psi g_{\psi^*}(Z_1)\|^4]^{1/2}} \end{aligned}$$

□

III.C Proofs for Neural Ratio Estimation

III.C.1 Auxiliary Lemmas for NRE

Below, we formalize the link between NRE and Noise Contrastive Estimation (NCE). This link will allow us to use the theoretical results available for NCE to study the efficiency of NRE.

Lemma III.C.1. *It holds that*

$$\begin{aligned}\bar{J}_{\text{NRE}}(\psi) &= \bar{J}_{\text{NCE}, p(x, \theta), p_{\mathcal{X}} \otimes \pi, p_{1,\bullet}}(\psi) \\ &= \mathbb{E} \left[\bar{J}_{\text{NCE}, p(\cdot|X), \pi, p_{2,X}, \bullet}(\psi) \mid X_1 \right] \\ &= \mathbb{E} \left[\bar{J}_{\text{NCE}, p(\cdot|\theta), p_{\mathcal{X}}, p_{3,\theta}, \bullet}(\psi) \mid \theta_1 \right]\end{aligned}$$

where $p_{1,\psi}(x, \theta) := \exp(g_\psi(x, \theta)) \times (\pi \otimes p_{\mathcal{X}})$, $p_{2,x,\psi}(\theta) := \exp(g_\psi(x, \theta)) \times \pi$, $p_{3,\theta,\psi} := \exp(g_\psi(x, \theta)) \times p_{\mathcal{X}}$.

Proof. The first equality holds by identification of the quantities involved in the NCE objective function with the ones of the NRE objective function. In particular, by construction, we have

$$\log \frac{dp_{1,\psi}}{d(\pi \otimes p_{\mathcal{X}})}(x, \theta) = g_\psi(x, \theta), \quad \forall x, \theta, \psi.$$

The second equality is obtained by the law of total expectations:

$$\bar{J}_{\text{NRE}}(\psi) := \mathbb{E} \left[\mathbb{E} \left[\log(h(g_\psi(X_1, \theta_1))) \mid X_1 \right] + \mathbb{E} \left[\log((1 - h(g_\psi(X_1, \theta_2))) \mid X_1) \right] \right]$$

noting that

$$\log \frac{dp_{2,x,\psi}}{d\pi}(\theta) = g_\psi(x, \theta), \quad \forall x, \theta, \psi.$$

The third equality is obtained similarly. \square

Proof of Lemma III.3.1

Proof of Lemma III.3.1. Let ψ^* such that $p_{\psi^*}(\theta = \bullet | X_1) = p(\theta = \bullet | X_1)$ almost surely. Let $\psi \in \Psi$. Then, by Lemma III.B.4,

$$\bar{J}_{\text{NCE}, p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1), \pi, p_{2,X_1}, \bullet}(\psi^*) \geq \bar{J}_{\text{NCE}, p(\cdot | X_1), \pi, p_{2,X_1}, \bullet}(\psi) \quad \text{almost surely}$$

by noting that $p_\psi(\theta | x) = p_{2,\psi}(\theta | x)$ for all x, θ, ψ . Thus,

$$\begin{aligned}\bar{J}_{\text{NRE}}(\psi^*) &= \mathbb{E} \bar{J}_{\text{NCE}, p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1), \pi, p_{2,X_1}, \bullet}(\psi^*) \geq \mathbb{E} \bar{J}_{\text{NCE}, p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1), \pi, p_{2,X_1}, \bullet}(\psi) \\ &= \bar{J}_{\text{NRE}}(\psi)\end{aligned}$$

And the inequality is strict if $\mathbb{P} \left[\exp(g_\psi(X_1, \theta_1 = \bullet)) \neq \frac{dp_{\Theta|\mathcal{X}}}{d\pi}(X_1, \theta_1) \right] > 0$. \square

Lemma III.C.2. *Assume A2. Then, it holds that, almost surely,*

$$\begin{aligned}\mathbb{E} \left[P(X_1, \theta_1) \nabla_\psi g_{\psi^*}(X_1, \theta_1) \mid \theta_1 \right] + \mathbb{E} \left[(1 - P(X_2, \theta_1)) \nabla_\psi g_{\psi^*}(X_2, \theta_1) \mid \theta_1 \right] &= 0 \\ \mathbb{E} \left[P(X_1, \theta_1) \nabla_\psi g_{\psi^*}(X_1, \theta_1) \mid X_1 \right] + \mathbb{E} \left[(1 - P(X_1, \theta_2)) \nabla_\psi g_{\psi^*}(X_1, \theta_2) \mid X_1 \right] &= 0.\end{aligned}$$

Proof. Assume the first equation does not hold. Let

$$\begin{aligned}A := \left\{ \theta_1 : \mathbb{E} \left[P(X_1, \theta_1) \nabla_\psi g_{\psi^*}(X_1, \theta_1) \mid \theta_1 \right] \right. \\ \left. + \mathbb{E} \left[(1 - P(X_2, \theta_1)) \nabla_\psi g_{\psi^*}(X_2, \theta_1) \mid \theta_1 \right] = 0 \right\} = 0\end{aligned}$$

which must have positive probability. By Lemma III.B.4, for all $\theta \in A$, we must have $p_{2,\theta,\psi^*} \neq p_{\mathcal{X}|\Theta}(x \in \cdot \mid \theta)$, which must imply there exists some $B(\theta)$ such that

$$\begin{aligned}\int \mathbb{I}_{\{x \in B(\theta)\}} p_{2,\theta,\psi^*}(dx, \theta) &\neq \int \mathbb{I}_{x \in B(\theta)} p_{\Theta|\mathcal{X}}(dx \mid \theta) \\ \implies \iint \mathbb{I}_{\{\theta \in A\}} \mathbb{I}_{\{x \in B(\theta)\}} p_{1,\theta,\psi^*}(dx, \theta) &\neq \iint \mathbb{I}_{\{\theta \in A\}} \mathbb{I}_{\{x \in B(\theta)\}} p_{\mathcal{X},\Theta}(dx, d\theta) \\ \implies \iint \mathbb{I}_{\{(\theta,x) \in C\}} p_{1,\theta,\psi^*}(dx, d\theta) &\neq \iint \mathbb{I}_{\{(x,\theta) \in C\}} p_{\mathcal{X},\Theta}(dx d\theta)\end{aligned}$$

where we noted $C := \{(\theta, x) : \theta \in A, x \in B(\theta)\}$. This in turns implies that on such C , we must have $\exp(g_{\psi^*}(x, \theta)) \neq \frac{dp_{\mathcal{X},\Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta)$, which is forbidden by assumption A2. The other equation is obtained similarly. \square

III.C.2 Proof of Corollary III.4.1

We begin by restating the proposition using the notations of the appendix.

Corollary III.C.3 (Restatement of Proposition III.4.1). *Assume that the triplet $(p_\bullet, (X_1, \theta_1), (X_1, \theta_2))$, where p_\bullet is given in Equation III.2 satisfies A1 and A3. Assume moreover that there exists a ψ^* such that $p_{\psi^*}(\theta \in \bullet \mid X_1) = p_{\Theta|\mathcal{X}}(\theta \in \bullet \mid X_1)$ almost surely (e.g. A2 for almost all settings of X). Finally, assume that J_{NRE} satisfies A4. Then it holds that*

- (i) $\lim_{N \rightarrow \infty} \mathbb{P} \left[(\nabla_\psi J_{\text{NRE}})^{-1}(\{0\}) \neq \emptyset \right] = 1$
- (ii) $\psi_{\text{NRE}, N} \xrightarrow{P} \psi^*$.

Proof of Corollary III.4.1. Let

$$p_{1,\psi} := \exp(g_\psi(x, \theta)) \times d(\pi \times p_{\mathcal{X}}).$$

The following facts on p_1 hold:

- We have

$$\nabla_\psi^{(k)} \log \frac{dp_\psi}{dp_{\mathcal{X}}}(x, \theta) = \nabla_\psi^{(k)} \log \frac{dp_{1,\psi}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta)$$

and as $(p_\bullet, (X_1, \theta_1), (X_1, \theta_2))$ satisfies **AA1**, **AA3**, so does the triplet $(p_{1,\bullet}, (X_1, \theta_1), (X_1, \theta_2))$

- $\frac{dp_{1,\psi^*}}{dp_{\mathcal{X}} \otimes \pi}(X_1, \theta_1) = \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(x, \theta)(X_1, \theta_1)$ almost surely, as by assumption, $p_{\psi^*}(\theta_1 = \bullet | X_1) = p_{\Theta|\mathcal{X}}(\theta \in \bullet | X_1)$ almost surely. Thus, $(p_{1,\bullet}, (X_1, \theta_1), (X_1, \theta_2))$ satisfies **AA2**.
- $g_\psi(x, \theta) = g_{\psi, 1, p, p_\psi(\theta \in \bullet | x)}(\theta) = g_{\psi, 1, p_{\mathcal{X}} \otimes \pi, p_{1,\bullet}}(x, \theta)$

Finally, by Lemma III.C.1, it holds that

$$\mathbb{E}[J_{\text{NRE}}(\psi)](\psi) = \bar{J}_{\text{NCE}, p_{\mathcal{X}, \Theta}, p_{\mathcal{X}} \otimes \pi, p_{1,\bullet}}(\psi)$$

Similarly, as $g_\psi(X_1, \theta_1)$ and $g_\psi(X_1, \theta_2)$ is almost surely thrice differentiable, J_{NRE} are almost surely thrice differentiable. Following Lemma III.D.1, direct differentiation produces

$$\begin{aligned} \nabla_\psi^{(k)} J_{\text{NRE}}(\psi) &= \frac{1}{N} \sum_{i=1}^N \nabla_\psi^{(k)} f_\psi(1, (X_i, \theta_i)) - \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} \nabla_\psi^{(k)} f_\psi(-1, (X_i, \theta_j)) \\ &= \frac{1}{N} \sum_{i=1}^N \nabla_\psi^{(k)} f_\psi(1, (X_i, \theta_i)) \\ &\quad - \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} (\nabla_\psi^{(k)} f_\psi(-1, (X_i, \theta_j)) + \nabla_\psi^{(k)} f_\psi(-1, (X_j, \theta_i))) \end{aligned} \tag{III.35}$$

where $f_\psi(y, (x, \theta)) := -\text{sp}(-yg_\psi(x, \theta))$. As $g_\psi(x, \theta) = g_{\psi, 1, p_{\mathcal{X}} \otimes \pi, p_{1,\psi}}(x, \theta)$, by the

strong law of large number for U-statistics [108], it holds that

$$\nabla_{\psi}^{(k)} J_{\text{NRE}}(\psi) \rightarrow \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}, p(x, \theta), p_{\mathcal{X}} \otimes \pi, p_{1,\bullet}}(\psi) \quad \text{almost surely}$$

implying in particular, $\nabla_{\psi}^{(k)} J_{\text{NRE}}(\psi) \xrightarrow{p} \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}, p(x, \theta), p_{\mathcal{X}} \otimes \pi, p_{1,\bullet}}(\psi)$. Moreover, as **AA1** holds, using the notations of Equation III.50, it holds that

$$\|\nabla_{\psi}^{(3)} \bar{J}_{\text{NCE}}(\tilde{\psi}_1)\| \leq (1 + v) \times \mathbb{E}[a_3(X_i, \theta_i)] < +\infty, \quad i \in \{1, 2\}$$

Moreover, we have using again the law of large numbers for U-statistics [108].

$$\begin{aligned} \|\nabla_{\psi}^{(3)} J_{\text{NRE}}(\tilde{\psi}_2)\| &\leq \frac{1}{N} \sum_{i=1}^N a_3(1, (X_i, \theta_i)) + \sum_{1 \leq i \neq j \leq N} a_3(-1, (X_i, \theta_j)) \\ &= \mathbb{E}[a_3(X_1, \theta_1) + a_3(X_1, \theta_2)] + o_p(1) = O_p(1). \end{aligned} \quad (\text{III.36})$$

All in all, the tuple $(p_{\mathcal{X}, \Theta}, p_{\mathcal{X}} \otimes \pi, p_1, J_{\text{NRE}}(\psi), \bar{J}_{\text{NRE}}(\psi))$ satisfies **AA1**, **AA2**, **AA3** and **AA4** for $v = 1$, as well as the conditions of Theorem III.A.1. Thus, we can invoke Theorem III.A.1 to conclude. \square

Proof of Corollary III.4.1, incomplete case

Proof of Corollary III.4.1, incomplete case. Similarly, we have

$$\nabla_{\psi}^{(k)} J_{\text{INRE}}(\psi) = \frac{1}{r} \sum_{i=1}^N \nabla_{\psi}^{(k)} f_{\psi}(1, (X_i, \theta_i)) - \frac{1}{r} \sum_{(i,j) \in \mathcal{D}} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \quad (\text{III.37})$$

Now, we have

$$\begin{aligned} &\mathbb{V} \left[\frac{1}{r} \sum_{(i,j) \in \mathcal{D}} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \right] \\ &= \mathbb{V} \left[\mathbb{E} \left[\frac{1}{r} \sum_{(i,j) \in \mathcal{D}} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \middle| \{X_i, \theta_i\}_{i=1}^N \right] \right] \\ &\quad + \mathbb{E} \left[\mathbb{V} \left[\frac{1}{r} \sum_{(i,j) \in \mathcal{D}} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \middle| \{X_i, \theta_i\}_{i=1}^N \right] \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{V} \left[\frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \right] \\
&\quad + \frac{p(1-p)}{r} \mathbb{E} \left[\nabla_{\psi}^{(k)} f_{\psi}(-1, (X_1, \theta_2))^2 \right] \\
&= \mathbb{V} \left[\frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_i, \theta_j)) \right] \\
&\quad + \frac{p(1-p)}{r} \mathbb{E} \left[\mathbb{V} \left[\frac{1}{N(N-1)} \sum_{i \leq i \neq j \leq N} (Z_{ij} - p) \nabla_{\psi}^{(k)} f_{\psi}(-1, (X_1, \theta_2))^2 \right] \right] \\
&\xrightarrow[N \rightarrow \infty]{} 0,
\end{aligned}$$

where we denoted $Z_{ij} := \mathbb{I}_{\{(i,j) \in \mathcal{D}\}} \sim \mathcal{B}\left(0, \frac{r}{N(N-1)}\right)$ the Bernoulli random variable independent of the data indicating whether the pair (i, j) is in \mathcal{D} , and where we noted $p = r/(N(N-1))$. Thus, we have that $\nabla_{\psi}^{(k)} J_{\text{INRE}}(\psi) \xrightarrow{p} \nabla_{\psi}^{(k)} \bar{J}_{\text{NCE}, p(x, \theta), p(x)\pi(\theta), p_{1,\bullet}}(\psi)$. Carrying out a similar reasoning to handle the third-order derivative, allows to conclude. \square

III.C.3 Proof of Corollary III.4.2

Next, we will derive asymptotic properties of the NRE estimator $\psi_{\text{NRE}, N}$ and $\psi_{\text{INRE}, N}$. To do so, we first derive the first and second-order moment of $\nabla_{\psi} J_{\text{NRE}}(\psi^*)$. Define

$$\begin{aligned}
A(x_1, \theta_2) &:= (1 - P(x_1, \theta_2)) \nabla_{\psi} g_{\psi^*}(x_1, \theta_2) \\
B(x_1, \theta_1) &:= P(x_1, \theta_1) \nabla_{\psi} g_{\psi^*}(x_1, \theta_1)
\end{aligned} \tag{III.38}$$

In the following, we will note, to simplify notations, $A(X_1) := \mathbb{E}[A(X_1, \theta_2) | X_1]$, $A(\theta_1) := \mathbb{E}[A(X_2, \theta_1) | \theta_1]$, $A := \mathbb{E}[A(X_2, \theta_1)]$, and similarly for $B(X_1, \theta_1)$. By construction, we have $\nabla_{\psi} J_{\text{NRE}}(\psi^*) = \frac{1}{N} \sum_{i=1}^N A(X_i, \theta_i) + \frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} B(X_i, \theta_j)$. Let (U, V) be the random vector defined by:

$$\begin{aligned}
U &= \frac{1}{N} \times \sum_{i=1}^N u_1(X_i, \theta_i), & u_1(x_1, \theta_1) &:= B(x_1, \theta_1) \\
V &= \binom{N}{2}^{-1} \times \sum_{1 \leq i < j \leq N} u_2((X_i, \theta_i), (X_j, \theta_j)) & u_2((x_1, \theta_1), (x_2, \theta_2)) &:= \frac{1}{2}(A(x_1, \theta_2) + A(x_2, \theta_1))
\end{aligned} \tag{III.39}$$

By construction again, it holds that

$$\begin{aligned}\text{Cov} [\nabla_\psi J_{2,\text{NRE}}(\psi)] &= \text{Cov}[U + V] = \text{Cov} \left[M \begin{bmatrix} U \\ V \end{bmatrix} \right] \\ &= M \text{Cov} \left[\begin{bmatrix} U \\ V \end{bmatrix} \right] M^\top, \quad M = [I_d \ I_d]\end{aligned}$$

where I_d is the identity matrix with d rows and columns. We now study $\text{Cov} \left[\begin{bmatrix} U \\ V \end{bmatrix} \right]$. The vector (U, V) is a multivariate U-statistic. We derive its variance and asymptotic distribution, thanks to which we will derive the asymptotic distribution of $\nabla_\psi J_{\text{NRE}}(\psi^*)$, and ultimately of $\psi_{\text{NRE},N} - \psi^*$.

Lemma III.C.4. *It holds that*

$$\sqrt{N} \times (U, V) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \Sigma := \begin{bmatrix} \zeta_{1,11} & \zeta_{1,12} \\ \zeta_{1,12}^\top & \zeta_{1,22} \end{bmatrix}$$

where

$$\begin{aligned}\zeta_{1,11} &= \mathbb{E} [B(X_1, \theta_1)B(X_1, \theta_1)^\top] - \mathbb{E}[B(X_1, \theta_1)]\mathbb{E}[B(X_1, \theta_1)]^\top \\ \zeta_{1,22} &= \mathbb{E} [A(X_1)A(X_1)^\top] + \mathbb{E} [A(\theta_1)A(\theta_1)^\top] + \mathbb{E} [A(X_1)A(\theta_1)^\top] \\ &\quad + \mathbb{E} [A(\theta_1)A(X_1)^\top] - 4AA^\top \\ \zeta_{1,12} &= 2AA^\top - \mathbb{E} [A(X_1)A(X_1)^\top] - \mathbb{E} [A(\theta_1)A(\theta_1)^\top]\end{aligned}$$

Proof. By [109, Theorem 7.1], we have

$$(U, V) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} \zeta_{1,11} & \zeta_{1,12} \\ \zeta_{1,12}^\top & \zeta_{1,22} \end{bmatrix}$$

and

$$\begin{aligned}\zeta_{1,11} &= \text{Cov}[u_1(X_1, \theta_1), u_1(X_1, \theta_1)] \\ \zeta_{1,22} &= 4 \times \text{Cov}[\mathbb{E}[u_2((X_1, \theta_1), (X_2, \theta_2)) | (X_1, \theta_1)]] \\ \zeta_{1,12} &= 2 \times \text{Cov}[u_1(X_1, \theta_1), \mathbb{E}[u_2((X_1, \theta_2), (X_2, \theta_1)) | (X_1, \theta_1)]]\end{aligned}$$

The final expressions for $\zeta_{1,11}$ and $\zeta_{1,22}$ follow by plugging in the expression of u_1 and u_2 , and marginalizing out the unused variables. For $\zeta_{1,12}$, which equals

$$\begin{aligned}\zeta_{1,12} &= 2 \times \left(\mathbb{E} \left[u_1(X_1, \theta_1) \mathbb{E} [u_2((X_1, \theta_2), (X_2, \theta_1)) | (X_1, \theta_1)]^\top \right] \right. \\ &\quad \left. - \mathbb{E} [u_1(X_1, \theta_1)] \mathbb{E} [u_2((X_1, \theta_2), (X_2, \theta_1))]^\top \right) \\ &= \mathbb{E} \left[B(X_1, \theta_1) (A(X_1) + A(\theta_1))^\top \right] - 2BA^\top \\ &= \left(\mathbb{E} \left[B(X_1) A(X_1)^\top \right] + \mathbb{E} \left[B(\theta_1) A(\theta_1)^\top \right] - 2BA^\top \right)\end{aligned}$$

We additionally use the fact that, at $\psi = \psi^*$, it holds

$$A = -B$$

$$A(X_1) = -B(X_1) \quad \text{a.s.}$$

$$A(\theta_1) = -B(\theta_1) \quad \text{a.s.}$$

leading to the final expression displayed in the Lemma above.

□

Lemma III.C.5. *It holds that*

$$\sqrt{N} \times \nabla_\psi J_{\text{NRE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda_{\text{NRE}})$$

where

$$\Lambda_{\text{NRE}} := \mathbb{E} \left[B(X_1, \theta_1) B(X_1, \theta_1)^\top \right] - \mathbb{E} \left[(B(X_1) - B(\theta_1))(B(X_1) - B(\theta_1))^\top \right] - BB^\top$$

Proof. By Lemma III.C.4, it holds that

$$\sqrt{N} \times (U, V) \xrightarrow{d} \mathcal{N}(0, \Sigma), \quad \Sigma := \begin{bmatrix} \zeta_{1,11} & \zeta_{1,12} \\ \zeta_{1,12}^\top & \zeta_{1,22} \end{bmatrix}$$

As $\nabla_{\psi} J_{\text{NRE}}(\psi^*) = M \begin{bmatrix} U \\ V \end{bmatrix}$, by the continuous mapping theorem, it holds that:

$$\sqrt{N} \nabla_{\psi} J_{\text{NRE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, M \Sigma M^\top) = \mathcal{N}(0, \zeta_{1,11} + \zeta_{1,22} + \zeta_{1,12} + \zeta_{1,21})$$

and

$$\begin{aligned} \zeta_{1,11} + \zeta_{1,22} + \zeta_{1,12} + \zeta_{1,21} &= \mathbb{E} \left[B(X_1, \theta_1) B(X_1, \theta_1)^\top \right] \\ &- \mathbb{E} \left[A(X_1) A(X_1)^\top + A(\theta_1) A(\theta_1)^\top \right] + \mathbb{E} \left[A(X_1) A(\theta_1)^\top + A(\theta_1) A(X_1)^\top \right] - AA^\top \end{aligned} \quad (\text{III.40})$$

or, using the identities linking A and B :

$$\begin{aligned} \zeta_{1,11} + \zeta_{1,22} + \zeta_{1,12} + \zeta_{1,21} &= \mathbb{E} \left[B(X_1, \theta_1) B(X_1, \theta_1)^\top \right] - \mathbb{E} \left[(B(X_1) - B(X_1))(B(X_1) - B(\theta_1))^\top \right] - BB^\top \end{aligned} \quad (\text{III.41})$$

□

We establish a similar Lemma for Incomplete NRE. Recall that incomplete NRE is given by:

$$\begin{aligned} \psi_{N, \text{INRE}} &:= \arg \min_{\psi \in \Psi} J_{\text{INRE}}(\psi) \\ J_{\text{INRE}}(\psi) &:= \frac{1}{N} \sum_{i=1}^N \log(h(g_\psi(X_i, \theta_i))) + \frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} \log(1 - h(g_\psi(X_i, \theta_j))) \end{aligned} \quad (\text{III.42})$$

Lemma III.C.6. *Assume \mathcal{D} is formed of $r(N)$ tuples (i, j) drawn from \mathcal{D}_c with replacement, such that $\lim_{N \rightarrow \infty} \frac{r}{N} \rightarrow \alpha$. Then it holds that*

$$\sqrt{N} \times \nabla_{\psi} J_{\text{INRE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda_{\text{INRE}}).$$

where

$$\begin{aligned}\Lambda_{\text{INRE}} &= \Lambda_{\text{NRE}} + \alpha^{-1} \times \text{Cov}[A(X_1, \theta_2)] \\ &= (1 - \alpha^{-1})\mathcal{I}_{\text{NRE},2} + \alpha^{-1}\mathcal{I}_{\text{NRE},1} - \mathbb{E}\left[(B(X_1) - B(\theta_1))(B(X_1) - B(\theta_1))^\top\right] \\ &\quad - (1 + \alpha^{-1})BB^\top.\end{aligned}$$

Proof. Let us use the notations of the previous Lemma, and denote

$$V_{\text{inc}} = \frac{1}{r} \times \sum_{(i,j) \in \mathcal{D}} u_2((X_i, \theta_i), (X_j, \theta_j))$$

where we recall from Equation III.39 and III.38 that

$$\begin{aligned}u_2((X_i, \theta_i), (X_j, \theta_j)) &= \frac{1}{2}(A(X_i, \theta_j) + A(X_j, \theta_i)) \\ A(x_i, \theta_j) &\coloneqq (1 - P(x_i, \theta_j))\nabla_\psi g_{\psi^*}(x_i, \theta_j)\end{aligned}\quad (\text{III.43})$$

We will show convergence in distribution through Levy's continuity theorem by studying the characteristic function of $\nabla_\psi J_{\text{INRE}}(\psi^*)$. We first note that

$$\begin{aligned}\nabla_\psi J_{\text{INRE}}(\psi^*) &= \nabla_\psi J_{\text{NRE}}(\psi^*) + (J_{\text{INRE}}(\psi^*) - \nabla_\psi J_{\text{NRE}}(\psi^*)) \\ \implies e^{-\langle \xi, \nabla_\psi J_{\text{INRE}}(\psi^*) \rangle} &= e^{-\langle \xi, \nabla_\psi J_{\text{NRE}}(\psi^*) \rangle} e^{-\langle \xi, (\nabla_\psi J_{\text{INRE}}(\psi^*) - \nabla_\psi J_{\text{NRE}}(\psi^*)) \rangle} \\ &= e^{-\langle \xi, \nabla_\psi J_{\text{NRE}}(\psi^*) \rangle} e^{-\langle \xi, (V_{\text{inc}} - V) \rangle}\end{aligned}$$

and thus

$$\begin{aligned}\mathbb{E}\left[e^{-\sqrt{N}\langle \xi, \nabla_\psi J_{\text{INRE}}(\psi^*) \rangle}\right] &= \mathbb{E}\left[e^{-\sqrt{N}\langle \xi, \nabla_\psi J_{\text{NRE}}(\psi^*) \rangle} \mathbb{E}\left[e^{-\sqrt{r}\langle \xi, \sqrt{\frac{N}{r}}(V_{\text{inc}} - V) \rangle} \middle| \{X_i, \theta_i\}_{i=1}^N\right]\right]\end{aligned}\quad (\text{III.44})$$

Let us denote $Z_{ij} := \mathbb{I}_{\{(i,j) \in \mathcal{D}\}} \sim \mathcal{B}\left(0, \frac{r}{N(N-1)}\right)$ the Bernoulli random variable independent from the data indicating whether the pair (i, j) is in \mathcal{D} . By construction, we

have

$$\begin{aligned}
V_{\text{inc}} &= \frac{1}{r} \times \sum_{(i,j) \in \mathcal{D}} Z_{ij} A(X_i, \theta_j) \\
\implies V_{\text{inc}} - \mathbb{E}V &= \frac{1}{r} \times \sum_{(i,j) \in \mathcal{D}_c} Z_{ij} (A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j)) \\
\implies V_{\text{inc}} - V &= V_{\text{inc}} - \mathbb{E}V - (V - \mathbb{E}V) \\
&= \frac{1}{r} \sum_{1 \leq i \neq j \leq N} (Z_{ij} - p)(A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j))
\end{aligned}$$

By Lemma A of [149], we have that

$$\begin{aligned}
\mathbb{E} \left[e^{\sqrt{r} \times \left\langle \xi, \sqrt{\frac{N}{r}} (V_{\text{inc}} - V) \right\rangle} \mid \{X_i, \theta_i\}_{i=1}^N \right] &= \mathbb{E} \left[e^{\sqrt{r} \times \left\langle \left\langle \xi, \sqrt{\frac{N}{r}} (V_{\text{inc}} - V) \right\rangle \right\rangle} \mid \{X_i, \theta_i\}_{i=1}^N \right] \\
&\xrightarrow[r \rightarrow \infty]{} e^{-\frac{1}{2} \frac{\sigma_\xi}{\alpha}}
\end{aligned}$$

where

$$\begin{aligned}
\sigma_\xi &= \lim_{r \rightarrow \infty} \sum_{(i,j) \in \mathcal{D}} \langle \xi, A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j) \rangle^2 \\
&= \lim_{r \rightarrow \infty} \sum_{(i,j) \in \mathcal{D}} \left\langle \xi, (A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j))(A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j))^\top \xi \right\rangle \\
&= \lim_{r \rightarrow \infty} \left\langle \xi, \lim_{r \rightarrow \infty} \sum_{(i,j) \in \mathcal{D}} (A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j))(A(X_i, \theta_j) - \mathbb{E}A(X_i, \theta_j))^\top, \xi \right\rangle \\
&\stackrel{\text{a.s}}{=} \langle \xi, \text{Cov}[A(X_i, \theta_j)] \xi \rangle = \langle \xi, \text{Cov}[A(X_1, \theta_2)] \xi \rangle
\end{aligned}$$

where the last line holds by the strong law of large numbers. Consequently, we have

$$\begin{aligned}
\mathbb{E} \left[e^{-\sqrt{N} \langle \xi, \nabla_{\psi} J_{\text{NRE}}(\psi^*) \rangle} \mathbb{E} \left[e^{-\sqrt{r} \left\langle \xi, \sqrt{\frac{N}{r}} (V_{\text{inc}} - V) \right\rangle} \mid \{X_i, \theta_i\}_{i=1}^N \right] \right] \\
&= \mathbb{E} \left[e^{i\sqrt{N} \langle \xi, \nabla_{\psi} J_{\text{NRE}}(\psi^*) \rangle} \right] e^{-\frac{1}{2\alpha} \langle \xi, \text{Cov}[A(X_1, \theta_2)] \xi \rangle} \\
&= e^{-\frac{1}{2} \left\langle \xi, \left(\Lambda_{\text{NRE}} + \frac{\text{Cov}[A(X_1, \theta_2)]}{\alpha} \right) \xi \right\rangle},
\end{aligned}$$

where, in the last line, we used Lemma III.C.5. The result follows by Levy's continuity theorem. The second formula for Λ_{INRE} follows by plugging the expression of $\text{Cov}[A(X_1, \theta_2)]$ derived in Equation III.32, using $v = 1$.

□

Proof of Proposition III.4.2

Proof of Proposition III.4.2. By Lemma III.C.5, it holds that

$$\sqrt{N} \times \nabla_{\psi} J_{\text{NRE}}(\psi^*) \xrightarrow{d} \mathcal{N}(0, \Lambda_{\text{NRE}})$$

Moreover, by Lemma III.C.1, it holds that

$$\bar{J}_{\text{NRE}}(\psi) = \bar{J}_{\text{NCE}, p(x, \theta), p(x)\pi(\theta), p_{1,\bullet}}(\psi)$$

$$\begin{aligned} \nabla_{\psi}^{(2)} \bar{J}_{\text{NRE}} &= \nabla_{\psi}^{(2)} \bar{J}_{\text{NCE}, p(x, \theta), p(x)\pi(\theta), p_{1,\bullet}}(\psi) \\ &= \mathbb{E} \left[\frac{1}{1 + \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(X_1, \theta_1)} \nabla_{\psi} g_{\psi, 1, p_{\mathcal{X}} \otimes \pi, p_{1,\psi}}(X_1, \theta_1) \right. \\ &\quad \times \left. \nabla_{\psi} g_{\psi, 1, p_{\mathcal{X}} \otimes \pi, p_{1,\psi}}(X_1, \theta_1)^{\top} \right] \\ &= \mathbb{E} \left[\frac{1}{1 + \frac{dp_{\mathcal{X}, \Theta}}{d(p_{\mathcal{X}} \otimes \pi)}(X_1, \theta_1)} \nabla_{\psi} \log \frac{dp_{\psi^*}}{dp}(X_1 | \theta_1) \nabla_{\psi} \log \frac{dp_{\psi^*}}{dp}(X_1, \theta_1)^{\top} \right] \end{aligned}$$

which is full rank by assumption. Thus, Theorem III.A.2 holds, and we have

$$\sqrt{N}(\psi^* - \psi_{\text{NRE}, N}) \xrightarrow{d} \mathcal{N}(0, \nabla_{\psi}^{(2)} \bar{J}_{\text{NRE}}(\psi^*)^{-1} \Lambda_{\text{NRE}} \nabla_{\psi}^{(2)} \bar{J}_{\text{NRE}}(\psi^*)^{-1}).$$

The final result follows by plugging the formula of each term involved, and the same reasoning can be done for INRE. □

III.C.4 Proof of Lemma III.5.3

Proof of III.5.3. First, note that

$$\begin{aligned} \mathbb{E} \left[(B(X_1) - B(\theta_1))(B(X_1) - B(\theta_1))^{\top} \right] \\ \preceq 2 \left(\mathbb{E} \left[B(X_1) B(X_1)^{\top} \right] + \mathbb{E} \left[B(\theta_1) B(\theta_1)^{\top} \right] \right) \quad (\text{III.45}) \end{aligned}$$

Then,

$$\begin{aligned}
& B(X_1)B(X_1)^\top \\
&= \int P_{\text{NRE}}(x, \theta) \nabla_\psi g_{\psi^*}(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \\
&\quad \times \left(\int P_{\text{NRE}}(x, \theta) \nabla_\psi g_{\psi^*}(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \right)^\top \\
&\leq \int P_{\text{NRE}}(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \int P_{\text{NRE}}(x, \theta) \nabla_\psi g_{\psi^*} \nabla_\psi g_{\psi^*}^\top(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta)
\end{aligned}$$

Now

$$\begin{aligned}
\int P_{\text{NRE}}(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) &= \int \frac{1}{1 + \frac{\mathrm{d}p_{\mathcal{X}|\Theta}}{\mathrm{d}p_{\mathcal{X}}}(x, \theta)} p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \\
&\leq \int \min \left(\frac{\mathrm{d}p_{\mathcal{X}|\Theta}}{\mathrm{d}p_{\mathcal{X}}}(x, \theta), 1 \right) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \\
&\leq \int \left(\frac{\mathrm{d}p_{\mathcal{X}|\Theta}}{\mathrm{d}p_{\mathcal{X}}}(x, \theta) \right)^\alpha p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta), \quad \alpha \in (0, 1) \\
&= \int e^{\alpha \log \left(\frac{\mathrm{d}p_{\mathcal{X}|\Theta}}{\mathrm{d}p_{\mathcal{X}}}(x, \theta) \right)} p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \\
&\leq e^{\alpha \int \log \left(\frac{\mathrm{d}p_{\mathcal{X}|\Theta}}{\mathrm{d}p_{\mathcal{X}}}(x, \theta) p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \right)} \\
&= e^{-\alpha \text{KL}(p_{\mathcal{X}|\Theta}(x | \theta), p_{\mathcal{X}})} \\
&= e^{-\alpha d \text{KL}((p_{\mathcal{X}|\Theta})_1(\bullet | \theta), (p_{\mathcal{X}})_1(x))} \\
&\leq e^{-\alpha d \beta}
\end{aligned}$$

Consequently, we have:

$$\begin{aligned}
\text{Tr} \left(B(X_1)B(X_1)^\top \right) &\leq \int e^{-\frac{d\beta}{2}} \left(\int P_{\text{NRE}}(x, \theta) \|\nabla_\psi g_{\psi^*}\|^2 p_{\mathcal{X}|\Theta}(\mathrm{d}x | \theta) \right) \pi(\mathrm{d}\theta) \\
&= e^{-\frac{\beta d}{2}} \mathcal{I}_{\text{NRE}, 1}
\end{aligned}$$

The same reasoning applies to $\mathbb{E}[B(\theta_1)B(\theta_1)^\top]$. \square

III.D Auxiliary Lemmas

III.D.1 Differentiability of the NCE Loss

Notations For any rank-3 tensor $A \in \mathbb{R}^{m \times n \times p}$ and a vector $x \in \mathbb{R}^p$, we note:

$$Ax \in \mathbb{R}^{m \times n}, \quad (Ax)_{ij} = \sum_{k=1}^p A_{ijk}x_k, \quad \forall i \in [m], j \in [n] \quad (\text{III.46})$$

the “batch” matrix vector product of A and x . Unless specified, the default norm $\|x\|$ of some rank- k tensor, $k \in [3]$, is set to be the standard euclidean norm for vectors, the operator norm for matrices

$$\|A\| = \max_{\|x\|=1} \|Ax\|, A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n$$

and the operator norm for rank-3 tensors:

$$\|A\| = \max_{\|x\|=1} \|Ax\|, A \in \mathbb{R}^{m \times n \times p}, x \in \mathbb{R}^p.$$

For some p times differentiable $f \in \Psi \mapsto \mathbb{R}$, $k \in \mathbb{N}$, and some $\psi \in \text{int}(\Psi)$, we define $(\nabla_\psi^{(k)} f(\psi)) \in \overbrace{\mathbb{R}^d \times \cdots \times d}^{k \text{ times}}$, (with the convention $\nabla^{(0)} f = f$) such that $(\nabla_\psi^{(k)} f(\psi))_\alpha = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$, for any p -dimensional multi-index α s.t. $|\alpha| = k$. Note that it holds that $Df = \nabla_\psi f^\top D^{(k)} f = D\nabla_\psi^{(k-1)} f = \nabla_\psi^{(k)} f$, for $1 \leq k \leq p$, where D is the differential operator. We moreover have, for all $0 \leq k \leq p$, $\|D^{(k)} f\| = \|\nabla_\psi^{(k)} f\|$.

Lemma III.D.1. Assume AA1. Define $f_\psi(y, z) := -\text{sp}(-yg_\psi(z))$, such that $\bar{J}_{\text{NCE}} = (1 + v) \times \mathbb{E} f_\psi(Y, Z)$. Then it holds that:

$$\begin{aligned} \nabla_\psi f_\psi(y, z) &= \nabla_\psi g_\psi(z) \times y \times \text{sp}'(-yg_\psi(z)). \\ \nabla_\psi^2 f_\psi(y, z) &= \nabla_\psi^2 g_\psi(z) \times y \times \text{sp}'(-yg_\psi(z)) \\ &\quad - (\nabla_\psi g_\psi(z) \otimes \nabla_\psi g_\psi(z)) \times \text{sp}''(-yg_\psi(z)). \\ \nabla_\psi^3 f_\psi(y, z) &= -\text{sp}''(-yg_\psi(z)) \times (\nabla_\psi \otimes \nabla_\psi^{(2)}) g(z) \\ &\quad + \nabla_\psi^{(3)} g_\psi(z) \times \text{sp}'(-yg_\psi(z)) \\ &\quad + (\nabla_\psi g_\psi(z) \otimes \nabla_\psi g_\psi(z) \otimes \nabla_\psi g_\psi(z)) \times y \times \text{sp}'''(-yg_\psi(z)) \end{aligned} \quad (\text{III.47})$$

where we noted

$$(\nabla_{\psi} \otimes \nabla_{\psi}^{(2)})g(z) := \left((\nabla_{\psi} g_{\psi}(z) \otimes \nabla_{\psi}^2 g_{\psi}(z))^{\top_{1 \leftrightarrow 3}} + (\nabla_{\psi} g_{\psi} \otimes \nabla_{\psi}^2 g_{\psi})^{\top_{1 \leftrightarrow 2}} + (\nabla_{\psi} g_{\psi}(z) \otimes \nabla_{\psi}^2 g_{\psi}(z)) \right) \quad (\text{III.48})$$

Moreover, \bar{J}_{NCE} is thrice-differentiable on $\text{int}(\Psi)$, and we have, for $k \in \{1, 2, 3\}$,

$$\nabla_{\psi}^{(k)} J_{\text{NCE}} = (1 + v) \times \mathbb{E} \left[\nabla_{\psi}^{(k)} f_{\psi}(Y, Z) \right] \quad (\text{III.49})$$

Proof. Equation III.47 follows from direct applications of the chain rule, the fact that $y^2 = 1$. Thus, it holds that:

$$\begin{aligned} \|\nabla_{\psi} f_{\psi}(y, z)\| &\leq l_1(z) := a_1(z) \\ \|\nabla_{\psi}^2 f_{\psi}(y, z)\| &\leq l_1(z)^2 + l_2(z) := a_2(z) \\ \|\nabla_{\psi}^3 f_{\psi}(y, z)\| &\leq l_1(z)^3 + 3l_2 \times l_1(z) + l_3(z) \leq l_1(z)^3 + \frac{3}{2} (l_1(z)^2 + l_2(z)^2) := a_3(z) \end{aligned} \quad (\text{III.50})$$

where, in (ii), (iii), we used Lemma III.D.5, the triangle inequality, the fact that $\|a \otimes b\| = \|a\| \|b\|$, in (iii) we used the inequality $\prod_{b=1}^B a_b \leq \frac{1}{B} \sum_{b=1}^B a_b^B$ and finally, in each inequality, we used AA1. We show the differentiability of J_{NCE} by induction. Note that J_{NCE} is 0 times differentiable. Now, assume that J_{NCE} is k -times differentiable for some $k \in \{0, 1, 2\}$. Let $\psi \in \text{int}(\Psi)$. Define

$$\Delta_{\psi}^{(k)}(y, z; h) := \frac{\nabla_{\psi}^{(k)} f_{\psi+h}(y, z) - \nabla_{\psi}^{(k)} f_{\psi}(y, z) - D \nabla_{\psi}^{(k)} f_{\psi}(y, z) h}{\|h\|}$$

And let $(h_n)_{n \in \mathbb{N}} \in (\Psi - \psi)^{\mathbb{N}}$ such that $\|h_n\| \rightarrow 0$ as $n \rightarrow +\infty$. As f is $(k+1)$ -times differentiable, it holds by the sequential characterization of limits that

$\lim_{\|h\| \rightarrow 0} \Delta(\psi; h) = 0$. Moreover, for all h s.t. $\psi + h \in \Psi$, we have

$$\begin{aligned}\|\Delta_\psi(y, z; h)\| &\leq \frac{\|\nabla_\psi^{(k)} f_{\psi+h}(y, z) - \nabla_\psi^{(k)} f_\psi(y, z)\|}{\|h\|} + \left\| \nabla_\psi^{(k+1)} f_\psi(y, z) \right\| \\ &\leq \frac{\left\| \int_0^1 \mathbf{D}\nabla_\psi^{(k)} f_{\psi+th}(y, z) h dt \right\|}{\|h\|} + a_{k+1}(z) \\ &\leq 2 \times a_{k+1}(z).\end{aligned}$$

As $\mathbb{E}[a_k(Z)] < +\infty$, by the dominated convergence theorem, it follows that:

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{\left\| \nabla_\psi^k \bar{J}_{\text{NCE}}(\psi + h_n) - \nabla_\psi^{(k)} \bar{J}_{\text{NCE}}(\psi) - \mathbb{E} \left[\mathbf{D}\nabla_\psi^{(k)} f_\psi(Y, Z) h_n \right] \right\|}{\|h_n\|} \\ &= \lim_{n \rightarrow \infty} \mathbb{E} \left[\Delta_\psi^{(k)}(Y, Z; h_n) \right] \\ &= \mathbb{E} \left[\lim_{n \rightarrow \infty} \Delta^{(k)}(Y, Z; h_n) \right] = 0.\end{aligned}$$

By the sequential characterization of limits, we have that $\bar{J}_{\text{NCE}}(\psi)$ is $(k+1)$ -times differentiable, with $\nabla_\psi^{(k+1)} \bar{J}_{\text{NCE}}(\psi) = \mathbb{E} \left[\nabla_\psi^{(k+1)} f_\psi(Y, Z) \right]$. Invoking the recursion up to $k = 2$ concludes the proof. \square

Similarly, direct differentiation yield

$$\nabla_\psi^{(k)} J_{\text{NCE}}(\psi) = \frac{1}{N} \left(\sum_{i=1}^N \nabla_\psi^{(k)} f_\psi(1, Z_i) + \sum_{i=1}^M \nabla_\psi^{(k)} f_\psi(-1, \tilde{Z}_i) \right) \quad (\text{III.51})$$

III.D.2 Important Identities of NCE losses

Lemma III.D.2. *For any function f , it holds that:*

$$\mathbb{E} [Y f(Z) \text{sp}'(-Y g_{\psi^*}(Z))] = 0 \quad (\text{III.52})$$

and

$$\mathbb{E} [f(Z_1) \text{sp}'(-g_{\psi^*}(Z_1))^2] + v \mathbb{E} [f(\tilde{Z}_1) \text{sp}'(g_{\psi^*}(\tilde{Z}_1))^2] = \mathbb{E} [\text{sp}'(-g_{\psi^*}(Z_1)) f(Z_1)] \quad (\text{III.53})$$

Proof. For the first equality, we have

$$\begin{aligned} & \mathbb{E} [Y \nabla_{\psi} f(Z) \text{sp}'(Y g_{\psi^*}(Z))] \\ &= \frac{1}{1+v} \mathbb{E} [f(Z_1) \text{sp}'(Y g_{\psi^*}(Z_1))] - \frac{v}{1+v} \mathbb{E} [f(\tilde{Z}_1) \text{sp}'(-g_{\psi^*}(\tilde{Z}_1))] \\ &\stackrel{(a)}{=} \frac{1}{1+v} \left(\int \frac{f(z)1}{1+\frac{1}{v}\frac{d\mu}{d\gamma}} \mu(dz) - \int \frac{f(z)v}{1+v \times \frac{d\gamma}{d\mu}} \gamma(dz) \right) = 0 \end{aligned}$$

where in (a), we used Lemma III.D.5. For the second equality, we have

$$\begin{aligned} & \mathbb{E} [f(Z_1) \text{sp}'(g_{\psi^*}(Z_1))^2] + v \mathbb{E} [f(\tilde{Z}_1) \text{sp}'(-g_{\psi^*}(\tilde{Z}_1))^2] \\ &= \int f(z) \frac{1}{\left(1+\frac{1}{v}\frac{d\mu}{d\gamma}(z)\right)^2} \mu(dz) + \int \frac{f(z)v}{(1+v\frac{d\gamma}{d\mu}(z)^2)} \gamma(dz) \\ &= \int \frac{f(z)(1+\frac{1}{v}\frac{d\mu}{d\gamma}(z))}{\left(1+\frac{1}{v}\frac{d\mu}{d\gamma}\right)^2} \mu(dz) \\ &= \int f(z) \frac{1}{1+\frac{d\mu}{d\gamma}(z)} \mu(dz) \end{aligned}$$

□

Lemma III.D.3. Let A_n be a sequence of random matrices such that $A_n \xrightarrow{p} A$ where A is invertible. Let B_n be a sequence of random matrices, such that $A_n B_n \xrightarrow{d} C$ where C is a random matrix. Then it holds that $B_n \xrightarrow{d} A^{-1}C$.

Proof. We can write:

$$B_n = A_n^{-1}C \times 1_{\{A_n \text{ invertible}\}} + B_n \times 1_{\{A_n \text{ not invertible}\}}$$

We first show that first term converges in distribution to $A^{-1}C$. Let $\epsilon > 0$. As $\text{GL}_d(\mathbb{R})$ is open, we can find some δ such that $\overline{\mathcal{B}}(A, \delta) \subset \text{GL}_d(\mathbb{R})$. Since the inverse function is continuous on that set, there exists a $\delta_1 \in (0, \delta)$ such that $\|A' - A\| \leq \delta_1 \implies$

$\|(A')^{-1} - A^{-1}\| \leq \varepsilon$. Consequently, we have

$$\begin{aligned} & \mathbb{P} [\{\|A_n^{-1} 1_{\{A_n \text{ invertible}\}} - A^{-1}\| \leq \varepsilon\}] \\ & \geq \mathbb{P} [\{\|A_n^{-1} 1_{\{A_n \text{ invertible}\}} - A^{-1}\| \leq \varepsilon\} \cap \{\|A_n - A\| \leq \delta_1\}] \\ & = \mathbb{P} [\{\|A_n - A\| \leq \delta_1\}] \xrightarrow{n \rightarrow \infty} 1 \end{aligned}$$

since as $A_n \xrightarrow{p} A$. By the sandwich rule, it holds that $A_n^{-1} 1_{\{A_n \text{ invertible}\}} \xrightarrow{p} A^{-1}$, and, by Slutsky's theorem, we have

$$A_n^{-1} 1_{\{A_n \text{ invertible}\}} C \xrightarrow{d} A^{-1} C.$$

We now show that the second term converges to 0 in probability: indeed, we have:

$$\mathbb{P} [B_n 1_{\{A_n \text{ not invertible}\}} \leq \varepsilon] \geq \mathbb{P} [\{A_n \text{ invertible}\}] \geq \mathbb{P} [\{\|A_n - A\| \leq \delta_1\}] \xrightarrow{n \rightarrow \infty} 1.$$

By Slutsky's theorem, we thus have that $B_n \xrightarrow{d} A^{-1} C$. \square

Lemma III.D.4. *Let E, F be finite dimensional normed vector spaces, and $f : E \rightarrow F$ be Fréchet differentiable. Let G be a bounded subset of E . Define*

$$\sigma_{\min} = \min_{g \in G} \min_{\|e\|=1} \|\nabla f(e)g\|, \quad \sigma_{\max} = \max_{g \in G} \max_{\|e\|=1} \|\nabla f(e)g\|$$

Then we have, for all $e_1, e_2 \in G$:

$$\sigma_{\min} \|e_1 - e_2\| \leq \|f(e_1) - f(e_2)\| \leq \sigma_{\max} \|e_1 - e_2\|$$

Proof. Let $\varphi : t \in [0, 1] = f(e_1 + t(e_2 - e_1))$.

$$\begin{aligned}
 \|\varphi(1) - \varphi(0)\| &= \left\| \int_0^1 \varphi'(t) dt \right\| \\
 &\leq \int_0^1 \|\varphi'(t)\| dt \\
 &= \int_0^1 \|\nabla f(e_1 + t(e_2 - e_1))(e_2 - e_1)\| dt \\
 &= \|e_2 - e_1\| \int_0^1 \left\| \nabla f(e_1 + t(e_2 - e_1)) \frac{(e_2 - e_1)}{\|e_2 - e_1\|} \right\| dt \\
 &\leq \|e_2 - e_1\| \int_0^1 \sigma_{\max} dt \\
 &= \sigma_{\max} \|e_2 - e_1\|
 \end{aligned}$$

and the lower bound inequality is obtained similarly. \square

Lemma III.D.5. Let $h(x) := \frac{1}{1+\exp(-x)}$, defined as in Section III.3, and recall that $\text{sp}(x) := \log(1 + \exp(x))$. The following holds:

- (i) $\text{sp}'(x) = 1 - h(x)$
- (ii) $\text{sp}''(x) = h(x) - h(x)^2$
- (iii) $\text{sp}'''(x) = h'(x) \tanh(\frac{x}{2})$
- (iv) $h(x) + h(-x) = 1$

Consequently, we have $h^{(k)}(x) \in (-1, 1)$ for $k \in [0, 3]$.

Proof. Omitted. \square

III.E Background on finite-sample bounds for Logistic Regression

Finite-sample bounds for logistic regression are available for various setting of the dimension d and sample size n , and can be used to analyze the behavior of estimators when the dimension d grows with n , such as the traditional high-dimensional regime $d/n \rightarrow c \in (0, 1)$, or milder form of high dimensionality, $n, d \rightarrow \infty, d/n \rightarrow 0$. In addition to n and d , a key parameter dictating estimation bounds in logistic regression is the norm $\|\psi^*\|$ of the linear separator ψ^* , sometimes referred to as the signal-to-

noise ratio. Let us note $D = \text{diam}(\Psi)$, which acts as a uniform upper bound on the parameter norm $\|\psi\|$ for all $\psi \in \Psi$. We list a few regimes of interest

- $N \geq d \times e^{cD}$ for some $c > 0$. This regime is a large sample size regime, in which the obtained bounds resemble known large sample asymptotic results [256].
- $N = O(e^{cD})$. This regime is a “moderate” sample size scenario which may encompass many cases of interest when D is large. In this paper, we construct a set of distributions for which $D = \Omega(d)$, meaning this regime becomes relevant even for $d \approx 10$. Many results are available in such cases:
 - [100] prove a worst-case lower bound of $\Omega((D/N)^{1/2})$ in that regime.
 - [185] show that if $N \geq c \log^4(D) D^8 d \log(1/\delta)$, then w.p $1 - \delta$, the MLE exists and is upper bounded by $D^3 \log(1/\delta) \times \frac{d}{N}$.
 - [139] showed that if $n \geq cD \times d \log n + \log(1/\delta)$, the direction of ψ^* can be well estimated, and the norm of ψ^* can be estimated not too poorly

The foundational work of [100] has established that when $\|X\| \leq R$ almost surely and $N = O(e^{RD})$, the worst-case classification excess risk of the logistic regression estimator will be no better than $\Omega((D/N)^{1/2})$, which is slower than the minimax rate $O(N^{-1})$ [2, 179]. Regarding upper bounds, [139] established bounds on the (misspecified) logistic estimator used on data satisfying the probit model when the covariates satisfy a Gaussian distribution. For high $\|\psi^*\|$ (as a function of n, d), the probability of obtaining an incorrect label is low. In other words, the two populations of samples are “easy” to classify, in the sense that the two populations rarely overlap. In such settings, while estimating the direction of ψ^* (i.e. $\psi^*/\|\psi^*\|$) is possible, estimating $\|\psi^*\|$ is hard: in particular,

- In “low” signal-to-noise scenarios, e.g. $\|\psi^*\| \leq \frac{n}{p \log n}$, bounds on the difference $\|\|\psi_N\| - \|\psi^*\|\|$ typically scale with $\|\psi^*\|^{3/2}$ [36, 139, 185].
- For “high” signal-to-noise scenarios, e.g. $\|\psi^*\| \geq \frac{n}{p \log n}$, bounds only on $\|\psi_N\|$ (not even $\|\|\psi_N\| - \|\psi^*\|\|$) are available [139].

While estimating the parameter direction is good enough to obtain a low classification error (the canonical objective in logistic regression), it is not good enough, by definition, to estimate p_{ψ^*} when it depends on $\|\psi^*\|$, which is the goal of noise-contrastive estimation. The “hard” distributions we construct to prove lower bounds on the variance of the NRE estimators are precisely estimators with high parameter norm $\|\psi^*\|$. In that sense, we thus exploit a defect of logistic regression which was previously documented, and position it in a density estimation setting, where the gold standard to compare against is the MLE estimator of the data distribution. On the other hand, such results ought to be compared with our asymptotic lower bonds established in this paper. In particular, our lower bounds are consistent with the high signal-to-noise regime, where consistency of the parameter norm estimator is not achieved. On the other hand, our bounds are (a prior) not consistent with the low signal-to-noise regime, where consistency is obtaining with (assuming $\|\psi^*\| = O(d)$) a polynomial dependence on d , not an exponential one as in our setting. Such difference does not infirm the results obtain in neither work, but highlight the difference in asymptotic vs. non-asymptotic analysis, and why asymptotic analyses are sometimes not enough.

CHAPTER IV

Near Optimality of Contrastive Divergence Algorithms

This Chapter is based on the following work:

Pierre Glaser, Kevin Han Huang, and Arthur Gretton. Near-optimality of contrastive divergence algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Q74JVgKCP6>

Abstract

We perform a non-asymptotic analysis of the contrastive divergence (CD) algorithm, a training method for unnormalized models. While prior work has established that (for exponential family distributions) the CD iterates asymptotically converge at an $O(n^{-1/3})$ rate to the true parameter of the data distribution, we show, under some regularity assumptions, that CD can achieve the parametric rate $O(n^{-1/2})$. Our analysis provides results for various data batching schemes, including the fully online and minibatch ones. We additionally show that CD can be near-optimal, in the sense that its asymptotic variance is close to the Cramér-Rao lower bound.

IV.1 Introduction

Describing data using probability distributions is a central task in multiple scientific and industrial disciplines [47, 202, 201]. Since the true distribution of the data is generally unknown, such a task requires finding an estimator of the true distribution among a model class that best describes the available data. An estimator can be characterized at multiple levels of granularity: at the highest level lies *consistency* [256], a property which states that as the number of available data points increases, a given estimator will converge to the one best describing the data distribution. At a lower level, a consistent estimator can be further characterized by its convergence rate, a quantity upper-bounding its distance to the true distribution as a function of

the number of samples. A convergence rate can be either *asymptotic*, e.g. hold only in the limit of an infinite sample size, or *non-asymptotic*, in which case the rate also holds for finite sample sizes. In their simplest form, convergence rates are provided in big- O notation, discarding finer grained information such as asymptotically dominated quantities as well as multiplicative constants. These constants play a role in the so-called *asymptotic variance* of the estimator, which is a precise descriptor of an estimator's statistical efficiency. Convergence rates and asymptotic variances have been the subject of extensive research in the statistical literature; in particular, well-known lower bounds exist regarding both the best possible (asymptotic) convergence rate of an estimator and its best possible asymptotic variance. These results set a clear frame of reference to interpret individual convergence rates, and are routinely present in the analysis of modern statistical algorithms such as noise-contrastive estimation [151, 92] or score matching [117, 136, 187].

In this work, we focus on cases where (1) the true data distribution admits a density with respect to some known base measure, and (2) the model class is parametrized by a finite-dimensional parameter. In this setting, provided that the true distribution belongs to the model class, a celebrated result in statistical estimation states that the model maximizing the average log-likelihood both achieves the best possible asymptotic convergence rate (called the *parametric rate*) and the best possible asymptotic variance, called the Cramér-Rao bound (see, e.g. [35]). While this result shows that Maximum Likelihood Estimators (MLE) are asymptotically optimal, fitting them is complicated by computational hurdles when using models with intractable normalizing constants. Such *unnormalized models* are common in the Machine Learning literature due to their high flexibility [146, 62]; their weakness however lies in the fact that expectations under these models have no unbiased approximation. For this reason, popular approximation algorithms such as unbiased gradient-based stochastic optimization of the empirical log-likelihood cannot *a priori* be used, as the gradient of the normalizing constant is given by an expectation under the model distribution.

The Contrastive Divergence (CD) algorithm [105] is a popular approach that cir-

cumvents this issue by using a Markov Chain Monte Carlo (MCMC) algorithm to approximate the gradient of the log-likelihood. Unnormalized models trained with Contrastive Divergence have been shown to reach competitive performance in high-dimensional tasks such as image [183, 182, 63], text [199], and protein modeling [137, 259], or neuroscience [249]. A consistency analysis of the Contrastive Divergence algorithm is delicate, however: indeed, the *optimization error* e.g. the difference between the estimate returned by CD and the MLE, is likely to be non-negligible as compared with *statistical error* – the distance between the MLE and the true distribution – and thus cannot be discarded, as often done when analyzing estimators that minimize tractable objectives [136, 151]. Recent work [121] elegantly established asymptotic $O(n^{-1/3})$ –consistency of the CD estimator for unnormalized exponential families when using only a *finite* number of MCMC steps. Key to their argument is the fact that the bias of the CD gradient estimate decreases as iterates approach the data distribution. However, as noted by the authors, their work left open whether and under what conditions CD might achieve $O(n^{-1/2})$ –consistency.

Contributions In this work, we answer this question by providing a non-asymptotic analysis of the CD algorithm for unnormalized exponential families. While existing convergence bounds [121] were derived for the “full batch” setting, where the CD gradient is estimated using the full dataset at each iteration, our analysis covers both the online setting (where data points are processed one at a time without replacement), and the offline setting with multiple data reuse strategies (including full batch).

In the online case (Section IV.3), we show, under a restricted set of assumptions compared to Jiang et al. [121], that the CD iterates can converge to the true distribution at the parametric $O(n^{-1/2})$ rate. Our analysis reveals that CD contains two sources of approximation: a bias term, and a variance term. These sources are almost independent of each other, in the sense that decreasing the bias by increasing the number of MCMC steps will not decrease the variance. The impact of these two sources of approximation transparently propagates in our resulting bounds: in particular, as the bias of the CD algorithm goes to 0, our bounds recover well-known results in online stochastic optimization [169]. Finally, we study the asymptotic

variance of an estimator obtained by averaging the CD iterates, a classic acceleration technique in stochastic optimization [196]. We show that provided that the number of steps m is sufficiently large, the *asymptotic* variance of this estimator matches (up to a factor 4) the Cramér-Rao bound.

Next, we study the offline setting (Section IV.4), where the CD gradient is estimated by reusing (potentially random) subsets of a finite dataset. We show that a similar result to the online setup holds, up to an additional correlation term that arises from data reuse, and present several approaches to control this term. We improve over the results of [121] by showing a non-asymptotic and near-parametric rate at $O((\log n)^{1/2} n^{-1/2})$ under their conditions, and also illustrate how different rates can be obtained under a variety of conditions. Our results also show an interesting trade-off between the effect of initialization and the statistical error as a function of batch size.

In summary, we establish the near-optimality of a variety of Contrastive Divergence algorithms for unnormalized exponential families in the so called “long-run” regime, where the number of MCMC steps is high enough to ensure that the CD gradient bias is sufficiently offset by the convexity of the negative log-likelihood.

IV.2 Contrastive Divergence in Unnormalized Exponential Families

Unnormalized Exponential Families Exponential families (EF) [32, 264] form a well-studied class of probability distributions, given by

$$p_\psi(dx) := e^{\psi^\top \phi(x) - \log Z(\psi)} c(dx), \quad Z(\psi) := \int_{\mathcal{X}} e^{\psi^\top \phi(x)} c(dx). \quad (\text{IV.1})$$

Here, $\mathcal{X} \ni x$ is the *data* or *sample space*, which we set to be a subset of \mathbb{R}^d for some $d \in \mathbb{N}^*$, although our results are readily extendable to more general measurable spaces. c is a measure on \mathcal{X} called the *base* or *carrier* measure. When $\mathcal{X} \subseteq \mathbb{R}^d$, c is often set to be the corresponding Lebesgue measure. $\psi \in \Psi \subseteq \mathbb{R}^p$ is a finite-dimensional parameter called the *natural parameter*, and $\phi : \mathbb{R}^d \mapsto \mathbb{R}^p$ is a function called the *sufficient statistics*, which, alongside with the base measure,

fully describes an exponential family. Finally, $\log Z(\psi)$, the *log-normalizing* (or *cumulant*) function, is a quantity ensuring that p_ψ integrates to 1 over \mathcal{X} . Crucially, we will not assume that $\log Z(\psi)$ admits a closed form expression for all ψ . The latter fact provides the practitioner with a great deal of flexibility in designing the model class: indeed, the only requirement that should be satisfied prior to performing statistical estimation is to have $Z(\psi) < +\infty$ for all ψ , something that can be readily verified and is often the case in practice. The drawback of unnormalized EFs is the fact that sampling (and thus approximating expectations under the model) cannot usually be performed in an unbiased manner. Instead, inference in unnormalized EFs is often performed using tools from the Bayesian Inference literature, such as MCMC [75]. Unnormalized EFs belong to the larger class of *unnormalized models* [148, 228, 117, 92], of the form $e^{-E_\psi(x)-\log Z(\psi)}c(dx)$, $Z(\psi) = \int e^{-E_\psi(x)}c(dx)$, for some parametrized function $E_\psi : \mathbb{R}^d \mapsto \mathbb{R}$ referred to as the *energy*. Unnormalized models thus take the flexibility of unnormalized EFs one step further by allowing the (negative) unnormalized log-density to be an arbitrary function E_ψ of x and ψ , instead of requiring a linear dependence on ψ as in Equation V.45. We focus in this work on unnormalized EFs due to the multiple computational benefits they provide, as explained in the next section, but we believe that extending our analysis to more general unnormalized models is an interesting avenue for future work.

Statistical Estimation in Unnormalized Exponential Families using Contrastive Divergence We now review the Contrastive Divergence algorithm, an algorithm used to fit unnormalized models, and our main object of study in this work. The general setting is the following: we assume access to n i.i.d. samples (X_1, \dots, X_n) drawn from some unknown distribution p^* , which we assume belongs to \mathcal{P}_ψ , e.g. $p^* = p_{\psi^*}$ for some $\psi^* \in \Psi$. Given these samples, we aim to perform *statistical estimation*, e.g. find a parameter ψ_n within Ψ that should approach ψ^* as n grows.

The starting point of the Contrastive Divergence algorithm is the unfortunate realization that Maximum Likelihood Estimation, which corresponds to minimizing the cross-entropy $\mathcal{L}(\psi) := -\mathbb{E}_{p_n} \log dp_\psi/dc$ between the model p_ψ and the empirical data distribution $p_n := 1/n \sum_{i=1}^n \delta_{X_i}$, cannot be performed using exact (possibly

stochastic) gradient-based optimization, as the gradient $\nabla_\psi \mathcal{L}(\psi)$ of \mathcal{L} with respect to the parameter ψ contains an expectation under the model distribution p_ψ . Indeed, the cross entropy and its gradient are given by

$$\begin{cases} \mathcal{L}(\psi) &= -\frac{1}{n} \sum_{i=1}^n \phi(X_i)^\top \psi + \log Z(\psi) \\ \nabla_\psi \mathcal{L}(\psi) &= -\frac{1}{n} \sum_{i=1}^n \phi(X_i) + \mathbb{E}[\phi(X^\psi)], \quad X^\psi \sim p_\psi. \end{cases} \quad (\text{IV.2})$$

The second line follows from the well known identity $\nabla_\psi \log Z(\psi) := \mathbb{E}[\phi(X^\psi)]$; we refer to [264, Proposition 3.1] for a proof. The Contrastive Divergence algorithm circumvents this issue by running approximate stochastic gradient descent (SGD) on \mathcal{L} , where the intractable expectation in $\nabla_\psi \log Z$ is estimated using an MCMC algorithm initialized at the empirical data distribution. In more details, given a number of epochs T , a sequence of data batches $B_{t,j}$ of size B (e.g. $B_{t,j} \in \llbracket 1, n \rrbracket^B, 1 \leq t \leq T, 1 \leq j \leq N \lceil n/B \rceil$), and a family of *Markov kernels* $\{k_\psi, \psi \in \Psi\}$ each with invariant distribution p_ψ , at the j^{th} minibatch of epoch t , $\nabla_\psi \log Z(\psi_{t,j-1})$ is approximated by $\frac{1}{B} \sum_{i \in B_{t,j}} \phi(\tilde{X}_i^m)$, where \tilde{X}_i^m is produced by running the recursion $\tilde{X}_i^k \sim k_{\psi_t}(\tilde{X}_i^{k-1}, \cdot)$, $\tilde{X}_i^0 = X_i$ up to $k = m$. Throughout the paper, we will refer to the conditional distribution of \tilde{X}_i^m given X_i as $k_\psi^m(X_i, \cdot)$. The resulting gradient estimate arising from combining this approximation with the other (tractable) sum over the data samples present in $\nabla_\psi \mathcal{L}(\psi)$, which we refer to as the *CD gradient* and denote as h_t , is thus

$$h_{t,j} := \frac{1}{B} \sum_{i \in B_{t,j}} \phi(X_i) - \frac{1}{B} \sum_{i \in B_{t,j}} \phi(\tilde{X}_i^m) = \frac{1}{B} \sum_{i \in B_{t,j}} (\phi(X_i) - \phi(\tilde{X}_i^m)). \quad (\text{IV.3})$$

Key to the behavior and analysis of the CD algorithm is the strategy employed to generate minibatches $B_{t,j}$. The case where $T = 1$, $B = 1$, and $B_{1,j} = \{j\}$ will be referred to as *online* CD, while the variant where $T > 1$, and each batch $B_{t,j}$ draws B indices (with or without replacement) from $\llbracket 1, n \rrbracket$ will be referred to as *offline* CD. In online CD, each data point is present in one and one batch only, while in offline CD, data points are reused across batches. From a statistical perspective, we will see that online CD can be analyzed in a remarkably simple way, while offline CD

introduces additional correlations that require care to be controlled. Both settings come with their advantages and drawbacks, as we will see in the next section. The CD algorithms we study will employ decreasing step size schedules $(\eta_t)_{t \geq 0}$ of the form $\eta_t = Ct^{-\beta}$, where $C > 0$ is the initial leaning rate and $\beta \in [0, 1]$. We lay out online CD and offline CD in Algorithms 1 and 2. Note that our algorithms include a projection step on the parameter space Ψ to account for the case where Ψ is compact. In the case $\Psi = \mathbb{R}^p$, this step can be omitted. Next we depart from the setting of [121] and start by analyzing online CD.

Algorithm 1 Online CD

Input: $(X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} p_{\psi^*}$
Parameters: Model class $\{p_\psi, \psi \in \Psi\}$, Markov kernels $\{k_\psi, \psi \in \Psi\}$, number of MCMC steps m , learning rate schedule $\eta_t := Ct^{-\beta}$, $\beta \in [0, 1]$, $C > 0$, initial parameter ψ_0
for $t = 1, \dots, n$ **do**
 //Approx. sample from $p_{\psi_{t-1}}$
 $\tilde{X}_t^m \sim k_{\psi_{t-1}}^m(X_t, \cdot)$
 $h_t := \phi(X_t) - \phi(\tilde{X}_t^m)$
 $\psi_t \leftarrow \psi_{t-1} - \eta_t h_t$
 $\psi_t \leftarrow \text{Proj}_\Psi(\psi_t)$
end for
Return ψ_n

Algorithm 2 Offline CD

Input: $(X_1, \dots, X_n) \stackrel{\text{i.i.d.}}{\sim} p_{\psi^*}$
Parameters: Same as Algorithm 1, plus number of epochs T , batch size B , batching schedule $B_{t,j}$, initial parameter $\psi_{1,0}$
for $t = 1, \dots, T$ **do**
 for $j = 1, \dots, \lceil n/B \rceil$ **do**
 $\tilde{X}_{t,i,j}^m \sim k_{\psi_{t-1}}^m(X_i, \cdot)$ **for** i in $B_{t,j}$
 $h_{t,j} := \frac{1}{B} \sum_{i \in B_{t,j}} (\phi(X_i) - \phi(\tilde{X}_{t,i,j}^m))$
 $\psi_{t,j} \leftarrow \psi_{t,j-1} - \eta_t h_{t,j}$
 $\psi_{t,j} \leftarrow \text{Proj}_\Psi(\psi_{t,j})$
 end for
 $\psi_{t+1,0} \leftarrow \psi_{t,\lceil n/B \rceil}$
end for
Return $\psi_{T,\lceil n/B \rceil}$

IV.3 Non-asymptotic analysis of Online CD

IV.3.1 Preliminaries and Assumptions

Recall that the chi-squared divergence between two probability measures p and q is defined as: $\chi^2(p, q) := \int (\frac{dp}{dq}(x) - 1)^2 q(dx)$ if $p \ll q$, and $+\infty$ otherwise. Here, $p \ll q$ denotes that p is absolutely continuous with respect to q and dp/dq is the Radon-Nikodym derivative [61] of p with respect to q . Let $L^2(p_\psi)$ be the space of square-integrable functions with respect to p_ψ . For a function $f \in L^2(p_\psi)$, we define

$$\alpha(f, \psi) = \frac{\left(\int \left(\int (f - \mathbb{E}[f(X^\psi)])(y) k_\psi(x, dy) \right)^2 p_\psi(dx) \right)^{1/2}}{\left(\int (f - \mathbb{E}[f(X^\psi)])(x)^2 p_\psi(dx) \right)^{1/2}} \quad (\text{IV.4})$$

which is a measure of how quick a Markov chain with kernel k_ψ mixes, relative to the function f [152]. With these definitions in hand, we now state the assumptions required by our analysis of online CD. These assumptions form a strict subset of the

assumptions considered in prior work [121], which required additional regularity and tail conditions on the Markov kernels k_ψ .

Assumption A0. \mathcal{P}_ψ is a subset of a regular and minimal [264, Section 3.2] exponential family with natural parameter domain $\mathcal{D} \subseteq \mathbb{R}^p$, Ψ is a convex and compact subset of \mathcal{D} , and ψ^* lies in the interior of Ψ .

Assumption A1. There exists a constant $C_\chi > 0$ such that $\chi^2(p_{\psi^*}, p_\psi) \leq C_\chi^2 \|\psi - \psi^*\|^2$

Assumption A2. $\alpha := \sup\{\alpha(f, \psi), f \in \{\phi_i\}_{i=1}^p \cup \{\phi_i \phi_j\}_{i,j=1}^p, \psi \in \Psi\} < 1$, where ϕ_i is the i -th component of the function ϕ , and ϕ_i^2 is the i -th component of the function $x \mapsto \phi(x)^2$.

A well known property of EFs [264, Proposition 3.1] is that their negative cross-entropy (against any other measure) is C^∞ , convex, and strictly so if the exponential family is minimal (meaning that the set of sufficient statistic functions ϕ_i are not linearly dependent). Leaving aside the issue of intractable expectations, this convexity suggests that \mathcal{L} can be efficiently minimized using stochastic approximation algorithms [180, 169]. The compactness of Ψ provided by Assumption A0 thus ensures, by the extreme value theorem [97], the existence of finite positive constants μ and L defined as:

$$\mu := \min_{\psi \in \Psi} \lambda_{\min} (\nabla_\psi^2 \mathcal{L}(\psi)), \quad L := \max_{\psi \in \Psi} \lambda_{\max} (\nabla_\psi^2 \mathcal{L}(\psi)), \quad (\text{IV.5})$$

where $\nabla_\psi^2 \mathcal{L}$ is the Hessian of \mathcal{L} with respect to ψ . μ (called the *strong convexity* constant) and L (a bound controlling the smoothness of the problem) play a critical role in the analysis of convex optimization algorithms [180]. While it is possible to obtain convergence rates in non-smooth or non-strongly-convex settings, our analysis follows the spirit of [121] by leveraging the strong convexity of the problem to compensate for the bias introduced by using CD gradients instead of unbiased stochastic gradients.

Assumption A1 allows link variations in distribution space to variations in parameter space, and will be instrumental to control the bias of the CD gradient. Note that since

$\chi^2(p_{\psi^*}, p_\psi) = e^{\log Z(2\psi - \psi^*) - (2\log Z(\psi) - \log Z(\psi^*))} - 1$ provided that $2\psi - \psi^* \in \mathcal{D}$ (see [181, Lemma 1]), we expect Assumption A1 to hold in many cases of interests. On the other hand, the possible exponential scaling of C_χ w.r.t $\log Z$ suggests that this constant may be large in some instances.

Assumption A2 is a *restricted* spectral gap condition: it guarantees that the time required by the MCMC algorithm to estimate expectations of ϕ and ϕ^2 under p_ψ will be uniformly bounded. This assumption is weaker than the (unrestricted) uniform spectral gap condition of [121], which requires that α controls the convergence rate of *all* functions in $L^2(p_\psi)$. Note that standard results in stochastic analysis [16] guarantee that $\alpha \leq 1$: thus, it only remains to ensure that α is strictly less than 1. Spectral gaps are strongly dependent on two properties of distribution: their tail behavior and their multimodality. While multimodality poses the risk of pushing the constant α close to 1, very heavy tails distributions may not verify the spectral gap condition at all.

IV.3.2 Results

IV.3.2.1 Parametric convergence of online CD

In this section, we show that under the assumptions stated in Section IV.3.1, the iterates ψ_t produced by the online CD algorithm described in Algorithm 1 will converge to the true parameter ψ^* at the parametric rate $O(n^{-1/2})$. To do so, we follow a well known paradigm in convex optimization [169] by deriving a recursion on the quantity $\delta_t := \mathbb{E} \|\psi_t - \psi^*\|^2$, which will allow, after unrolling, to obtain convergence rates for the iterates ψ_t . We aim to characterize precisely the impact of performing CD as opposed to performing online SGD on \mathcal{L} , which would consist of replacing the CD gradient h_t of Algorithm 1 by the unbiased (stochastic) gradient, given by:

$$g_t(\psi) := -\phi(X_t) + \nabla_\psi \log Z(\psi) \quad (\text{IV.6})$$

which satisfies $\mathbb{E} g_t(\psi) = \nabla_\psi \mathcal{L}(\psi)$. The only stochasticity in g_t comes from the sampling of a single data point x_t from the true distribution, which is unavoidable in the online setting, and we have $\mathbb{E} \|g_t(\psi^*)\|^2 = \text{Tr}(\text{Cov}[\phi(X_1)]) =: \sigma_*^2$. σ_*^2 plays a key role in the analysis of Stochastic Gradient Descent [169]. We expect that

replacing g_t by h_t will introduce two sources of approximation: a bias term coming from using a finite number of MCMC steps m , and an *additional* variance term, coming from using a single sample \tilde{x}_t^m to estimate $\nabla_{\psi} \log Z(\psi_t)$. With that in mind, we derive a recursion on δ_t in the following lemma.

Lemma IV.3.1. *Let $(\psi_t)_{0 \leq t \leq n}$ be the iterates from Algorithm 1. Denote $\delta_t = \mathbb{E}\|\psi_t - \psi^*\|^2$, $\sigma_* = (\mathbb{E}\|\phi(X_1) - \mathbb{E}[\phi(X_1)]\|^2)^{1/2}$, and $\sigma_t = (\mathbb{E}\|\phi(X^{\psi_t}) - \mathbb{E}[\phi(X^{\psi_t})]\|^2)^{1/2}$. Then, under A0, A1 and A2, for all $t \geq 1$,*

$$\delta_t \leq (1 - 2\eta_t \tilde{\mu}_{m,t-1} + 2\eta_t^2 L^2) \delta_{t-1} + 2\eta_t^2 \tilde{\sigma}_{m,t-1}^2 + 4\alpha^{m/2} \eta_t^2 \|\log Z\|_{3,\infty} C_\chi \delta_{t-1}^{1/2} \quad (\text{IV.7})$$

where $\|\log Z\|_{3,\infty}$ is a constant, $\tilde{\mu}_{m,t} := \mu - \alpha^m \sigma_t C_\chi$, and $\tilde{\sigma}_{m,t} := (\sigma_*^2 + \sigma_t^2 + 2\sigma_t^2 \alpha^{2m})^{1/2}$.

Lemma IV.3.1 is proved in Appendix IV.D.2, which details the form of $\|\log Z\|_{3,\infty}$, a constant that we expect to scale roughly as dL . Loosely speaking, this recursion suggests that as the learning rate η_t goes to 0, the two terms scaling in η_t^2 will be negligible, in which case we will have: $\delta_t \leq (1 - 2\eta_t \tilde{\mu}_{m,t-1}) \delta_{t-1} < \delta_{t-1}$, yielding convergence of δ_t to 0. We make these arguments formal in the next theorem. The reader familiar with the convex optimization literature will note the similarities between this recursion and the one derived in [169], which would apply as is to online SGD on \mathcal{L} using g_t . The difference between the two recursions is that the roles of the strong convexity constant μ and the noise σ_* are now played respectively by

$$\tilde{\mu}_{m,t-1} = \mu - \alpha^m \sigma_{t-1} C_\chi \quad \text{and} \quad \tilde{\sigma}_{m,t-1}^2 = \sigma_*^2 + \sigma_t^2 + 2\sigma_t^2 \alpha^{2m}$$

These two modifications respectively characterize the impact of the bias and the additional variance introduced by the CD gradient. The last term in Equation IV.7, scaling in $\alpha^{m/2} \eta_t^2 \sqrt{\delta_t}$, is a residual higher order mixed term coming from relating the variance of the Markov chain sample \tilde{x}_t^m to σ_t^2 . This term can be easily controlled as done next, and disappears as $m \rightarrow \infty$. Investigating the impact of m in the recursion, we notice that as $m \rightarrow \infty$, $\tilde{\mu}_{m,t} \rightarrow \mu$. As we will see later, this ensures that CD will

converge for a sufficiently high m . On the other hand, in that same regime, $\tilde{\sigma}_{m,t}$ does not converge to σ_* , but rather to $(\sigma_*^2 + \sigma_t^2)^{1/2}$, showing the irreducible impact of the variance term. While we precisely investigate the impact of the residual variance term in the next section, we now unify σ_* and σ_t by introducing

$$\sigma := \sup_{\psi \in \Psi} (\mathbb{E} [\|\phi(X^\psi) - \mathbb{E}[\phi(X^\psi)]\|^2])^{1/2}. \quad (\text{IV.8})$$

σ is an upper bound on the noise induced *both* by the CD gradient and by the online setup, and was used in prior work [121]. Note that by the properties of $\log Z$, σ^2 also equals $\sup_{\psi \in \Psi} \text{tr}(\nabla_\psi^2 \mathcal{L}(\psi))$, where $\text{tr}(A)$ is the trace of $A \in \mathbb{R}^{p \times p}$, and thus finite by the extreme value theorem. The following theorem is obtained by invoking standard unrolling arguments in the convex optimization literature. In the next result, we use the function $\varphi_\gamma(t)$, defined as $\varphi_\gamma(t) = \frac{t^\gamma - 1}{\gamma}$ if $\gamma \neq 0$, and $\log t$ if $\gamma = 0$.

Theorem IV.3.2. *Fix $n \geq 1$. Let $(\psi_t)_{0 \leq t \leq n}$ be the iterates produced by Algorithm 1, and define $\delta_t := \mathbb{E} \|\psi_t - \psi^*\|^2$. Moreover, assume that $m > \frac{\log(\sigma C_\chi / \mu)}{\log |\alpha|}$, i.e. $\tilde{\mu}_m := \mu - \alpha^m \sigma C_\chi > 0$. Then under Assumptions A0, A1 and A2, for $\eta_t = Ct^{-\beta}$ with $C > 0$, we have:*

$$\delta_n \leq \begin{cases} 2 \exp(4\tilde{L}C^2 \varphi_{1-2\beta}(n)) \exp\left(-\frac{\tilde{\mu}_m C}{4} n^{1-\beta}\right) \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m n^\beta}, & \text{if } 0 \leq \beta < 1 \\ \frac{\exp(2\tilde{L}^2 C^2)}{n^{\tilde{\mu}_m C}} \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + 2\tilde{\sigma}_m^2 C^2 \frac{\varphi_{\tilde{\mu}_m C/2-1}(n)}{n^{\tilde{\mu}_m C/2}}, & \text{if } \beta = 1, \end{cases}$$

where $\tilde{\sigma}_m = \sigma^2(2 + 2\alpha^{2m}) + \alpha^{m/2} \|\log Z\|_{3,\infty}^2 C_\chi^2$ and $\tilde{L} = (L^2 + \alpha^{m/2})^{1/2}$. Consequently, if $\eta_n = \frac{C}{n}$ with an initial learning rate $C > 2\tilde{\mu}_m^{-1}$, we have $\sqrt{\delta_n} \leq 2\tilde{\sigma}_m C \sqrt{\frac{\tilde{\mu}_m C}{\tilde{\mu}_m C - 2}} \frac{1}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$.

Theorem V.D.1 is proved in Appendix IV.D.3. It shows that the iterates produced by online CD will converge to the true parameter ψ^* at the rate $O(n^{-1/2})$ provided that the number of steps m is sufficiently large, improving over the asymptotic $O(n^{-1/3})$ rate of [121], while imposing slightly weaker conditions on the number of steps m (see [121, Theorem 2.1]). This proves that online CD can be asymptotically competitive with other methods for training unnormalized models, such as Noise Contrastive Estimation [92], or Score Matching [117]. However, the asymptotic

variance of ψ_t (e.g. the multiplicative factor in front of the $O(n^{-1/2})$ term) is likely to be suboptimal, e.g. much larger than the Crámer-Rao bound, given by the trace of the inverse of the Fisher information matrix [264]. Given the statistical optimality of MLE, and the fact that CD is an approximate MLE method, this motivates the further goal of obtaining a CD estimator with near-optimal statistical properties. In the next section, we achieve this goal by showing that averaging the iterates ψ_t will produce a near statistically-optimal estimator, in a sense that we will make precise.

IV.3.2.2 Towards statistical optimality with averaging

Polyak-Ruppert averaging [196] is a simple yet surprisingly effective way to construct an asymptotically optimal estimator $\bar{\psi}_n := \frac{1}{n} \sum_{i=1}^n \psi_i$ from a sequence of iterates $(\psi_t)_{0 \leq t \leq n}$ obtained by running a standard online SGD algorithm [169]. As shown in [169], when the objective is the cross-entropy of a model, and assuming the unbiased stochastic gradients are available, averaging yields an estimator $\bar{\psi}$ with the asymptotic variance $\text{tr}(\mathcal{I}(\psi^*)^{-1})/n$, where $\mathcal{I}(\psi) := \text{Cov}[\phi(X_1)]$ is the Fisher information matrix of the data distribution p_{ψ^*} . $\mathcal{I}(\psi^*)^{-1}$ being the Cramér-Rao lower bound on asymptotic variances of statistical estimators [35], this estimator $\bar{\psi}_n$ is asymptotically optimal. The following theorem shows conditions under which averaging CD iterates can give rise to a near-optimal estimator.

Theorem IV.3.3 (Contrastive Divergence with Polyak-Ruppert averaging). *Let $(\psi_t)_{t \geq 0}$ the sequence of iterates obtained by running the CD algorithm with a learning rate $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$. Define $\bar{\psi}_n := \frac{1}{n} \sum_{i=1}^n \psi_i$. Then, under the same assumptions as Theorem V.D.1, and assuming additionally that $m := m(n) > \frac{(1-\beta)\log n}{2|\log \alpha|}$, we have, for all $n \geq 1$,*

$$(\mathbb{E} \|\bar{\psi}_n - \psi^*\|^2)^{1/2} \leq 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n}} + o(n^{-1/2})$$

Consequently, we have that $\limsup_{n \rightarrow \infty} n \mathbb{E}(\|\bar{\psi}_n - \psi^*\|^2) \leq 4 \times \text{tr}(\mathcal{I}(\psi^*)^{-1})$.

Theorem IV.3.3, alongside with a statement which includes the asymptotic order of the residual term, is proved in Appendix IV.D.4. It shows that at the cost of an increase in computational complexity of the entire algorithm from $O(n)$ to $O(n \log n)$,

$\bar{\psi}_n$ will be a near-optimal statistical estimator of ψ^* . While this increase in complexity emerges from the bias of CD, the additional variance of CD results in an asymptotic variance inflated by a factor of 4 compared to the Cramér-Rao bound.

Theorem IV.3.3 concludes our analysis of online CD. Despite their asymptotic near-optimality, the bounds provided for online CD and its averaged version have weaknesses: the online CD iterates are not robust to choices of C . On the other hand, as shown in Appendix IV.D.4, the bound of the averaged iterates contain higher-order terms that could be large in intermediate sample regimes. Next, we show that offline CD, which processes data points multiple times, can alleviate these issues.

IV.4 Non-asymptotic analysis of offline CD

In practice, CD gradient approximation schemes are commonly used within an offline stochastic gradient descent (SGD) algorithm, where one is given the full size- n dataset upfront and each update uses some stochastic subset of the data. We study CD under offline SGD with replacement (SGDw), i.e. Algorithm 2 with batches $B_{t,j}$ being i.i.d. uniform draws of size- B subsets of $[n]$, and include SGD without replacement in Section IV.B.2. To do so, we follow the setting of prior work on offline CD [121], which established its asymptotic $O(n^{-\frac{1}{3}})$ consistency. We show that by slightly strengthening a moment assumption used in [121], the offline CD iterates converge to the true parameter at a near-parametric $O((\log n)^{\frac{1}{2}} n^{-\frac{1}{2}})$ rate. Our proof proceeds by controlling a “tail probability” term specific to the offline setting which characterizes the strength of the correlations between the offline CD iterates and the training data. While, as we show, the assumptions of [121] provide a tail control sufficient to obtain a near-parametric rate, other strategies are possible to obtain convergence guarantees. In particular, we show that non-asymptotic convergence can be obtained by either (1) relaxing assumptions on the Markov kernel required by prior work, or (2) making a specific mixing assumption the Markov chain.

IV.4.1 Background: Asymptotic consistency of offline CD in subexponential settings

Prior work [121] has established *asymptotic* $O(n^{-\frac{1}{3}})$ consistency of the (averaged) offline CD iterates in the full-batch case. We summarize their results and assumptions below. In the following, we write $K_\psi^m(x) \sim k_\psi^m(x, \cdot)$ to keep the dependence of the MCMC sample on ψ and x explicit.

Assumption A3. *There exists $v \geq 2$ s.t. for all $m \in \mathbb{N}$, there is $\kappa_{v;m} < \infty$ s.t.*

$$\sup_{x \in \mathcal{X}} \sup_{\psi \in \Psi} (\mathbb{E} \|\phi(K_\psi^m(x)) - \mathbb{E}[\phi(K_\psi^m(x))]\|^v)^{1/v} \leq \kappa_{v;m}.$$

Assumption A4. *There exists some $C_m > 0$ such that, for all $\psi_1, \psi_2 \in \Psi$, $\sup_{x \in \mathcal{X}} \|\mathbb{E}[\phi(K_{\psi_1}^m(x))] - \mathbb{E}[\phi(K_{\psi_2}^m(x))]\| \leq C_m \|\psi_1 - \psi_2\|$.*

Assumption A5. *There exist some $\sigma_m, \zeta_m > 0$ such that, for any $z \in \mathbb{R}^p$ with $\|z\| \leq \zeta_m$, $\mathbb{E}[e^{z^\top(\phi(K_{\psi^*}^m(X_1)) - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \leq e^{\sigma_m^2 \|z\|^2 / 2}$.*

Theorem IV.4.1 (Theorem 2.1 of [121]). *Assume assumptions A0, A1, A2, A3 (for $v = 2$), A4 and A5. Let $\psi_{t,1}$ be the t -th iterate of offline CD with full-batch gradient descent and constant step size $\eta_t = C$, i.e, the iterates produced by Algorithm 2 using $B_{t,1} = [[1, n]]$. Then for any learning rate C and number of Markov kernel steps m satisfying $\mu - \alpha^m \sigma C_\chi - \frac{C}{2}(L + \alpha^m \sigma C_\chi)^2 > 0$, we have, for some $A_m > 0$,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\limsup_{T \rightarrow \infty} \left\| \frac{1}{T} \sum_{t=1}^T \psi_{t,1}^{\text{SGDw}} - \psi^* \right\| > A_m n^{-\frac{1}{3}} \right) = 0$$

This result shows convergence of the *averaged* full-batch CD iterates to the true parameter in the large n and T limit. As discussed, this result is asymptotic both in n and T : the probability of the error exceeding $A_m n^{-\frac{1}{3}}$ goes to 0 as $n \rightarrow \infty$ and $T \rightarrow \infty$, but at an unknown rate. Moreover, the $O(n^{-\frac{1}{3}})$ does not match the optimal $O(n^{-\frac{1}{2}})$ rate.

IV.4.2 Sharpening offline CD bounds in subexponential settings

IV.4.2.1 Non-asymptotic $\tilde{O}(n^{-1/2})$ -consistency

As a first result, we show that under the assumptions of [121] (except for a slightly stronger $v > 2$ moment assumption in A3), $\psi_{T,N}^{\text{SGDw}}$ in fact achieves a near-parametric rate. The most general version of our result holds for any learning rate schedule of the form $Ct^{-\beta}$, $\beta \in [0, 1]$, and for offline SGD with arbitrary batch sizes B , with data drawn either with or without replacement across batches. For simplicity, we first present our result assuming full batch ($B = n, N = 1, \psi_{t,j}^{\text{SGDw}} = \psi_{t,1}^{\text{SGDw}}$ for $t \geq 1$) SGD with constant step sizes $\eta_t = C$, which is the setting of [121]. Analogue bounds holding for the other mentioned batching and step sizes schedules can be found in Appendix IV.B.

Theorem IV.4.2. *Assume the setup of Theorem IV.4.1, except that Assumption A3 holds for some $v > 2$, and that $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 4CL^2$. Let $\delta_{t,j}^{\text{SGDw}} := \mathbb{E}\|\psi_{t,j}^{\text{SGDw}} - \psi^*\|^2$. Then, we have:*

$$\sqrt{\delta_{T,1}^{\text{SGDw}}} \leq E_1^{T,1} \sqrt{\delta_{0,0}^{\text{SGDw}}} + C'(p, v, m, \Psi) \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right) \left(\frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} + \frac{E_2^{T,1}}{L^2 C^2} \right) \quad (\text{IV.9})$$

where $E_1^{T,1}, E_2^{T,1}$ are functions decreasing exponentially in T , and $C'(p, v, m, \Psi)$ is a constant in n, T . Consequently,

$$\lim_{T \rightarrow \infty} \sqrt{\delta_{T,1}^{\text{SGDw}}} \leq \frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} C'(p, v, m, \Psi, \beta) \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{n}} \right).$$

The precise values of all the constants can be found in Theorem IV.B.1 (for $E_1^{T,1}, E_2^{T,1}$) and Lemma IV.B.3 (for $C'(p, v, m, \Psi, \beta)$), including their expressions for $N > 1$ and $\beta \in [0, 1]$. We comment on the main differences between our result and the one of [121]. First our bound holds for *any* epoch T and number of samples n . Second, fixing n but taking $T \rightarrow \infty$, the final bound matches the parametric $O(\sqrt{n})$ up to a $\sqrt{\log(n)}$ factor, a significant improvement over the $O(n^{-\frac{1}{3}})$ rate of [121]. Finally, we control an L_2 error, which is a stronger control than a high probability bound by Markov's inequality; we hypothesize this is the reason why a slightly stronger

moment assumption is required for our setup, compared to the one used for the high probability bound in [121].

Inspecting Equation IV.9, we notice the presence of two *transient* terms, and a *stationary term*, reminiscent of the structure of upper bound of Theorem V.D.1. The transient terms (i.e. the ones containing $E_1^{T,1}$ and $E_2^{T,1}$) vanish exponentially fast in the total number of CD updates T . However, unlike in online CD where the number of updates and the number of samples are tied (e.g. $T = n$), these two values are now *decoupled*, and these terms can be made arbitrarily small by increasing the number of gradient steps T without having to collect more samples n . The stationary term, which is the only one remaining in the limit of $T \rightarrow \infty$, decreases with n at a rate that is independent of hyperparameters like the step size C or the learning rate schedule β (see Lemma IV.B.3). In that sense, offline CD compares favorably to online CD, whose rate is sensitive to β and C , and averaged online CD, whose bound contains higher-order (in n) terms which can be large in the moderate n regime. On the other hand, the stationary term in offline CD is asymptotically suboptimal: its rate is larger (while only up to a log factor) than the best-case $O(\sqrt{n})$ one achieved by online CD algorithms, and the leading constant does not match the optimal one.

IV.4.2.2 Proof of Theorem IV.4.2

The high-level proof of Theorem IV.4.2 follows a similar strategy as the online one: first, derive a recursion for the quantity $\delta_{t,1}^{\text{SGDw}} := \mathbb{E}\|\psi_{t,1}^{\text{SGDw}} - \psi^*\|^2$, then unroll it explicitly to obtain a final bound on $\delta_{T,1}^{\text{SGDw}}$. The main difference to online CD is the presence of an additional offline-specific correlation between the iterates and the data. We thus break down the proof into three steps: (1) deriving a controllable, uniform-in-time upper bound of the data-iterate correlations, (2) deriving and unrolling a recursion on $\delta_{t,1}^{\text{SGDw}}$ containing this new term, and (3) controlling that term to obtain a final bound on $\delta_{T,1}^{\text{SGDw}}$.

Step 1: characterizing the data-iterate correlations in offline CD In offline CD, at each epoch $t \geq 1$, the iterate $\psi_{t-1,1}^{\text{SGDw}}$ and the data samples X_i are correlated: this is because these samples may have been used in previous epochs $t' < t - 1$ to obtain the $\psi_{t',1}^{\text{SGDw}}$, which themselves influenced $\psi_{t-1,1}$. With such correla-

tions, we now have $\mathbb{P}[X_i \in \bullet | \psi_{t-1,1}^{\text{SGDw}}] \neq \mathbb{P}[X_i \in \bullet]$, preventing us from obtaining an “unrollable” recursion on $\delta_{t,1}^{\text{SGDw}}$ by first marginalizing X_i out to obtain an upper bound of $\mathbb{E}[\|\psi_{t,1}^{\text{SGDw}} - \psi^*\|^2 | \psi_{t-1,1}^{\text{SGDw}}]$ that only depends on $\|\psi_{t-1,1}^{\text{SGDw}} - \psi^*\|$, and then marginalizing over $\psi_{t-1,1}^{\text{SGDw}}$ to obtain a recursion as in Lemma IV.3.1. As this problem would not have occurred had we used “fresh samples” (e.g. i.i.d copies of X_i not present in the training data) to perform our update, the core of the proof lies in controlling the following quantity:

$$\begin{aligned} & \Delta(\psi_{t,1}^{\text{SGDw}}) \\ &:= \left\| \frac{1}{n} \sum_{i \leq n} (\mathbb{E}[\phi(K_{i;\psi_{t,1}^{\text{SGDw}}}(X_i)) | \psi_{t,1}^{\text{SGDw}}, X_i] - \mathbb{E}[\phi(K_{i;\psi_{t,1}^{\text{SGDw}}}(X'_1)) | \psi_{t,1}^{\text{SGDw}}]) \right\| \end{aligned}$$

where X'_1 is an i.i.d. copy of X_1 . $\Delta(\psi_{t,1}^{\text{SGDw}})$ is the expected (over the data and iterates) error between a quantity that allows to obtain a recursion (the rightmost term) and the one actually used by offline CD (the leftmost term). To control it, we upper-bound it using a tail decomposition:

$$\begin{aligned} & \mathbb{E}[\Delta(\psi_{t,1}^{\text{SGDw}})^2] \\ & \leq \varepsilon^2 + (\sup_t \mathbb{E}[\Delta(\psi_{t,1}^{\text{SGDw}})^\nu])^{2/\nu} \sup_t \mathbb{P}(\Delta(\psi_{t,1}^{\text{SGDw}}) > \varepsilon)^{\frac{\nu-2}{\nu}} := \varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon)^2 \quad (\text{IV.10}) \end{aligned}$$

We invoke an additional assumption to ensure that $(\mathbb{E}[\Delta(\psi_{t,1}^{\text{SGDw}})^\nu])^{2/\nu}$ is finite; in the results of [121], this is automatically implied by assumptions A3 and A5. For simplicity, we assume the same bounding constant $\kappa_{\nu;m}$.

Assumption A6. *There exists $\nu \geq 2$ s.t. for all $m \in \mathbb{N}$, $\kappa_{\nu;m}$ from A3 moreover verifies*

$$\sup_{\psi \in \Psi} (\mathbb{E}\|\phi(K_\psi^m(X_1)) - \mathbb{E}[\phi(K_\psi^m(X_1))]\|^\nu)^{1/\nu} \leq \kappa_{\nu;m}.$$

Note the similarity of this assumption with assumption A3: the only difference is that X_1 is now a random training point instead of a deterministic (arbitrary) one. Ensuring assumption A6 in addition to assumption A3 thus requires controlling a ν -th order moment, instead of all moments as implied by assumption A5.

Step 2: Deriving and unrolling the recursion on $\delta_{t,1}^{\text{SGDw}}$ The right-hand side of Equation (IV.10) does not depend on t , allowing for the derivation of an “unrollable” recursion on $\delta_{t,1}^{\text{SGDw}}$ and its subsequent unrolling, which is performed in the following theorem. For simplicity, we again assume $\beta = 0$ and $N = 1$ and defer the general case to Theorem IV.B.1 in appendix.

Theorem IV.4.3 (Convergence up to a tail control). *Assume A0, A1, A2, A3 and A6. Let $\eta_t = C$ for some $C > 0$, and assume that $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 4CL^2$. Then for any $\varepsilon > 0$,*

$$\sqrt{\delta_{T,1}^{\text{SGDw}}} \leq E_1^{T,1} \sqrt{\delta_{0,0}^{\text{SGDw}}} + C \left(\varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{\nu,m}}{\sqrt{n}} \right) \left(\frac{e^{\frac{\tilde{\mu}_m C}{2}}}{\tilde{\mu}_m C} + \frac{E_2^{T,1}}{L^2 C^2} \right)$$

where $\varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon)$ is defined in Equation (IV.10).

Note that in the general, non-full batch $B \leq n$ case, $\frac{5\sigma + 5\kappa_{\nu,m}}{\sqrt{n}}$ is replaced by $\frac{5\sigma + 5\kappa_{\nu,m}}{\sqrt{B}}$ (see Theorem IV.B.1). Under our bounds, obtaining consistency thus requires setting

$$B \equiv B(n) \xrightarrow[n \rightarrow \infty]{} +\infty.$$

Step 3: Controlling the tail probability term Theorem IV.4.3 is just one step away from the final bound of Theorem IV.4.2: it remains to control the tail term $\varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon)$. Under the assumptions of [121], minimizing $\varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon)$ over ε yields the following result:

Lemma IV.4.4. *Assume the setup of Theorem IV.4.2. Let $n \in \mathbb{N}$ be sufficiently large s.t. $\frac{\log n}{n} < \frac{\sigma_m^2 \zeta_m^2}{p+\nu-2}$. Denote r_Ψ as the radius of the smallest sphere in \mathbb{R}^p that contains Ψ , which is finite under A0. Then*

$$\inf_{\varepsilon > 0} \varepsilon_{\nu;n,m,T}^{\text{SGDw}}(\varepsilon) \leq \frac{\sqrt{\log n}}{\sqrt{n}} \times \left(\frac{3\sigma_m \sqrt{p((\nu-2)p+2\nu)}}{\sqrt{\nu-2}} + \kappa_{\nu,m} 2^{\frac{\nu-2}{2\nu}} (r_\Psi)^{\frac{(\nu-2)p}{2\nu}} \left(1 + \frac{2C_m(\nu-2)^{1/2}}{\sigma_m p^{1/2} ((\nu-2)p+2\nu)^{1/2}} \right)^{\frac{(\nu-2)p}{2\nu}} \right).$$

To obtain this result, we control the moment term ($\sup_t \mathbb{E}[\Delta(\psi_{t,1}^{\text{SGDw}})^\nu]$) using A6, and we control the tail probability term $\sup_t \mathbb{P}(\Delta(\psi_{t,1}^{\text{SGDw}}) > \varepsilon)$ as in [121, Lemma

3.1] using a union bound, a covering argument and A5. Theorem IV.4.2 then follows by plugging Lemma IV.4.4 into Theorem IV.4.3.

IV.4.3 Consistency of offline CD: beyond subexponential tails.

As discussed above, the general unrolling result of Theorem IV.4.3 holds without the subexponentiality assumption A5; this assumption was only used in Lemma IV.4.4 to control $\varepsilon_{n,m,T;\nu}^{\text{SGDw}}(\varepsilon)$. We now discuss two alternative ways to control this quantity without requiring subexponential tails. The first generalizes the idea of Jiang et al. [121], while the second exploits mixing of the Markov chain $K_\Psi^m(x)$ as $m \rightarrow \infty$. As before we only state partial results (full batch, $\beta = 0$) and defer the full explicit bounds to Section IV.B.3.

Control via Markov Inequality Our first alternative uses Markov Inequality to yield the following.

Theorem IV.4.5. *Assume the setup of Theorem IV.4.3 and additionally that A4 holds.*

Then

$$\inf_{\varepsilon > 0} \varepsilon_{\nu;n,m,T}^{\text{SGDw}}(\varepsilon) \leq \tilde{C}(p, \nu, m, \Psi) n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}},$$

$$\lim_{T \rightarrow \infty} \sqrt{\delta_{T,1}^{\text{SGDw}}} \leq \tilde{C}'(p, \nu, m, \Psi) \left(n^{-\frac{(\nu-2)\nu}{2(\nu^2+(\nu-2)p)}} + \frac{1}{\sqrt{n}} \right),$$

where \tilde{C} and \tilde{C}' are functions whose explicit expressions are given in Lemma IV.B.4 in the appendix.

In the case $p = 1$ and $\nu = 3$, the suboptimal error from Theorem IV.4.5 reads $O(n^{-3/20})$. Theorems IV.4.2 and IV.4.5 reveal that, depending on the tail condition imposed on the noise introduced by the Markov kernel, the convergence rate of offline CD varies: A subexponential tail, as assumed in prior work, in fact leads to near-parametric rate. Meanwhile, consistency can be obtained without assuming subexponentiality, albeit at a suboptimal rate.

Control via Markov chain mixing. Alternatively, notice that $\mathbb{E}[\Delta(\psi_{t,1}^{\text{SGDw}})^2]$ involves an average of $\mathbb{E}[\phi(K_{\psi_{t,1}^{\text{SGDw}}}(X_i)^2) | X_i, \psi_{t,1}^{\text{SGDw}}] - \mathbb{E}[\phi(K_{\psi_{t,1}^{\text{SGDw}}}(X'_1)) | \psi_{t,1}^{\text{SGDw}}]$. When $m \rightarrow \infty$, the effect of initialization vanishes, and one may expect the difference

to converge to zero. We defer to Lemma IV.B.5 in the appendix to show that, under a ϕ -discrepancy mixing condition ([200]) with a mixing coefficient $\tilde{\alpha} \in [0, 1]$,

$$\inf_{\varepsilon > 0} \mathcal{E}_{V;n,m,T}^{\text{SGDw}}(\varepsilon) = O(\kappa_{V;m} \tilde{\alpha}^{\frac{(V-2)m}{3V-2}}) \quad \lim_{T \rightarrow \infty} \sqrt{\delta_{T,1}^{\text{SGDw}}} = O\left(\kappa_{V;m} \tilde{\alpha}^{\frac{(V-2)m}{3V-2}} + \frac{\sigma + \kappa_{V;m}}{\sqrt{n}}\right).$$

As $m \rightarrow \infty$, this recovers the parametric rate $O(n^{-1/2})$. This alternative convergence guarantee comes at the cost of requiring m , the number of Markov chain steps, to grow with the sample size n .

Remark (Examples). In our main results (Theorems IV.3.3, IV.4.3 and V.D.1) and the tail condition for offline SGD (Theorem IV.4.2), we employed a weaker set of assumptions than those in [121] (except for the mild $V > 2$ moment assumption in (A3)). Consequently, our results apply to all three examples studied in [121]: A bivariate Gaussian model with unknown mean and random-scan Gibbs sampler, a fully visible Boltzmann machine with random-scan Gibbs sampler, and an exponential-family random graph model with a Metropolis-Hastings sampler.

IV.5 Related Work

Central to this paper is the prior work of Jiang et al. [121], which provided a rigorous theoretical foundation to analyze the convergence of full-batch CD, and which we refine. The study of optimization with biased gradient descent has attracted a lot of attention in recent years [133, 235, 58, 15]. These works, while closely connected to ours, analyze algorithms with different implementation choices than the CD algorithm: i.i.d. noise setup [133], or setup where a persistent Markov chain is maintained through the iterations [133, 235, 58, 15]. The latter is akin to a variant of the CD algorithm, called the persistent CD [251]. In contrast, our analysis focus on the CD algorithm that restarts a batch of Markov chains from the data distribution at every iteration. Finally, there is a rich body of work on convergence guarantees for offline multi-pass SGD [65, 96, 153, 194, 171, 280]. A notable difference of our analysis is that we are primarily concerned with statistical errors associated with convergence to the true parameter ψ^* in number of samples n , and not the commonly studied convergence rate in number of epochs T . Consequently, most of our work

for the offline setup goes into handling the correlations that accumulate by reusing data across epochs.

IV.6 Discussion

In this work, we provide a non-asymptotic analysis of the Contrastive Divergence algorithms, showing, in the online setting, their potential to converge at the parametric rate and to have near-optimal asymptotic variance, and proving a near-parametric rates in the offline setting, significantly extending prior results. Our results apply to unnormalized exponential families: despite their flexibility, these models only cover log-densities with linear relationships on the model parameters. We believe that extending our results to more general forms of unnormalized models is an important direction for future work.

Acknowledgments and Disclosure of Funding

All authors acknowledge support from the Gatsby Charitable Foundation.

Appendix

Supplementary Material for the paper Near Optimality of Contrastive Divergence Algorithms

The supplementary material provides the proofs of the main results of the paper:

Section IV.B states full explicit bounds for the offline CD algorithm.

Section IV.C collects a list of useful tools for our proofs. These include the properties of φ_γ introduced before Theorem V.D.1 in the main text, as well as several contraction and integrability results.

Section IV.D provides the proofs for the online CD algorithm.

Sections IV.E, IV.F and IV.G contain the proofs about the offline CD algorithm and the tail control.

IV.A Notations

Throughout the proofs, we will denote by P_ψ^m the following operator from $L^2(p_\psi)$ to itself:

$$P_\psi^m f(x) := \int k^m(x, x') f(x') p_\psi(x') dx'. \quad (\text{IV.11})$$

Here, $k^m(x, x')$ is the m -iterated version of some Markov transition kernel k_ψ , e.g.:

$$k_\psi^m(x, x') := \int k_\psi(x, x_1) \dots k_\psi(x_{m-2}, x_{m-1}) \dots k_\psi(x_{m-1}, x') dx_1 \dots dx_{m-1}. \quad (\text{IV.12})$$

$\text{Proj}_\Psi : \mathbb{R}^p \mapsto \Psi$ denotes the projection operator onto the convex set Ψ , e.g.

$$\text{Proj}_\Psi(\psi) := \arg \min_{\psi' \in \Psi} \|\psi - \psi'\|.$$

We also frequently use the following function, used in standard convex optimization results [169].

$$\varphi_\gamma(t) = \begin{cases} \frac{t^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ \log t & \text{if } \gamma = 0 \end{cases}$$

which is defined on $\mathbb{R}_+ \setminus \{0\}$.

Finally, we recall the notation $K_\psi^m(x) \sim k_\psi^m(x, \cdot)$. We will note, in the proofs regarding online CD, X^ψ a sample drawn independently of X_1, \dots, X_n and on a given $\psi \in \Psi$,

as

$$X^\psi \sim p_\psi .$$

Notations for Offline CD For offline CD, we will note by $X_1^\psi, \dots, X_n^\psi$ the i.i.d. samples, drawn independently of X_1, \dots, X_n and on a given $\psi \in \Psi$, as

$$X_1^\psi, \dots, X_n^\psi \stackrel{\text{i.i.d.}}{\sim} p_\psi .$$

We also write, for $m \in \mathbb{N} \cup \{\infty\}$,

$$K_{1;\psi}^m(x), \dots, K_{n;\psi}^m(x) \stackrel{\text{i.i.d.}}{\sim} k_\psi^m(x, \cdot)$$

IV.B Additional results for offline SGD

In this section, we provide the full statements on error bounds for SGD with replacement (SGDw), SGD with reshuffling (SGDo) and tail moment bounds, which complement the results in Section IV.4. Proofs are deferred to Section IV.F, which make use of L_2 approximation by auxiliary gradient updates derived in Section IV.E.

Notations Denote the SGDw iterates by $(\psi_{t,j}^{\text{SGDw}})_{t \in \mathbb{N}, j \leq N}$ and let X'_1 be an i.i.d. copy of X_1 . Throughout the remaining of the appendix, we define given $\varepsilon > 0$ and $n \in \mathbb{N}$

$$\begin{aligned} \vartheta_{n,m,T}^{\text{SGDw}}(\varepsilon) \\ := \sup_{\substack{t \in [T] \\ j \in [N]}} \mathbb{P} \left(\frac{\left\| \sum_{i=1}^n \left(\mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X_i) \right) \middle| X_i, \psi_{t-1,j}^{\text{SGDw}} \right] - \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X'_1) \right) \middle| \psi_{t-1,j}^{\text{SGDw}} \right] \right) \right\|}{n} > \varepsilon \right) \\ = \sup_{t,j} \mathbb{P}(\Delta(\psi_{t,j}^{\text{SGDw}}) > \varepsilon) \end{aligned}$$

and $\vartheta_{n,m,T}^{\text{SGDo}}(\varepsilon)$ analogously. Using these notations, we can redefine the quantity $\varepsilon_{n,m,T;v}^{\text{SGDw}}(\varepsilon)$ in the main as $\varepsilon_{n,m,T;v}^{\text{SGDw}}(\varepsilon) := \sqrt{\varepsilon^2 + \kappa_{v;m}^2 (\vartheta_{n,m,T}^{\text{SGDw}}(\varepsilon))^{\frac{v-2}{v}}} .$

IV.B.1 An explicit finite-sample bound for SGDW

In the result below, we write $\delta_{t,j}^{\text{SGDW}} := \mathbb{E} \|\psi_{t,j}^{\text{SGDW}} - \psi^*\|^2$ and, for a fixed $\varepsilon > 0$, the quantity

$$\sigma_{n,T}^{\text{SGDW}} = \varepsilon_{n,m,T;\nu}^{\text{SGDW}}(\varepsilon) + \frac{5\sigma + 5\kappa_m}{\sqrt{B}} = \sqrt{\varepsilon^2 + \kappa_m^2 (\vartheta_{n,m,T}^{\text{SGDW}}(\varepsilon))^{\frac{\nu-2}{\nu}}} + \frac{5\sigma + 5\kappa_m}{\sqrt{B}}.$$

Theorem IV.B.1. Assume A0 (where Ψ may be non-compact), A1, A2, A3 and A6. Let $\eta_t = Ct^{-\beta}$ for some $\beta \in [0, 1]$ and $C > 0$, and assume that $m > \frac{\log(\sigma C_\chi / \mu)}{\log |\alpha|}$ s.t. $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 0$ as in Theorem V.D.1. Then for any $\varepsilon > 0$, $\sqrt{\delta_{T,N}^{\text{SGDW}}}$ is upper bounded by

$$\left\{ \begin{array}{ll} E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDW}}} + C \sigma_{n,T}^{\text{SGDW}} \left(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1 + \tilde{\mu}_m C)^{N-1} \varphi_{\frac{1}{2}-L^2C^2N}(T+1) E_2^{T,N} \right) & \text{for } \beta = \frac{1}{2}, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDW}}} + C \sigma_{n,T}^{\text{SGDW}} \left(\frac{4}{\tilde{\mu}_m C} + \frac{3N \left(1 + \frac{L^2C^2}{2}\right)^{N-1} e^{2L^2C^2N} \log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}} \right) & \text{for } \beta = 1, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDW}}} + C \sigma_{n,T}^{\text{SGDW}} \left(\frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)} \frac{N}{(T+1)\beta}} + \frac{3^\beta (1 + \tilde{\mu}_m C)^{N-1} (T+2)^\beta}{L^2 C^2} E_2^{T,N} \right) & \text{otherwise,} \end{array} \right.$$

where $E_1^{T,N}$ and $E_2^{T,N}$ are two decreasing functions in T defined by

$$\begin{aligned} E_1^{T,N} &:= \exp \left(1 - N \tilde{\mu}_m C \varphi_{1-\beta}(T+1) + \frac{NL^2C^2}{2} \varphi_{1-2\beta}(T+1) \right), \\ E_2^{T,N} &:= \exp \left(- \frac{N \tilde{\mu}_m C}{2} \varphi_{1-\beta}(T+1) + 2NL^2C^2 \varphi_{1-2\beta}(T+1) \right). \end{aligned}$$

We emphasize that the full result above holds for any $\beta \in [0, 1]$, which in particular includes the constant step size $\beta = 0$ regime considered by [121]. When $\beta = 0$, for $E_1^{T,N}$ and $E_2^{T,N}$ to decay to zero as $T \rightarrow \infty$, we additionally need the condition

$$\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 4CL^2.$$

This is almost identical to the condition used in [121, Equation 2.5, Theorem 2.1],

except that $4L^2$ gets replaced by $\frac{1}{2}(L + \alpha^m \sigma C_\chi)^2$. Notably this says that an additional step size condition is needed for our results to hold in the constant step size regime, but not necessary for a decreasing step size.

IV.B.2 Results for SGDo

SGD with reshuffling (SGDo, also called SGD without replacement) is an optimization scheme that is also widely used in practice compared to SGDo and online SGD. In the context of CD, it corresponds to Algorithm 2 with batches chosen as

$$(B_{t,1}, \dots, B_{t,N}) = \pi(\{1, \dots, n\}),$$

where π is a uniform draw of the permutation group on n elements. We denote the iterates of SGDo $(\psi_{t,j}^{\text{SGDo}})_{t \in \mathbb{N}, j \in [N]}$. Analogously to $\vartheta_{n,m,T}^{\text{SGDw}}$, we define, for X'_1 an i.i.d. copy of X_1 , $\varepsilon > 0$ and $n \in \mathbb{N}$, the tail probability term

$$\vartheta_{n,m,T}^{\text{SGDo}}(\varepsilon) := \sup_{\substack{t \in [T] \\ j \in [N]}} \mathbb{P} \left(\frac{\left\| \sum_{i=1}^n \left(\mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDo}}}^m(X_i) \right) \middle| X_i, \psi_{t-1,j}^{\text{SGDo}} \right] - \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDo}}}^m(X'_1) \right) \middle| \psi_{t-1,j}^{\text{SGDo}} \right] \right) \right\|}{n} > \varepsilon \right).$$

Also denote $\epsilon_{n,m,T;v}^{\text{SGDo}}(\varepsilon) = \sqrt{\varepsilon^2 + \kappa_m^2 (\vartheta_{n,m,T}^{\text{SGDo}}(\varepsilon))^{\frac{v-2}{v}}}$ and $\sigma_{n,T}^{\text{SGDw}} = \epsilon_{n,m,T;v}^{\text{SGDo}}(\varepsilon) + \frac{5\sigma + 5\kappa_m}{\sqrt{B}}$. The following result says that $\psi_{t,j}^{\text{SGDo}}$ enjoys exactly the same convergence guarantee as $\psi_{t,j}^{\text{SGDw}}$ in Theorem IV.4.3. The statement is identical to that of Theorem IV.B.1 and is stated in full for completeness; see Section IV.F.2 for the proof, which is a slight adaptation of the proof for Theorem IV.B.1. As before we write $\delta_{t,j}^{\text{SGDo}} := \mathbb{E} \|\psi_{t,j}^{\text{SGDo}} - \psi^*\|^2$.

Theorem IV.B.2 (Convergence of CD-SGDo). *Assume A0 (where Ψ may be non-compact), A1, A2, A3 and A6. Let $\eta_t = Ct^{-\beta}$ for some $\beta \in [0, 1]$ and $C > 0$, and assume that $m > \frac{\log(\sigma C_\chi / \mu)}{\log |\alpha|}$ s.t. $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi > 0$ as in Theorem V.D.1. Then*

for any $\varepsilon > 0$, $\sqrt{\delta_{T,N}^{\text{SGDo}}}$ is upper bounded by

$$\left\{ \begin{array}{l} E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDo}}} + C\sigma_{n,T}^{\text{SGDo}} \left(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1+\tilde{\mu}_m C)^{N-1} \varphi_{\frac{1}{2}-L^2C^2N}(T+1) E_2^{T,N} \right) \\ \quad \text{for } \beta = \frac{1}{2}, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDo}}} + C\sigma_{n,T}^{\text{SGDo}} \left(\frac{4}{\tilde{\mu}_m C} + \frac{3N(1+\frac{L^2C^2}{2})^{N-1} e^{2L^2C^2N} \log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}} \right) \quad \text{for } \beta = 1, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDo}}} + C\sigma_{n,T}^{\text{SGDo}} \left(\frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)} \frac{N}{(T+1)\beta}} + \frac{3^\beta (1+\tilde{\mu}_m C)^{N-1} (T+2)^\beta}{L^2C^2} E_2^{T,N} \right) \\ \quad \text{otherwise,} \end{array} \right.$$

where $E_1^{T,N}$ and $E_2^{T,N}$ are two decreasing functions in T defined by

$$\begin{aligned} E_1^{T,N} &:= \exp \left(1 - N\tilde{\mu}_m C \varphi_{1-\beta}(T+1) + \frac{NL^2C^2}{2} \varphi_{1-2\beta}(T+1) \right), \\ E_2^{T,N} &:= \exp \left(- \frac{N\tilde{\mu}_m C}{2} \varphi_{1-\beta}(T+1) + 2NL^2C^2 \varphi_{1-2\beta}(T+1) \right). \end{aligned}$$

Remark. We also remark that existing works [177, 204] show that the standard SGDo typically gives a faster convergence rate in T than SGDw. An analogous result for the CD setup would involve additional technical hurdles of jointly controlling the correlations across minibatches and from reusing data samples, and we defer this to future work.

IV.B.3 Explicit tail control

We now provide the full explicit tail control bounds. All results in this section hold directly for $\varepsilon_{v;n,m,T}^{\text{SGDo}}(\varepsilon)$ and $\delta_{T,N}^{\text{SGDo}}$, and we omit them here. In the result below, we denote r_Ψ as the radius of the smallest sphere in \mathbb{R}^p that contains Ψ , which is finite under A0.

Lemma IV.B.3. *Assume A4 and A5. Let $n \in \mathbb{N}$ be sufficiently large s.t. $\frac{\log n}{n} < \frac{\sigma_m^2 \zeta_m^2}{p+v-2}$.*

Then

$$\inf_{\varepsilon > 0} \varepsilon_{v;n,m,T}^{\text{SGDw}}(\varepsilon) \leq \frac{\sqrt{\log n}}{\sqrt{n}} \times$$

$$\left(\frac{3\sigma_m \sqrt{p((v-2)p+2v)}}{\sqrt{v-2}} + \kappa_{v;m} 2^{\frac{v-2}{2v}} (r_\Psi)^{\frac{(v-2)p}{2v}} \left(1 + \frac{2C_m(v-2)^{1/2}}{\sigma_m p^{1/2} ((v-2)p+2v)^{1/2}} \right)^{\frac{(v-2)p}{2v}} \right).$$

In particular, if we additionally assume the conditions of Theorem IV.B.1, we have

$$\lim_{T \rightarrow \infty} \sqrt{\delta_{T,N}^{\text{SGDw}}} \leq C'(p, v, m, \Psi, \beta) \left(\frac{\sqrt{\log n}}{\sqrt{n}} + \frac{1}{\sqrt{B}} \right)$$

where

$$\begin{aligned} C'(p, v, m, \Psi, \beta) &:= \frac{8(1+5\sigma+5\kappa_m)}{\tilde{\mu}_m} \\ &\times \left(\frac{3\sigma_m \sqrt{p((v-2)p+2v)}}{\sqrt{v-2}} + \kappa_{v;m} 2^{\frac{v-2}{2v}} (r_\Psi)^{\frac{(v-2)p}{2v}} \left(1 + \frac{2C_m(v-2)^{1/2}}{\sigma_m p^{1/2} ((v-2)p+2v)^{1/2}} \right)^{\frac{(v-2)p}{2v}} \right). \end{aligned}$$

Lemma IV.B.4. Assume the conditions of Theorem IV.4.3 and additionally that A4 holds. Then

$$\begin{aligned} \inf_{\epsilon>0} \epsilon_{v;n,m,T}^{\text{SGDw}}(\epsilon) &\leq (3C_m + \kappa_{v;m}^{v/2} (r_\Psi)^{\frac{(v-2)p}{2v}} C_m^{-\frac{v-2}{2}} 3^{\frac{(v-2)p}{2v}}) n^{-\frac{(v-2)v}{2(v^2+(v-2)p)}}, \\ \lim_{T \rightarrow \infty} (\delta_{T,N}^{\text{SGDw}})^{1/2} &\leq \tilde{C}'(p, v, m, \Psi, \beta) (n^{-\frac{(v-2)v}{2(v^2+(v-2)p)}} + B^{-1/2}), \end{aligned}$$

where

$$\tilde{C}'(p, v, m, \Psi) := \frac{8(1+5\sigma+5\kappa_m)}{\tilde{\mu}_m} (3C_m + \kappa_{v;m}^{v/2} (r_\Psi)^{\frac{(v-2)p}{2v}} C_m^{-\frac{v-2}{2}} 3^{\frac{(v-2)p}{2v}}).$$

The next result considers a ϕ -discrepancy mixing condition ([200]), which is a mixing assumption on K_ψ^m but with respect to a specific test function ϕ , and we impose it uniformly over $\psi \in \Psi$. We also recall that $X_1^\psi \sim p_\psi$.

Lemma IV.B.5. Assume that there exist $\tilde{\alpha} \in [0, 1)$ and $\tilde{C}_K > 0$ such that, for all $\psi \in \Psi$ and $x \in \mathcal{X}$, $\|\mathbb{E}[\phi(K_\psi^m(x))] - \mathbb{E}[\phi(X_1^\psi)]\| \leq \tilde{C}_K \tilde{\alpha}^m$. Then

$$\inf_{\epsilon>0} \epsilon_{v;n,m,T}^{\text{SGDw}}(\epsilon) \leq (1 + 2^{\frac{v-2}{2v}} \kappa_{v;m} (\tilde{C}_K)^{\frac{v-2}{2v}}) \tilde{\alpha}^{\frac{(v-2)m}{3v-2}}.$$

In particular, if we additionally assume the conditions of Theorem IV.4.3, we have

$$\lim_{T \rightarrow \infty} \sqrt{\delta_{T,N}^{\text{SGDw}}} \leq \frac{8}{\mu - \alpha^m \sigma C_\chi} \left((1 + 2^{\frac{v-2}{2v}} \kappa_{v,m} (\tilde{C}_K)^{\frac{v-2}{2v}}) \tilde{\alpha}^{\frac{(v-2)m}{3v-2}} + \frac{5\sigma + 5\kappa_m}{\sqrt{B}} \right).$$

IV.C Auxiliary Tools

IV.C.1 Properties of φ_γ

The following lemma collects some identities used in [169].

Lemma IV.C.1. φ_γ satisfies the following properties:

- (i) φ_γ is increasing on \mathbb{R}_+ for all γ ;
- (ii) $\varphi_\gamma(t) \leq \frac{t^\gamma}{\gamma}$ for $\gamma > 0$, and $\varphi_\gamma(t) \geq -\frac{1}{\gamma}$ for $\gamma < 0$;
- (iii) $\varphi_{1-\beta}(t) \geq t^{1-\beta}$ for $\beta \in (0, 1]$;
- (iv) $\varphi_\gamma(t) - \varphi_\gamma(\frac{t}{2}) \geq \frac{1}{2}x^\gamma$ for $\gamma \in (0, 1]$.

The next lemma provides some additional results on φ_γ .

Lemma IV.C.2. φ_γ satisfies the following properties:

- (i) φ_γ is positive on $t > 0$ and increasing for every $\gamma \in \mathbb{R}$;
- (ii) For $1 \leq t_1 \leq t_2$ and $\beta \geq 0$, we have

$$\varphi_{1-\beta}(t_2 + 1) - \varphi_{1-\beta}(t_1) \leq \sum_{t=t_1}^{t_2} t^{-\beta} \leq 2(\varphi_{1-\beta}(t_2 + 1) - \varphi_{1-\beta}(t_1)).$$

If instead $\beta < 0$, we have

$$\frac{1}{2}(\varphi_{1-\beta}(t_2 + 1) - \varphi_{1-\beta}(t_1)) \leq \sum_{t=t_1}^{t_2} t^{-\beta} \leq \varphi_{1-\beta}(t_2 + 1) - \varphi_{1-\beta}(t_1);$$

- (iii) For $1 \leq t_1 \leq t_2$ and $\gamma \neq 0$, we have

$$\begin{aligned} t_1^{\gamma-1} &\leq \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \leq t_2^{\gamma-1} && \text{if } \gamma \geq 1, \\ t_2^{-(1-\gamma)} &\leq \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \leq t_1^{-(1-\gamma)} && \text{if } \gamma \leq 1; \end{aligned}$$

(iv) Let $1 \leq t_1 < t_2$ and $\kappa, \beta \geq 0$. If $\kappa \neq 1$ and $a > 0$, we have

$$\sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp(a \varphi_{1-\kappa}(t-1)) \leq \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{a} \exp(a \varphi_{1-\kappa}(t_2+1)),$$

and if $\kappa \neq 1$ and $a < 0$, we have

$$\sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp(a \varphi_{1-\kappa}(t)) \leq \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{(-a)} \exp(a \varphi_{1-\kappa}(t_1)).$$

Proof of Lemma IV.C.2. (i) follows from checking $\gamma > 0$, $\gamma = 0$ and $\gamma < 0$ respectively. The first set of bounds in (ii) follow by noting that $t \mapsto t^{-\beta}$ is decreasing for $\beta \geq 0$:

$$\begin{aligned} \sum_{t=t_1}^{t_2} t^{-\beta} &\geq \int_{t_1}^{t_2+1} t^{-\beta} dt = \varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1), \\ \sum_{t=t_1}^{t_2} t^{-\beta} &\leq 2 \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \leq 2 \int_{t_1-1}^{t_2} (t+1)^{-\beta} dt = 2(\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)). \end{aligned}$$

The second set of bounds follows from noting that $t \mapsto t^{-\beta}$ is increasing for $\beta < 0$:

$$\begin{aligned} \sum_{t=t_1}^{t_2} t^{-\beta} &\geq \frac{1}{2} \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \geq \frac{1}{2} \int_{t_1-1}^{t_2} (t+1)^{-\beta} dt = \frac{1}{2} (\varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1)), \\ \sum_{t=t_1}^{t_2} t^{-\beta} &\leq \int_{t_1}^{t_2+1} t^{-\beta} dt = \varphi_{1-\beta}(t_2+1) - \varphi_{1-\beta}(t_1). \end{aligned}$$

For (iii), we note that for $\gamma \neq 0$,

$$\varphi_\gamma(t_2) - \varphi_\gamma(t_1) = \frac{t_2^\gamma - t_1^\gamma}{\gamma},$$

so by the mean value theorem,

$$\inf_{t_1 \leq t \leq t_2} t^{\gamma-1} \leq \varphi_\gamma(t_2) - \varphi_\gamma(t_1) \leq \sup_{t_1 \leq t \leq t_2} t^{\gamma-1}.$$

The desired bounds then follow from an explicit computation of the infimum and the maximum in each of the two cases $\gamma \geq 1$ and $\gamma \leq 1$.

For (iv), we first consider the case $\kappa \neq 1$ and $a > 0$. Then

$$\begin{aligned} \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp(a \varphi_{1-\kappa}(t-1)) &= e^{-\frac{a}{1-\kappa}} \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right) \\ &\leq e^{-\frac{a}{1-\kappa}} \max_{t_1 \leq t \leq t_2} \left(\frac{(t+1)^\kappa}{(t+1)^\beta} \right) \sum_{t=t_1}^{t_2} (t+1)^{-\kappa} \exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right) \\ &\leq e^{-\frac{a}{1-\kappa}} (t_2+1)^{\max\{\kappa-\beta, 0\}} \sum_{t=t_1}^{t_2} (t+1)^{-\kappa} \exp\left(\frac{at^{1-\kappa}}{1-\kappa}\right). \end{aligned}$$

Since, for $x \geq 0$, $x \mapsto (x+1)^{-\kappa}$ is decreasing and $x \mapsto \exp(ax^{1-\kappa}/(1-\kappa))$ is increasing, we have that for $t_1 \leq t \leq t_2$ and $x \in [t, t+1]$,

$$(t+1)^{-\kappa} \leq x^{-\kappa} \quad \text{and} \quad \exp(at^{1-\kappa}/(1-\kappa)) \leq \exp(ax^{1-\kappa}/(1-\kappa)).$$

This implies that

$$\begin{aligned} \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp(a \varphi_{1-\kappa}(t-1)) &\leq (t_2+1)^{\max\{\kappa-\beta, 0\}} e^{-\frac{a}{1-\kappa}} \sum_{t=t_1}^{t_2} \int_t^{t+1} x^{-\kappa} \exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right) dx \\ &= (t_2+1)^{\max\{\kappa-\beta, 0\}} e^{-\frac{a}{1-\kappa}} \int_{t_1}^{t_2+1} x^{-\kappa} \exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right) dx \\ &\leq \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{a} e^{-\frac{a}{1-\kappa}} e^{\frac{a(t_2+1)^{1-\kappa}}{1-\kappa}} \\ &= \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{a} \exp(a \varphi_{1-\kappa}(t_2+1)). \end{aligned}$$

The main difference in the case $\kappa \neq 1$ and $a < 0$ is that we now use $x \mapsto \exp(a(x+1)^{1-\kappa}/(1-\kappa))$ is decreasing to obtain, for $t_1 \leq t \leq t_2$ and $x \in [t, t+1]$,

$$\exp(a(t+1)^{1-\kappa}/(1-\kappa)) \leq \exp(ax^{1-\kappa}/(1-\kappa)).$$

A similar argument then yields

$$\begin{aligned} \sum_{t=t_1}^{t_2} (t+1)^{-\beta} \exp(a \varphi_{1-\kappa}(t)) &\leq (t_2+1)^{\max\{\kappa-\beta, 0\}} e^{-\frac{a}{1-\kappa}} \sum_{t=t_1}^{t_2} (t+1)^{-\kappa} \exp\left(\frac{a(t+1)^{1-\kappa}}{1-\kappa}\right) \\ &\leq (t_2+1)^{\max\{\kappa-\beta, 0\}} e^{-\frac{a}{1-\kappa}} \int_{t_1}^{t_2+1} x^{-\kappa} \exp\left(\frac{ax^{1-\kappa}}{1-\kappa}\right) dx \end{aligned}$$

$$\begin{aligned}
&= \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{a} (\exp(a\varphi_{1-\kappa}(t_2+1)) - \exp(a\varphi_{1-\kappa}(t_1))) \\
&\leq \frac{(t_2+1)^{\max\{\kappa-\beta, 0\}}}{(-a)} \exp(a\varphi_{1-\kappa}(t_1)).
\end{aligned}$$

□

We also need the following lemma, which is useful for controlling the accumulation of errors from the noise terms over iterations.

Lemma IV.C.3. *For any $a, b \geq 0$, $T, N \in \mathbb{N}$ and $\kappa, \beta \geq 0$ such that $bt^{-\beta} - at^{-\kappa} \leq 1$ for all $1 \leq t \leq T$, we have that*

$$\prod_{t=1}^T (1 - bt^{-\beta} + at^{-\kappa})^N \leq \exp(-bN\varphi_{1-\beta}(T+1) +aN\varphi_{1-\kappa}(T+1)).$$

Moreover, for any $\zeta \geq 0$, we have that

$$\begin{aligned}
\sum_{t=1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^j \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N \\
\leq Q_1 + \exp\left(-\frac{bN}{2}\varphi_{1-\beta}(T+1) + 4aN\varphi_{1-\kappa}(T+1)\right) Q_2,
\end{aligned}$$

where

$$Q_1 := \begin{cases} \frac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta, 0\}}}{b} \exp\left(\frac{bN}{2(1-\beta)(T+1)^\beta}\right) & \text{if } \beta \neq 1, b > 0, \\ 2N\varphi_{1-\zeta+bN/2}(T+1) \exp\left(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\right) & \text{if } \beta = 1 \text{ or } b = 0, \end{cases}$$

and

$$Q_2 := \begin{cases} \frac{3^\zeta(1+a)^{N-1}}{2a}(T+2)^{\max\{\kappa-\zeta, 0\}} & \text{if } \kappa \neq 1 \text{ and } a > 0, \\ 2N(1+a)^{N-1}\varphi_{1-\zeta-2aN}(T+1) & \text{if } \kappa = 1 \text{ or } a = 0. \end{cases}$$

In the special case $\zeta = \beta = 1 < \kappa$, we have

$$\sum_{t=1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N$$

$$\leq \frac{4}{b} + \frac{3N(1+a)^{N-1} e^{\frac{4aN}{\kappa-1}} \log(T+1)}{(T+1)^{\frac{bN}{2}}}.$$

Proof of Lemma IV.C.3. By assumption, $bt^{-\beta} - at^{-\kappa} \leq 1$ for all $1 \leq t \leq T$. Since $0 \leq 1-x \leq e^{-x}$ for all $x \leq 1$, we have that for any $1 \leq t_1 \leq t_2 \leq T$,

$$\prod_{t=t_1}^{t_2} (1 - bt^{-\beta} + at^{-\kappa})^N \leq \exp \left(-bN \sum_{t=t_1}^{t_2} t^{-\beta} + aN \sum_{t=t_1}^{t_2} t^{-\kappa} \right). \quad (\text{IV.13})$$

Applying this to the first quantity of interest followed by noting that $a, b \geq 0$ and using Lemma IV.C.2(ii), we obtain the first bound that

$$\begin{aligned} \prod_{t=1}^T (1 - bt^{-\beta} + at^{-\kappa})^N &\leq \exp \left(-bN \sum_{t=1}^T t^{-\beta} + aN \sum_{t=1}^T t^{-\kappa} \right) \\ &\leq \exp(-bN \varphi_{1-\beta}(T+1) + aN \varphi_{1-\kappa}(T+1)). \end{aligned}$$

For the second bound, we define

$$t_0 := \sup \left\{ t \leq T \mid \frac{b}{2} \leq at^{-(\kappa-\beta)} \right\}.$$

Then by noting that $1 - bt^{-\beta} + at^{-\kappa} \geq 0$ for all $1 \leq t \leq T$ again, we can bound the quantity of interest as

$$\begin{aligned} &\sum_{t=1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N \\ &= \sum_{t=t_0+1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N \\ &\quad + \left(\prod_{s=t_0+1}^T (1 - bs^{-\beta} + as^{-\kappa}) \right) \\ &\quad \times \left(\sum_{t=1}^{t_0} t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{t_0} (1 - bs^{-\beta} + as^{-\kappa})^N \right) \\ &\leq \sum_{t=t_0+1}^T t^{-\zeta} \left(\sum_{j=1}^N \left(1 - \frac{b}{2} t^{-\beta} \right)^{j-1} \right) \prod_{s=t+1}^T \left(1 - \frac{b}{2} s^{-\beta} \right)^N \\ &\quad + \left(\prod_{s=t_0+1}^T \left(1 - \frac{b}{2} s^{-\beta} \right)^N \right) \left(\sum_{t=1}^{t_0} t^{-\zeta} \left(\sum_{j=1}^N (1 + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^{t_0} (1 + as^{-\kappa})^N \right) \\ &\leq N \times \underbrace{\sum_{t=t_0+1}^T t^{-\zeta} \prod_{s=t+1}^T \left(1 - \frac{b}{2} s^{-\beta} \right)^N}_{=: S_1} \\ &\quad + N(1+a)^{N-1} \times \underbrace{\left(\prod_{s=t_0+1}^T \left(1 - \frac{b}{2} s^{-\beta} \right)^N \right)}_{=: S_3} \times \underbrace{\left(\sum_{t=1}^{t_0} t^{-\zeta} \prod_{s=t+1}^{t_0} (1 + as^{-\kappa})^N \right)}_{=: S_2}. \end{aligned}$$

In the last line, we have used that $0 \leq 1 - \frac{b}{2}t^{-\beta} \leq 1$ for $t \geq t_0 + 1$ and $1 + at^{-\kappa} \leq 1 + a$.

To control the three quantities, we first note that by (IV.13), we have

$$\begin{aligned} S_3 &\leq \exp\left(-\frac{bN}{2}\sum_{s=1}^T s^{-\beta}\right) \exp\left(\frac{bN}{2}\sum_{s=1}^{t_0} s^{-\beta}\right) \\ &\stackrel{(a)}{\leq} \exp\left(-\frac{bN}{2}\sum_{s=1}^T s^{-\beta}\right) \exp\left(aN\sum_{s=1}^{t_0} s^{-\kappa}\right) \\ &\stackrel{(b)}{\leq} \exp\left(-\frac{bN}{2}\varphi_{1-\beta}(T+1) + 2aN\varphi_{1-\kappa}(T+1)\right). \end{aligned}$$

In (a) above, we have noted that $\frac{b}{2} \leq as^{-(\kappa-\beta)}$ for $s \leq t_0$; in (b), we have used $t_0 \leq T$ and Lemma IV.C.2(ii) with $a, b \geq 0$. In the special case $\beta = 1 < \kappa$, the above yields

$$\begin{aligned} S_3 &\leq (T+1)^{-\frac{bN}{2}} \exp\left(2aN\frac{1-(T+1)^{-(\kappa-1)}}{\kappa-1}\right) \\ &\leq (T+1)^{-\frac{bN}{2}} e^{\frac{2aN}{\kappa-1}}. \end{aligned}$$

We now control S_2 . By (IV.13) again, we have

$$\begin{aligned} S_2 &\leq \sum_{t=1}^{t_0} t^{-\zeta} \exp\left(aN\sum_{s=t+1}^{t_0} s^{-\kappa}\right) \\ &\leq \sum_{t=1}^T t^{-\zeta} \exp\left(aN\sum_{s=t+1}^T s^{-\kappa}\right) \\ &\stackrel{(c)}{\leq} \exp(2aN\varphi_{1-\kappa}(T+1)) \times \left(\sum_{t=1}^T t^{-\zeta} \exp(-2aN\varphi_{1-\kappa}(t+1))\right) \\ &\stackrel{(d)}{\leq} 3^\zeta \exp(2aN\varphi_{1-\kappa}(T+1)) \sum_{t=1}^T (t+2)^{-\zeta} \exp(-2aN\varphi_{1-\kappa}(t+1)). \end{aligned}$$

In (c) above, we have applied Lemma IV.C.2(ii); in (d), we have noted that $\sup_{t \in \mathbb{N}} (t+2)^\beta/t^\beta = 3^\beta$. If $\kappa \neq 1$ and $a > 0$, we can apply Lemma IV.C.2(iv) to get that

$$\begin{aligned} S_2 &\leq \frac{3^\zeta}{2aN} (T+2)^{\max\{\kappa-\zeta, 0\}} \exp(2aN\varphi_{1-\kappa}(T+1)) \\ &= \frac{Q_2}{N(1+a+c)^{N-1}} \exp(2aN\varphi_{1-\kappa}(T+1)). \end{aligned}$$

If $\kappa = 1$ or $a = 0$, the bound from (c) above reads

$$\begin{aligned} S_2 &\leq \exp(2aN\varphi_{1-\kappa}(T+1)) \sum_{t=1}^T t^{-\zeta} (t+1)^{-2aN} \\ &\leq \exp(2aN\varphi_{1-\kappa}(T+1)) \sum_{t=1}^T t^{-\zeta-2aN} \\ &\leq 2\varphi_{1-\zeta-2aN}(T+1) \exp(2aN\varphi_{1-\kappa}(T+1)) = \frac{\mathcal{Q}_2}{N(1+a)^{N-1}} \exp(2aN\varphi_{1-\kappa}(T+1)), \end{aligned}$$

where we have used Lemma IV.C.2(ii) in the last line. Now consider the special case with $\zeta = 1 < \kappa$, the bound from (d) becomes

$$\begin{aligned} S_2 &\leq 3 \exp(2aN\varphi_{1-\kappa}(T+1)) \left(\sum_{t=1}^T (t+2)^{-1} \exp(-2aN\varphi_{1-\kappa}(t+1)) \right) \\ &\leq 3 \exp\left(2aN \frac{1-(T+1)^{-(\kappa-1)}}{\kappa-1}\right) \sum_{t=1}^T (t+2)^{-1} \\ &\leq 3e^{\frac{2aN}{\kappa-1}} \log(T+1). \end{aligned}$$

We are left with controlling S_1 , which follows from a similar strategy as controlling S_2 :

$$\begin{aligned} S_1 &\stackrel{(IV.13)}{\leq} \sum_{t=t_0+1}^T t^{-\zeta} \exp\left(-\frac{bN}{2} \sum_{s=t+1}^T s^{-\beta}\right) \\ &\leq \sum_{t=1}^T t^{-\zeta} \exp\left(-\frac{bN}{2} \sum_{s=t+1}^T s^{-\beta}\right) \\ &\stackrel{(a)}{\leq} \exp\left(-\frac{bN}{2} \varphi_{1-\beta}(T+1)\right) \sum_{t=1}^T t^{-\zeta} \exp\left(\frac{bN}{2} \varphi_{1-\beta}(t+1)\right) \quad (IV.14) \\ &\leq 4^\zeta \exp\left(-\frac{bN}{2} \varphi_{1-\beta}(T+1)\right) \sum_{t=1}^T (t+3)^{-\zeta} \exp\left(\frac{bN}{2} \varphi_{1-\beta}(t+1)\right). \end{aligned}$$

In (a) above, we used Lemma IV.C.2(ii). For $\beta \neq 1$ and $b \neq 0$, we can apply Lemma IV.C.2(iv) with $\frac{b}{2} > 0$ to obtain

$$\begin{aligned} S_1 &\leq 4^\zeta \exp\left(-\frac{bN}{2} \varphi_{1-\beta}(T+1)\right) \frac{(T+3)^{\max\{\beta-\zeta, 0\}}}{bN/2} \exp\left(\frac{bN}{2} \varphi_{1-\beta}(T+3)\right) \\ &= \frac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta, 0\}}}{bN} \exp\left(\frac{bN}{2(1-\beta)} \left((T+3)^{1-\beta} - (T+1)^{1-\beta}\right)\right) \\ &\stackrel{(b)}{\leq} \frac{2^{2\zeta+1}(T+3)^{\max\{\beta-\zeta, 0\}}}{bN} \exp\left(\frac{bN}{2(1-\beta)(T+1)^\beta}\right) = \frac{\mathcal{Q}_1}{N}. \end{aligned}$$

In (b), we have used Lemma IV.C.2(iii) with $1-\beta \leq 1$. Meanwhile, if $\beta = 1$ or

$b = 0$, we have

$$\begin{aligned} S_1 &\leq \exp\left(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\right) \sum_{t=1}^T t^{-\zeta} (t+1)^{bN/2} \\ &\leq \exp\left(-\frac{b}{2}\varphi_{1-\beta}(T+1)\right) \sum_{t=1}^T t^{-\zeta+bN/2} \\ &\leq 2\varphi_{1-\zeta+bN/2}(T+1) \exp\left(-\frac{bN}{2}\varphi_{1-\beta}(T+1)\right) = \frac{Q_1}{N}. \end{aligned}$$

For the special case with $\zeta = \beta = 1$, the bound in (IV.14) becomes

$$\begin{aligned} S_1 &\leq \exp\left(-\frac{bN}{2}\varphi_0(T+1)\right) \sum_{t=1}^T t^{-1} \exp\left(\frac{bN}{2}\varphi_0(t+1)\right) \\ &= (T+1)^{-\frac{bN}{2}} \sum_{t=1}^T t^{-1} (t+1)^{\frac{bN}{2}} \\ &\leq (T+1)^{-\frac{bN}{2}} \sum_{t=1}^T t^{-(1-\frac{bN}{2})} \\ &\stackrel{(c)}{\leq} 2(T+1)^{-\frac{bN}{2}} \varphi_{\frac{bN}{2}}(T+1) = 2(T+1)^{-\frac{bN}{2}} \frac{(T+1)^{bN/2}-1}{bN/2} \leq \frac{4}{bN}. \end{aligned}$$

In (c), we have used Lemma IV.C.2(ii) for both the case $1 - \frac{bN}{2} \leq 0$ and $1 - \frac{bN}{2} \geq 0$.

Combining the bounds for the general cases, we obtain the first desired inequality that

$$\begin{aligned} &\sum_{t=1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N \\ &\leq Q_1 + \exp\left(-\frac{b}{2}\varphi_{1-\beta}(T+1) + u\varphi_{1-\xi}(T+3) + 4a\varphi_{1-\kappa}(T+1)\right) Q_2. \end{aligned}$$

For the special case $\zeta = \beta = 1 < \kappa, \gamma$, combining the earlier bounds gives

$$\begin{aligned} &\sum_{t=1}^T t^{-\zeta} \left(\sum_{j=1}^N (1 - bt^{-\beta} + at^{-\kappa})^{j-1} \right) \prod_{s=t+1}^T (1 - bs^{-\beta} + as^{-\kappa})^N \\ &\leq \frac{4}{b} + \frac{3N(1+a)^{N-1} e^{\frac{4aN}{\kappa-1}} \log(T+1)}{(T+1)^{\frac{bN}{2}}}. \end{aligned}$$

□

IV.C.2 Contraction and integrability results

The next result is a standard result in convex analysis, needed to handle projections performed in Algorithms 1 and 2.

Lemma IV.C.4. Let Ψ be convex subset of \mathbb{R}^p . Let $\psi^* \in \Psi$. Then, for all $\psi \in \mathbb{R}^p$, we have:

$$\|\text{Proj}_\Psi(\psi) - \psi^*\| \leq \|\psi - \psi^*\|$$

Proof. We have:

$$\begin{aligned} \|\psi - \psi^*\|^2 &= \|\psi - \text{Proj}_\Psi(\psi) + \text{Proj}_\Psi(\psi) - \psi^*\|^2 \\ &= \|\psi - \text{Proj}_\Psi(\psi)\|^2 + 2\langle \psi - \text{Proj}_\Psi(\psi), \text{Proj}_\Psi(\psi) - \psi^* \rangle + \|\text{Proj}_\Psi(\psi) - \psi^*\|^2 \end{aligned}$$

Since by [21, Proposition 1.1.9], we have:

$$\langle \psi - \text{Proj}_\Psi(\psi), \psi' - \text{Proj}_\Psi(\psi) \rangle \leq 0$$

for all $\psi' \in \Psi$, we can use this inequality at $\psi' = \psi^* \in \Psi$ to obtain:

$$\|\psi - \psi^*\|^2 \geq \|\psi - \text{Proj}_\Psi(\psi)\|^2$$

and the result follows by taking the square root. \square

We now state two lemmas that guarantee an amount of integrability sufficient to our analysis.

Lemma IV.C.5. Let $p, q \in \mathcal{P}(\mathcal{X})$ such that $\frac{dp}{dq}$ exists, and such that $\chi^2(p; q) < +\infty$. Assume that $f \in L^2(q)$. Then $|\int f dp| < +\infty$.

Proof. By assumption, $f \in L^2(q)$. Moreover, $\chi^2(p, q) < +\infty$, and thus we have $\frac{dp}{dq} - 1 \in L^2(q)$. Thus, the inner product is finite, and

$$\begin{aligned} \left| \int f \left(\frac{dp}{dq} - 1 \right) dq \right| &= \left| \int f dp - \int f dq \right| := M < +\infty \\ \implies M - \left| \int f dq \right| &< \int f dp < M + \left| \int f dq \right| \end{aligned}$$

\square

Lemma IV.C.6. For all $\psi \in \Psi$, for all $m \geq 1$, and for all $k \geq 1$, we have:

$$\mathbb{E} \left[\|P_\psi^m \phi(X_1)\|^k \right] < +\infty.$$

Proof. By analyticity of the log partition function $\psi \mapsto \log Z(\psi)$, we have $\int \|\phi\|^k dP_\psi < +\infty$ for all $\psi \in \Psi$, and thus, the function $x \mapsto \|\phi\|^k(x) \in L^2(p_\psi)$ for all ψ . Consequently, $P_\psi^m \|\phi\|^k \in L^2(p_\psi)$. We can apply Lemma IV.C.5 to $P_\psi^m \|\phi\|^k$ to obtain $\mathbb{E} [P_\psi^m \|\phi(X_1)\|^k] < +\infty$ for all $k \geq 1$ and for all $m \geq 0$. As a by-product, we obtain $P_\psi^m \|\phi\|^k \in L^2(p_{\psi^\star})$, and thus so $\|P_\psi^m \phi\|^k$. \square

The following lemma is used multiple times in our analysis.

Lemma IV.C.7. *Assume A2. Let q be a positive integer. Let $f := (f_1, \dots, f_q)$ such that $f_k \in \{\phi_i\}_{i=1}^p \cup \{\phi_i \phi_j\}_{i,j=1}^p$ for $k \in [q]$. Then, for all $\psi \in \Psi$, we have*

$$\left\| \mathbb{E} \left[P_\psi^m (f - \mathbb{E}[f(X^\psi)])(X_1) \right] \right\| \leq \alpha^m C_\chi \left(\mathbb{E} \left[\| (f - \mathbb{E}[f(X^\psi)])(X^\psi) \|^2 \right] \right)^{1/2} \|\psi - \psi^\star\|$$

Proof. Let us note first that

$$\begin{aligned} & \left\| \mathbb{E} P_\psi^m \left[(f - \mathbb{E}[f(X^\psi)])(X_1) \right] \right\|^2 \\ & \stackrel{(a)}{=} \sum_{i=1}^q \left(\int P_\psi^m (f_i - \mathbb{E}[f_i(X^\psi)])(x) (p_{\psi^\star}(x) - p_\psi(x)) dx \right)^2 \\ & = \sum_{i=1}^q \left(\int P_\psi^m (f_i - \mathbb{E}[f_i(X^\psi)])(x) \left(\frac{dp_{\psi^\star}}{dp_\psi}(x) - 1 \right) p_\psi(x) dx \right)^2 \\ & \stackrel{(b)}{\leq} \left(\int \left(\frac{dp_{\psi^\star}}{dp_\psi}(x) - 1 \right)^2 p_\psi(x) dx \right) \sum_{i=1}^q \int P_\psi^m (f_i - \mathbb{E}[f_i(X^\psi)])(x)^2 p_\psi(x) dx \\ & \leq \chi^2(p_\psi, p_{\psi^\star}) \sum_{i=1}^q \|P_\psi^m (f_i - \mathbb{E}[f_i(X^\psi)])\|_{L^2(p_\psi)} \\ & \stackrel{(c)}{\leq} \alpha^{2m} \chi^2(p_\psi, p_{\psi^\star}) \sum_{i=1}^q \|f_i - \mathbb{E}[f_i(X^\psi)]\|_{L^2(p_\psi)(\mathbb{R}^d)} \\ & \stackrel{(d)}{\leq} \alpha^{2m} C_\chi^2 \|\psi - \psi^\star\|^2 \mathbb{E} [\|(f - \mathbb{E}[f(X^\psi)])(X^\psi)\|^2]. \end{aligned}$$

Here, we used the fact that P_ψ^m admits p_ψ as an invariant measure in (a) [16, Eq. (1.2.2)], the Cauchy-Schwarz inequality in (b) \square

IV.C.3 Miscellaneous

Lemma IV.C.8. *Let $f : \Psi \rightarrow \mathbb{R}^p$ be a differentiable function in the interior of $\Psi \subseteq \mathbb{R}^p$. For $\psi \in \Psi$, define $\sigma_{\min}(\psi) := \inf_{\theta \in \Psi, \|\theta\|=1} \theta^\top \nabla f(\psi) \theta$ and $\sigma_{\max}(\psi) :=$*

$\sup_{\theta \in \Psi, \|\theta\|=1} \theta^\top \nabla f(\psi) \theta$ with respect to the Jacobian matrix $\nabla f(\psi)$. Then for any $\psi_1, \psi_2 \in \Psi$, we have that

$$\inf_{\psi \in \Psi} \sigma_{\min}(\psi) \leq (\psi_1 - \psi_2)^\top (f(\psi_1) - f(\psi_2)) \leq \sup_{\psi \in \Psi} \sigma_{\max}(\psi)$$

Proof of Lemma IV.C.8. By the mean value theorem, there exists some $a \in (0, 1)$ such that

$$\begin{aligned} (\psi_1 - \psi_2)^\top (f(\psi_1) - f(\psi_2)) &= (\psi_1 - \psi_2)^\top (f(1 \times \psi_1 + 0 \times \psi_2) - f(0 \times \psi_1 + 0 \times \psi_2)) \\ &= (\psi_1 - \psi_2)^\top \nabla f(a\psi_1 + (1-a)\psi_2)(\psi_1 - \psi_2) \\ &= \|\psi_1 - \psi_2\|^2 \frac{(\psi_1 - \psi_2)^\top}{\|\psi_1 - \psi_2\|} \nabla f(a\psi_1 + (1-a)\psi_2) \frac{\psi_1 - \psi_2}{\|\psi_1 - \psi_2\|}. \end{aligned}$$

Plugging in the definition of σ_{\max} gives the desired upper bound and similarly σ_{\min} implies the lower bound. \square

IV.D Proofs for Online CD

IV.D.1 Auxiliary Lemmas for Online CD

We recall the following notations: $\sigma_\psi := \mathbb{E}[\|(\phi - \mathbb{E}[\phi(X^\psi)])(X^\psi)\|^2]$, as well as $\sigma_* := \sigma_{\psi^*}$ and $\sigma := \sup_{\psi \in \Psi} \sigma_\psi$.

We now provide two intermediary lemmas necessary to analyze the impact of variance in the CD gradient. The strategy is similar in both of them: we change the integration from p_{ψ^*} to p_ψ to obtain contraction, at the cost of an additional term scaling with $C_\chi \|\psi - \psi^*\|$. We obtain exact constants that we choose to describe in terms of the smoothness parameters of the problem, e.g. the k^{th} derivatives of the log partition function $\log Z$, which, for $k \geq 2$, equals the k^{th} derivative of the negative cross-entropy model w.r.t p_{ψ^*} .

Second Moment convergence The following lemmas guarantee the second moment of a sample from $k_\psi^m p_{\psi^*}$ approaches the second moment of a sample from the target distribution p_ψ .

Lemma IV.D.1. Under A0, A1 and A2, for all $\psi \in \Psi$, we have:

$$\left| \mathbb{E} \left[P_\psi^m \|\phi(X_1)\|^2 \right] - \mathbb{E} \|\phi(X^\psi)\|^2 \right| \leq \alpha^m C_\chi \|\psi - \psi^*\| \|\log Z\|_{1,\infty},$$

where

$$\begin{aligned} \|\log Z\|_{1,\infty} &:= \sup_{\psi \in \Psi} \sum_{i=1}^p (4\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 2\partial_i^2 \log Z(\psi)^2 \\ &\quad + 4\partial_i^1 \log Z(\psi) \partial_i^3 \log Z(\psi) + \partial_i^4 \log Z(\psi))^{1/2} < +\infty \end{aligned}$$

Proof. Applying Lemma IV.C.7 to each $f_i := \phi_i^2$, we have

$$\begin{aligned} &\left| \mathbb{E} \left[P_\psi^m \phi_i(X_1)^2 \right] - \mathbb{E} [\phi_i(X_i^\psi)^2] \right| \\ &= \alpha^m C_\chi \|\psi - \psi^*\| \left(\mathbb{E} \left[(\phi_i(X^\psi)^2 - \mathbb{E} [\phi_i(X^\psi)^2])^2 \right] \right)^{1/2} \\ &= \alpha^m C_\chi \|\psi - \psi^*\| \left(\mathbb{E} [\phi_i(X^\psi)^4] - (\mathbb{E} [\phi_i^2(X^\psi)])^2 \right)^{1/2} \end{aligned}$$

We map the two moments to derivatives of $\log Z(\psi)$, since the k^{th} derivative of $\log Z(\psi)$ is the k^{th} cumulant. It can be shown, using the multivariate moment to cumulant mapping, that

$$\begin{aligned} &\mathbb{E} [\phi_i(X^\psi)^4] \\ &= \frac{\partial \log Z}{\partial \psi_i}^4 + 6 \frac{\partial \log Z}{\partial \psi_i}^2 \frac{\partial^2 \log Z}{\partial \psi_i^2} + 3 \left(\frac{\partial^2 \log Z}{\partial \psi_i^2} \right)^2 + 4 \frac{\partial \log Z}{\partial \psi_i} \frac{\partial^3 \log Z}{\partial \psi_i^3} + \frac{\partial^4 \log Z}{\partial \psi_i^4} \\ &= \partial_i^1 \log Z(\psi)^4 + 6\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 3\partial_i^2 \log Z(\psi)^2 \\ &\quad + 4\partial_i^1 \log Z(\psi) \partial_i^3 \log Z(\psi) + \partial_i^4 \log Z(\psi) \end{aligned}$$

where $\partial_i^k \log Z(\psi)$ denotes the k^{th} derivative of $\log Z$ with respect to ψ_i . On the other hand,

$$\begin{aligned} &\mathbb{E} [\phi_i(X_i^\psi)^2] = \partial_i^1 \log Z(\psi)^2 + \partial_i^2 \log Z(\psi) \\ \implies &(\mathbb{E} [\phi_i(X^\psi)^2])^2 = \partial_i^1 \log Z(\psi)^4 + 2\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + \partial_i^2 \log Z(\psi)^2 \end{aligned}$$

implying

$$\begin{aligned} & \mathbb{E} [\phi_i(X^\psi)^4] - (\mathbb{E} \phi_i(X^\psi))^2 \\ &= 4\partial_i^1 \log Z(\psi)^2 \partial_i^2 \log Z(\psi) + 2\partial_i^2 \log Z(\psi)^2 + 4\partial_i^1 \log Z(\psi) \partial_i^3 \log Z(\psi) \\ &+ \partial_i^4 \log Z(\psi) \end{aligned}$$

The result follows by summing over i , since:

$$\left| \mathbb{E} \left[P_\psi^m \|\phi(X_1)\|^2 \right] - \mathbb{E} [\|\phi(X^\psi)\|^2] \right| \leq \sum_{i=1}^d \left| \mathbb{E} \left[P_\psi^m \phi_i(X_1)^2 \right] - \mathbb{E} [\phi_i(X^\psi)^2] \right|$$

Note that $\|\log Z\|_{1,\infty}$ is finite since Ψ is compact and $\log Z$ is analytic. \square

Squared First Moment convergence The next lemma provides convergence (in squared absolute value) of the first moment of the m -iterated Markov kernel k_ψ^m .

Lemma IV.D.2. *Under A0, A1 and A2, for all $\psi \in \Psi$, we have*

$$\left| \mathbb{E} \left[\|\mathbf{P}_\psi^m \phi(X_1)\|^2 \right] - \mathbb{E} [\phi(X^\psi)] \right|^2 \leq \alpha^{2m} \sigma_\psi^2 + C_\chi \alpha^{m/2} \|\log Z\|_{2,\infty} \|\psi - \psi^*\|$$

where

$$\|\log Z\|_{2,\infty} := \sup_{\psi \in \Psi} \sum_{i=1}^p (F(\psi) \partial_i^2 \log Z(\psi))^{1/4} + 2 \left| \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \right|$$

and

$$F(\psi)$$

$$:= 15\partial_i^2 \log Z(\psi)^3 + 10\partial_i^3 \log Z(\psi)^2 + 15\partial_i^2 \log Z(\psi) \partial_i^4 \log Z(\psi) + \partial_i^6 \log Z(\psi)$$

Proof. We have:

$$\begin{aligned} \mathbb{E} [(P_\psi^m \phi_i(X_1))^2] &= \mathbb{E} [(P_\psi^m \phi_i(X_1) - \mathbb{E} [\phi_i(X^\psi)] + \mathbb{E} [\phi_i(X^\psi)])^2] \\ &= \mathbb{E} [(P_\psi^m \phi_i(X_1) - \mathbb{E} [\phi_i(X^\psi)])^2] \\ &+ 2 \times \mathbb{E} [\phi_i(X^\psi)] \mathbb{E} [P_\psi^m (\phi_i - \mathbb{E} [\phi_i(X^\psi)]) (X_1)] + (\mathbb{E} [\phi_i(X^\psi)])^2 \end{aligned}$$

Implying

$$\begin{aligned}
|\mathbb{E}[(P_\psi^m \phi_i(X_1))^2] - (\mathbb{E}[\phi_i(X^\psi)])^2| &\leq \underbrace{\left| \mathbb{E} \left[(P^m(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X_1))^2 \right] \right|}_{\Delta_1} \\
&+ 2 \underbrace{|\mathbb{E}[\phi_i(X^\psi)] \mathbb{E}[P^m(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X_1)]|}_{\Delta_2}
\end{aligned} \tag{IV.15}$$

where

$$\begin{aligned}
\Delta_1 &= \underbrace{\mathbb{E} \left[\left(P_\psi^m (\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi) \right)^2 \right]}_{\Delta_{1,1}} \\
&+ \underbrace{\mathbb{E} \left[\left(\left(P_\psi^m \phi_i - \mathbb{E}[\phi_i(X^\psi)] \right)^2 (X^\psi) \left(\frac{dp_{\psi^*}}{dp_\psi}(X^\psi) - 1 \right) \right) \right]}_{\Delta_{1,2}} \\
&\stackrel{(a)}{\leq} \alpha^{2m} \mathbb{E} [(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] \\
&+ C_\chi \|\psi - \psi^*\| (\mathbb{E} [(P^m(\phi_i - \mathbb{E}[\phi_i(X^\psi)]))(X^\psi)^4])^{1/2} \\
&\stackrel{(b)}{\leq} \alpha^{2m} \mathbb{E} [(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] \\
&+ C_\chi \|\psi - \psi^*\| \\
&\times \left(\mathbb{E} [(P^m(\phi_i - \mathbb{E}[\phi_i(X^\psi)]))(X^\psi)^6] \right)^{1/4} (\mathbb{E} [(P^m(\phi_i - \mathbb{E}[\phi_i(X^\psi)]))(X^\psi)^2])^{1/4} \\
&\stackrel{(c)}{\leq} \alpha^{2m} \mathbb{E} [(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] \\
&+ \alpha^{\frac{m}{2}} C_\chi \|\psi - \psi^*\| \left(\mathbb{E} [(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6] \right)^{1/4} \\
&\times (\mathbb{E} [(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2])^{1/4}
\end{aligned}$$

In (a), we used the restricted spectral gap Assumption A2 for $\Delta_{1,1}$, and the Cauchy-Schwarz inequality combined with Assumption A1 for $\Delta_{1,2}$. In (b), we used Cauchy-Schwarz once again, and in (c) we used the fact that P_ψ^m is a contraction in $L^6(p_\psi)$ and another invocation of the spectral gap assumption A2. As an aside, note that a simpler result can be obtained by making regularity assumption on the mapping $\psi \mapsto P_\psi^m$. Assuming that $\psi \mapsto P_\psi^m(x)$ is uniformly L_m -Lipschitz across $x \in \mathcal{X}$ for instance (as done in [121, Assumption 5]), the second term $\Delta_{1,2}$ of Δ_1 could have

been handled using

$$\begin{aligned}\Delta_{1,2} &\leq 2\mathbb{E} \|\mathbf{P}_{\psi}^m \phi(X_1) - \mathbb{E}[\phi(X_1)]\|^2 + 2 \|\mathbb{E}[\phi(X^\psi)] - \mathbb{E}[\phi(X_1)]\|^2 \\ &\leq 4(L_m \|\psi - \psi^*\| + \sigma_*^2 \alpha^{2m} + 2 \|\mathbb{E}[\phi(X^\psi)] - \mathbb{E}[\phi(X_1)]\|^2) \\ &\leq 4(L_m \|\psi - \psi^*\| + \sigma_*^2 \alpha^{2m} + 2L^2 \|\psi - \psi^*\|)\end{aligned}$$

Although this result does not require possibly large constants related to sixth-order moments, it is less tight in the sense that it does not go to 0 as $m \rightarrow \infty$. Back to the main proof, and Δ_2 in particular. Applying Lemma IV.C.7 to $f := \phi$, we have

$$\begin{aligned}\Delta_2 &\leq \alpha^m \mathbb{E}[\phi_i(X^\psi)] C_\chi \|\psi - \psi^*\| (\mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2])^{1/2} \\ &\leq \alpha^m C_\chi \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \|\psi - \psi^*\|\end{aligned}$$

Putting everything together, we have:

$$\begin{aligned}|\mathbb{E}_{p_{\psi^*}}(P_{\psi^*}^m \phi_i)^2 - \mathbb{E}_{p_\psi} \phi_i^2| &\leq \alpha^{2m} \mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] + C_\chi \|\psi - \psi^*\| \times \\ &\quad \left(\alpha^{\frac{m}{2}} \left(\mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6] \right)^{1/4} (\mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2])^{1/4} \right. \\ &\quad \left. + 2\alpha^m \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \right) \\ &\leq \alpha^{2m} \mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] + C_\chi \alpha^{m/2} \|\psi - \psi^*\| \\ &\quad \left(\left(\mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6] \right)^{1/4} (\mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2])^{1/4} \right. \\ &\quad \left. + 2 \left| \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \right| \right)\end{aligned}$$

Summing over i , we obtain

$$\begin{aligned}& \left| \mathbb{E} \left[\|\mathbf{P}_{\psi}^m \phi(X_1)\|^2 \right] - (\mathbb{E}[\phi(X^\psi)])^2 \right| \\ &\leq \sum_{i=1}^p \left| \mathbb{E} \left[(P_{\psi}^m \phi_i(X_1))^2 \right] - (\mathbb{E}[\phi_i(X^\psi)])^2 \right| \\ &\leq \alpha^{2m} \sum_{i=1}^p \mathbb{E}[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2] + C_\chi \alpha^{m/2} \|\log Z\|_{2,\infty} \|\psi - \psi^*\| \\ &\leq \alpha^{2m} \sigma_\psi^2 + C_\chi \alpha^{m/2} \|\log Z\|_{2,\infty} \|\psi - \psi^*\|\end{aligned}$$

where

$$\begin{aligned} & \|\log Z\|_{2,\infty} \\ &= \sup_{\psi \in \Psi} \sum_{i=1}^p \left(\mathbb{E} \left[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6 \right] \right)^{1/4} \left(\mathbb{E} \left[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^2 \right] \right)^{1/4} \\ &\quad + 2 \left| \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \right| \end{aligned}$$

Similarly to the previous lemma, one can upper bound $\mathbb{E} \left[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6 \right]$ using the *centered* moment to cumulant formula:

$$\begin{aligned} & \mathbb{E} \left[(\phi_i - \mathbb{E}[\phi_i(X^\psi)])(X^\psi)^6 \right] \\ &= 15\partial_i^2 \log Z(\psi)^3 + 10\partial_i^3 \log Z(\psi)^2 + 15\partial_i^2 \log Z(\psi) \partial_i^4 \log Z(\psi) + \partial_i^6 \log Z(\psi) \\ &=: F(\psi) \end{aligned}$$

To get a full description of $\|\log Z\|_{2,\infty}$:

$$\begin{aligned} & \|\log Z\|_{2,\infty} \\ &= \sup_{\psi \in \Psi} \sum_{i=1}^p \left(F(\psi) \partial_i^2 \log Z(\psi) \right)^{1/4} + 2 \left| \partial_i^1 \log Z(\psi) \partial_i^2 \log Z(\psi)^{1/2} \right|. \end{aligned}$$

□

We can now use the previous lemmas to obtain an expression on the second moment of the contrastive divergence gradient estimator, relating it to the one of the stochastic log-likelihood gradient estimator.

Lemma IV.D.3. *Under A0, A1 and A2, we have:*

$$\begin{aligned} & \mathbb{E} \|h_t(\psi, X_t)\|^2 \\ & \leq 2\sigma_*^2 + 2\sigma_\psi^2 + 2L^2 \|\psi - \psi^*\|^2 + 4(\sigma_\psi^2 \alpha^{2m} + \alpha^{m/2} \|\log Z\|_{3,\infty} C_\chi \|\psi - \psi^*\|) \end{aligned}$$

where $\|\log Z\|_{3,\infty} := 2 \max(\|\log Z\|_{1,\infty}, \|\log Z\|_{2,\infty})$.

Proof. We rely on the following decomposition:

$$\begin{aligned} h_t(\psi, X_t) &= \underbrace{(\phi(X_t) - \mathbb{E}[\phi(X_1)])}_{\Delta_{1,1}} + \underbrace{(\mathbb{E}[\phi(X_1)] - \mathbb{E}[\phi(X^\psi)])}_{\Delta_{1,2}} \\ &\quad + \underbrace{(\mathbb{E}[P_\psi^m \phi(X_t) | X_t] - \phi(K_\psi^m(X_t)))}_{\Delta_2} + \underbrace{(\mathbb{E}[\phi(X^\psi)] - \mathbb{E}[P_\psi^m \phi(X_t), | X_t])}_{\Delta_3} \end{aligned}$$

$\Delta_{1,1} + \Delta_{1,2}$ form the differentiable stochastic gradient g_t of Equation IV.6. Note that $\Delta_{1,1}$ is mean-zero, and Δ_2 is mean-zero conditionally on X_t . Consequently, $\mathbb{E}\langle \Delta_2, \Delta_3 \rangle = \mathbb{E}\langle \Delta_2, \Delta_{1,1} \rangle = \mathbb{E}\langle \Delta_{1,1}, \Delta_{1,2} \rangle = \mathbb{E}\langle \Delta_{1,1}, \Delta_2 \rangle = 0$, and the only mixed-terms that remain to be controlled are $\mathbb{E}\langle \Delta_{1,1}, \Delta_3 \rangle$ and $\mathbb{E}\langle \Delta_{1,2}, \Delta_3 \rangle$. We first control the unmixed terms, and the simple ones first: we have $\mathbb{E}\|\Delta_{1,1}\|^2 = \sigma_*^2$, as well as $\mathbb{E}\|\Delta_{1,2}\|^2 \leq L^2 \|\psi - \psi^*\|^2$. For Δ_2 , we have:

$$\mathbb{E}[\|\Delta_2\|^2 | X_t] = P_\psi^m \|\phi(X_t)\|^2 - \|P_\psi^m \phi(X_t)\|^2$$

We can invoke Lemmas IV.D.2 and IV.D.1, which guarantee

$$\begin{aligned} \left| \mathbb{E}[\|P_\psi^m \phi(X_1)\|^2] - \|\mathbb{E}[\phi(X^\psi)]\|^2 \right| &\leq \alpha^{2m} \sigma_\psi^2 + C_\chi \alpha^{m/2} \|\log Z\|_{2,\infty} \|\psi - \psi^*\| \\ \left| \mathbb{E}[P_\psi^m \|\phi(X_1)\|^2] - \mathbb{E}[\|\phi(X^\psi)\|^2] \right| &\leq \alpha^m C_\chi \|\log Z\|_{1,\infty} \|\psi - \psi^*\| \end{aligned}$$

to obtain

$$\begin{aligned} \mathbb{E}[\|\Delta_2\|^2] &= \mathbb{E}[\|\phi(X^\psi)\|^2] - \|\mathbb{E}[\phi(X^\psi)]\|^2 + \\ &\quad \alpha^{2m} \sigma_\psi^2 + C_\chi \alpha^{m/2} (\|\log Z\|_{1,\infty} + \|\log Z\|_{2,\infty}) \|\psi - \psi^*\| \\ &= \mathbb{E}[\|\phi(X^\psi)\|^2] - \|\mathbb{E}[\phi(X^\psi)]\|^2 + \alpha^{2m} \sigma_\psi^2 \\ &\quad + \alpha^{m/2} \|\log Z\|_{3,\infty} C_\chi \|\psi - \psi^*\| \end{aligned}$$

where $\|\log Z\|_{3,\infty} := 2 \max(\|\log Z\|_{1,\infty}, \|\log Z\|_{2,\infty})$.

For Δ_3 , notice that Δ_3 is precisely the term Δ_1 in Lemma IV.D.2, and we can thus

bound it by

$$\mathbb{E}[\|\Delta_3\|^2] \leq \sigma_\psi^2 \alpha^{2m} + \alpha^{m/2} \|\log Z\|_{2,\infty} C_\chi \|\psi - \psi^*\|$$

Finally, we use the simple bound $2\mathbb{E}[\langle \Delta_{1,1}, \Delta_3 \rangle]$ by $\mathbb{E}[\|\Delta_{1,1}\|^2] + \mathbb{E}[\|\Delta_3\|^2]$, and $2\mathbb{E}[\langle \Delta_{1,2}, \Delta_3 \rangle]$ by $\mathbb{E}[\|\Delta_{1,2}\|^2] + \mathbb{E}[\|\Delta_3\|^2]$. Putting everything together, we have:

$$\begin{aligned} & \mathbb{E}[\|h_t(\psi)\|^2] \\ &= \mathbb{E}[\|\Delta_{1,1}\|^2] + \mathbb{E}[\|\Delta_{1,2}\|^2] + \mathbb{E}[\|\Delta_2\|^2] + \mathbb{E}[\|\Delta_3\|^2] \\ &\quad + 2\mathbb{E}[\langle \Delta_{1,1}, \Delta_3 \rangle] + 2\mathbb{E}[\langle \Delta_{1,2}, \Delta_3 \rangle] \\ &\leq 2\mathbb{E}[\|\Delta_{1,1}\|^2] + 2\mathbb{E}[\|\Delta_{1,2}\|^2] + \mathbb{E}[\|\Delta_2\|^2] + 3\mathbb{E}[\|\Delta_3\|^2] \\ &\leq 2\sigma_*^2 + 2L^2 \|\psi - \psi^*\|^2 + \sigma_\psi^2 + 4(\sigma_\psi^2 \alpha^{2m} + \alpha^{m/2} \|\log Z\|_{3,\infty} C_\chi \|\psi - \psi^*\|) \\ &\leq 2\sigma_*^2 + 2\sigma_\psi^2 + 2L^2 \|\psi - \psi^*\|^2 + 4(\sigma_\psi^2 \alpha^{2m} + \alpha^{m/2} \|\log Z\|_{3,\infty} C_\chi \|\psi - \psi^*\|) \end{aligned}$$

□

IV.D.2 Proof of the SGD recursion (Lemma IV.3.1)

We are now ready to provide an SGD-style recursion for the expected squared distance to the optimum $\delta_t := \mathbb{E}[\|\psi_t - \psi^*\|^2]$.

Lemma (Restatement of Lemma IV.3.1). *Let ψ_t be the iterates produced by Algorithm 1. Let $\delta_t = \mathbb{E}[\|\psi_t - \psi^*\|^2]$, $\sigma_* = (\mathbb{E}[\|\phi(X_1) - \mathbb{E}[\phi(X_1)]\|^2])^{1/2}$, $\sigma_t = (\mathbb{E}[\|\phi(X^{\psi_t}) - \mathbb{E}[\phi(X^{\psi_t})]\|^2 | \mathcal{F}_t])^{1/2}$. Then, under Assumptions A0, A1 and A2, for all $t \geq 1$, we have:*

$$\delta_t \leq (1 - 2\eta_t \tilde{\mu}_{m,t-1} + 2\eta_t^2 L^2) \delta_{t-1} + \eta_t^2 \tilde{\sigma}_{m,t-1}^2 + 4\alpha^{m/2} \eta_t^2 \|\log Z\|_{2,\infty} C_\chi \delta_{t-1}^{1/2}$$

where $\|\log Z\|_{3,\infty}$ is a constant, $\tilde{\mu}_{m,t} := \mu - \alpha^m \sigma_t C_\chi$, and $\tilde{\sigma}_{m,t} := (\sigma_*^2 + \sigma_t^2 + 2\sigma_t^2 \alpha^{2m})^{1/2}$.

Proof. In this proof, we note $(\mathcal{F}_t)_{t \geq 0}$, the increasing family of σ -algebras generated by the random variables $(X_t)_{t \geq 0} \sim p_{\psi^*}$ and the Markov chain samples $\tilde{X}_t^m | X_t, \psi_t \sim$

$k_{\psi_t}^m(X_t, \cdot)$. We decompose the integrand of δ_t as follows:

$$\begin{aligned}\|\psi_t - \psi_n^*\|^2 &= \|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t(\psi_{t-1}, X_t)) - \psi_n^*\|^2 \\ &\leq \|\psi_{t-1} - \eta_t h_t(\psi_{t-1}, X_t) - \psi^*\|^2 \quad (\text{By Lemma IV.C.4}) \\ &= \|\psi_{t-1} - \psi^*\|^2 - 2\eta_t \langle h_t(\psi_{t-1}, X_t), \psi_{t-1} - \psi^* \rangle + \eta_t^2 \|h_t(\psi_{t-1}, X_t)\|^2\end{aligned}$$

The first term is (up to an averaging operation) the previous iterate. The middle term will ensure (provided m is large enough) contraction of the expected distance to the optimum. Finally, the third term can be described by Lemma IV.D.3, and essentially behaves like the second moment of a log-likelihood stochastic gradient. Indeed, noting

$$\bar{g}(\psi_{t-1}) := -\mathbb{E}[\phi(X_1)] + \mathbb{E}[\phi(X^{\psi_{t-1}})],$$

the expectation of g_t w.r.t x_t (which is the gradient of the negative cross-entropy between p_ψ and p_{ψ^*}), we have:

$$\begin{aligned}\langle h_t(\psi_{t-1}, X_t), \psi_{t-1} - \psi^* \rangle &= \langle \bar{g}(\psi_{t-1}), \psi_{t-1} - \psi^* \rangle \\ &\quad + \underbrace{\langle (h_t(\psi_{t-1}, X_t) - \bar{g}(\psi_{t-1})), \psi_{t-1} - \psi^* \rangle}_{\Delta}.\end{aligned}$$

Applying Lemma IV.C.7, we get that

$$\begin{aligned}h_t(\psi_{t-1}, X_t) - \bar{g}(\psi_{t-1}) &= \phi(k^m(X_t, \cdot)) - \mathbb{E}[\phi(X^{\psi_{t-1}}) | \mathcal{F}_{t-1}] \\ \implies \mathbb{E}[h_t(\psi_{t-1}, X_t) - \bar{g}(\psi_{t-1}) | \mathcal{F}_{t-1}] &= P_{\psi_{t-1}}^m \phi - \mathbb{E}[\phi(X_1^{\psi_{t-1}}) | \mathcal{F}_{t-1}],\end{aligned}$$

meaning

$$\begin{aligned}|\mathbb{E}[\Delta | \mathcal{F}_{t-1}]| &\leq \left\| \mathbb{E} \left[P_{\psi_{t-1}}^m (\phi - \mathbb{E}[\phi(X^{\psi_{t-1}})])(X_t) \middle| \mathcal{F}_{t-1} \right] \right\| \|\psi_{t-1} - \psi^*\| \\ &\leq \alpha^m \sigma_{t-1} C_\chi \|\psi_{t-1} - \psi^*\|^2\end{aligned}$$

On the other hand, by applying Lemma IV.C.8 to \bar{g} , we have:

$$\langle \bar{g}(\psi_{t-1}), \psi_{t-1} - \psi^* \rangle = \langle \bar{g}(\psi_{t-1}) - \bar{g}(\psi^*), \psi_{t-1} - \psi^* \rangle \geq \mu \|\psi_{t-1} - \psi^*\|^2$$

Combining the above results, we obtain:

$$\begin{aligned} & \mathbb{E} \left[\|\psi_t - \psi^*\|^2 | \mathcal{F}_{t-1} \right] \\ & \leq (1 - 2\eta_t(\mu - \alpha^m \sigma_{t-1} C_\chi)) \|\psi_{t-1} - \psi^*\|^2 \\ & \quad + \eta_t^2 (2\sigma_*^2 + 2\sigma_{t-1}^2 + 2L^2 \|\psi_{t-1} - \psi^*\|^2 \\ & \quad + 4(\sigma_{t-1}^2 \alpha^{2m} + \alpha^{m/2} \|\log Z\|_{3,\infty} C_\chi \|\psi_{t-1} - \psi^*\|)) \end{aligned}$$

And the result follows by integrating over \mathcal{F}_{t-1} . \square

IV.D.3 Proof of Online CD convergence

We now prove Theorem V.D.1. The recursion of Lemma IV.3.1 is almost identifiable, up to a cross-term of second order, with the one of an SGD algorithm as presented in the setting of [169, Theorem 1]. To make the identification exact, we use the bound $4\alpha^{m/2}\eta_t^2 \|\log Z\|_{3,\infty} C_\chi \delta_{t-1}^{1/2} \leq 2\alpha^{m/2}\eta_t^2 \delta_t + 2\alpha^{m/2}\eta_t^2 \|\log Z\|_{3,\infty}^2 C_\chi^2$, yielding the following recursion:

$$\begin{aligned} \delta_t \leq & (1 - 2\eta_t(\mu - \alpha^m \sigma C_\chi) + 2\eta_t^2(L^2 + \alpha^{m/2})) \delta_{t-1} + (\sigma^2(2 + 2\alpha^{2m}) \\ & + \alpha^{m/2} \|\log Z\|_{3,\infty}^2 C_\chi^2) \eta_t^2 \end{aligned} \quad (\text{IV.16})$$

where we used the fact that $\tilde{\sigma}_{m,t} \leq \sigma$. This recursion is of the same form as the one studied in [169, Equation 6, Theorem 1] given by:

$$\delta_t \leq (1 - 2\mu\gamma_t + 2L^2\gamma_t^2) \delta_{t-1} + 2\sigma^2\gamma_t^2$$

by identifying:

$$\begin{aligned} \sigma^2 & \leftarrow \sigma^2(2 + 2\alpha^{2m}) + \alpha^{m/2} \|\log Z\|_{3,\infty}^2 =: \tilde{\sigma}_m^2 \\ L^2 & \leftarrow (L^2 + \alpha^{m/2}) =: \tilde{L}^2 \\ \mu & \leftarrow \mu - \alpha^m \sigma C_\chi =: \tilde{\mu}_m \\ \gamma_t & \leftarrow \eta_t \end{aligned}$$

We can use the same unrolling strategy as theirs (the only condition required to

proceed is that $\tilde{\mu}_m < \tilde{L}$, which automatically holds since $\mu < L$), and we obtain

$$\delta_n \leq \begin{cases} 2 \exp(4\tilde{L}C^2\varphi_{1-2\beta}(n)) \exp\left(-\frac{\tilde{\mu}_m C}{4}n^{1-\beta}\right) \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m n^\beta}, & \text{if } 0 \leq \beta < 1 \\ \frac{\exp(2\tilde{L}^2C^2)}{n^{\tilde{\mu}_m C}} \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + 2\tilde{\sigma}_m^2 C^2 \frac{\varphi_{\tilde{\mu}_m C/2-1}(n)}{n^{\tilde{\mu}_m C/2}}, & \text{if } \beta = 1. \end{cases}$$

□

IV.D.4 Proof of online CD with averaging (Theorem IV.3.3)

We first restate the theorem in its complete form.

Theorem (Contrastive Divergence with Polyak-Ruppert averaging). *Let $(\psi_t)_{t \geq 0}$ the sequence of iterates obtained by running the CD algorithm with a learning rate $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$. Define $\bar{\psi}_n := \frac{1}{n} \sum_{i=1}^n \psi_i$. Then, under the same assumptions as Theorem V.D.1 we have, for all $n \geq 1$,*

$$\sqrt{\mathbb{E} [\|\bar{\psi}_n - \psi^*\|^2]} \leq n^2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi)^{-1})}{n}} + \mathcal{O}\left(n^{\max\left(-\left(\frac{1}{2} + \frac{\beta}{4}\right), -\beta, \frac{\beta}{2} - 1, -\left(\frac{\beta}{2} + m\frac{|\log \alpha|}{\log n}\right)\right)}\right)$$

Where $\mathcal{I}(\psi^*) := \text{Cov}[\phi(X_1)]$ is the Fisher information matrix of the data distribution. Additionally, if $m > \frac{(1-\beta)\log n}{2|\log \alpha|}$, we have $\sqrt{\mathbb{E} [\|\bar{\psi}_n - \psi^*\|^2]} \leq 2\sqrt{\frac{\text{tr}(\mathcal{I}(\psi)^{-1})}{n}} + o\left(n^{-1/2}\right)$.

Recall that we denote by h_n the standard online CD gradient defined in Equation IV.3:

$$h_n(\psi_{n-1}, X_n) = -\phi(X_n) + \phi(k_{\psi_{n-1}}^m(\cdot, X_n)), \quad X_n \sim p_{\psi^*}, \quad \forall n \in \mathbb{N} \setminus \{0\},$$

as well as

$$\bar{h}(\psi_{n-1}) := \mathbb{E}[h(\psi_{n-1}) | \mathcal{F}_{n-1}] = \mathbb{E}[\phi(X_n)] + \mathbb{E}\left[\phi(K_{\psi_{n-1}}^m(X_n, \cdot)) \mid \mathcal{F}_{n-1}\right].$$

In the following, we will hide the dependence on X_n of $h_n(\psi, X_n)$ by writing $h_n(\psi)$. We start by establishing some intermediate lemmas.

Lemma IV.D.4. *Under Assumptions A0, A1, A2, the online CD iterates produced by Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify*

$$\frac{1}{n} \sqrt{\sum_{i=1}^n \left(\mathbb{E} [\|\psi_i - \psi^*\|^2] \right)} = \mathcal{O}(n^{-\frac{1}{2}-\frac{\beta}{2}}).$$

Proof. Let us note $\delta_n = \mathbb{E} [\|\psi_n - \psi^*\|^2]$. Summing the r.h.s of Theorem V.D.1, we have

$$\sum_{i=1}^n \delta_i \leq \sum_{i=1}^n \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m i^\beta} + 2 \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right) \underbrace{\sum_{i=1}^n e^{4\tilde{L}C^2\varphi_{1-2\beta}(i)} e^{-\frac{\tilde{\mu}_m C}{4} n^{1-\beta}}}_{A_3}$$

$$\implies \frac{1}{n} \sqrt{\sum_{i=1}^n \delta_i} \leq \frac{1}{n} \sqrt{\frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m} \varphi_{1-\beta}(n)} + \frac{1}{n} \sqrt{\left(2(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}) A_3 \right)} = \mathcal{O}\left(n^{-\frac{1}{2}-\frac{\beta}{2}}\right).$$

where A_3 is finite if $\beta < 1$, and $A_3 = O(n)$ otherwise [169]. \square

Lemma IV.D.5. *Under Assumptions A0, A1, A2, the online CD iterates produced by Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify*

$$\frac{1}{n} \sqrt{\sum_{i=1}^n \left(\mathbb{E} [\|\psi_i - \psi^*\|^2] \right)^{1/2}} = \mathcal{O}(n^{-\frac{1}{2}-\frac{\beta}{4}}).$$

Proof. Let us note $\delta_n = \mathbb{E} [\|\psi_n - \psi^*\|^2]$. Applying $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ to the r.h.s of Theorem V.D.1, we have

$$\begin{aligned} \sum_{i=1}^n \delta_i^{1/2} &\leq \frac{2C^{1/2}\tilde{\sigma}_m}{2\tilde{\mu}_m^{1/2}} \sum_{i=1}^n i^{-\beta/2} + \sqrt{2 \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right) \underbrace{\sum_{i=1}^n e^{2\tilde{L}^2C^2\varphi_{1-2\beta}(i)} e^{-\frac{\tilde{\mu}_m C}{8} i^{1-\beta}}}_{A_4}} \\ \frac{1}{n} \sqrt{\sum_{i=1}^n \delta_i^{1/2}} &\leq \frac{1}{n} \sqrt{\frac{2C^{1/2}\tilde{\sigma}_m}{\tilde{\mu}_m^{1/2}} \varphi_{1-\beta/2}(n)} + \frac{1}{n} \left(\sqrt{2 \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right) A_4} \right)^{1/2} \\ &= \mathcal{O}\left(n^{-\frac{1}{2}-\frac{\beta}{4}}\right). \end{aligned}$$

where A_4 is finite if $\beta < 1$, and $A_4 = O(n)$ otherwise [169]. \square

Lemma IV.D.6. *Under Assumptions A0, A1, A2, the online CD iterates produced by*

Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify

$$\sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n h_i(\psi_{i-1}) \right\|^2 \right]} = \mathcal{O} \left(n^{\max(\frac{\beta}{2}-1, -\frac{2\beta}{2\beta+1})} \right).$$

Proof. The main subtlety consists in taking into account the projection step. Indeed, write $h_t(\psi_t)$ as:

$$\begin{aligned} \psi_t &= \text{Proj}_{\Psi}(\psi_{t-1} - \eta_t h_t(\psi_{t-1})) \\ &= \psi_{t-1} - \eta_t h_t(\psi_{t-1}) + \underbrace{(\text{Proj}_{\Psi}(\psi_{t-1} - \eta_t h_t(\psi_{t-1})) - (\psi_{t-1} - \eta_t h_t(\psi_{t-1})))}_{\Delta_t} \end{aligned}$$

Implying $h_t = \frac{1}{\eta_t} (\psi_{t-1} - \psi_t) + \frac{\Delta_t}{\eta_t}$, and thus

$$\begin{aligned} &\sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n h_i(\psi_{i-1}) \right\|^2 \right]} \\ &\leq \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_i} (\psi_{i-1} - \psi_i) \right\|^2 \right]} + \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\eta_i} \right\|^2 \right]} \end{aligned}$$

where the last line used Minkowski's inequality. We now control the two terms separately. The first term is well-known in the SGD literature, and can be controlled in a straightforward manner. The second term characterizes the deviation caused by the projection step, and we control it below.

Controlling the second term We use a similar strategy as in [81], but without requiring the assumption on the boundedness of the gradients. Let us use the notation $h_t := h_t(\psi_{t-1})$ for brevity. First, we have, using again Minkowski's inequality, that:

$$\sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\eta_i} \right\|^2 \right]} \leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_i} \sqrt{\mathbb{E} \left[\|\Delta_i\|^2 \right]}.$$

Note that Δ_t is 0 if $\psi_{t-1} - \eta_t h_t \in \Psi$, e.g.

$$\Delta_t = (\text{Proj}_{\Psi}(\psi_{t-1} - \eta_t h_t) - (\psi_{t-1} - \eta_t h_t)) \times \mathbb{I}_{\{\psi_{t-1} - \eta_t h_t \notin \Psi\}}$$

Which implies, using Hölder's inequality, that

$$\begin{aligned} \mathbb{E} \|\Delta_t\|^2 &= \mathbb{E} \left[\|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - (\psi_{t-1} - \eta_t h_t)\|^2 \times \mathbb{I}_{\{\psi_{t-1} - \eta_t h_t \notin \Psi\}} \right] \\ &\leq \mathbb{E} \left[\|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - (\psi_{t-1} - \eta_t h_t)\|^{2p} \right]^{\frac{1}{p}} \times \\ &\quad \mathbb{P}[\{\psi_{t-1} - \eta_t h_t \notin \Psi\}]^{\frac{1}{q}} \end{aligned}$$

For any pair (p, q) of Hölder conjugates, whose exact value will be determined later. We investigate the two terms separately; for the first term, note that

$$\begin{aligned} &\|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - (\psi_{t-1} - \eta_t h_t)\|^2 \\ &\leq 2 \|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - \psi_{t-1}\|^2 + 2 \|\eta_t h_t\|^2 \\ &\leq 2 \|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - \text{Proj}_\Psi(\psi_{t-1})\|^2 + 2 \eta_t^2 \|h_t\|^2 \\ &\stackrel{(a)}{\leq} 2 \|\psi_{t-1} - \eta_t h_t - \psi_{t-1}\|^2 + 2 \eta_t^2 \|h_t\|^2 \\ &\leq 4 \eta_t^2 \|h_t\|^2 \end{aligned}$$

where in (a), we used the fact that projections are 1-Lipschitz. Thus, we have:

$$\begin{aligned} &\left(\mathbb{E} \left[\|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - (\psi_{t-1} - \eta_t h_t)\|^{2p} \right] \right)^{1/p} \\ &\leq \left(4^p \eta_t^{2p} \times \mathbb{E} \|h_t\|^{2p} \right)^{1/p} \\ &\leq 16 \eta_t^2 \times \left(\tau_{2p}^{2p} + 2C_\chi r_\Psi \tau_{4p}^{2p} \right)^{1/p} \end{aligned}$$

where we noted, as in Lemma V.D.3, $\tau_p := \left(\sup_{\psi \in \Psi} \mathbb{E} \|\phi\|^p \right)^{1/p} < +\infty$, $r_\Psi = \sup_{\psi, \psi' \in \Psi} \|\psi - \psi'\|$, and we relied on the fact that

$$\begin{aligned} \mathbb{E} \left[\|h_t\|^{2p} \mid \mathcal{F}_{t-1} \right] &\leq 2^{2p-1} \left(\tau_{2p}^{2p} + 2C_\chi r_\Psi \tau_{4p}^{2p} \|\psi_{t-1} - \psi^\star\| \right) \\ &\leq 2^{2p-1} \left(\tau_{2p}^{2p} + 2C_\chi r_\Psi \tau_{4p}^{2p} \right) \end{aligned}$$

For the probability term, we have:

$$\begin{aligned} \mathbb{E} [\|\psi_t - \psi^*\|^4] &\geq \mathbb{E} \|(\psi_t - \psi^*) \times \mathbb{I}_{\{\psi_{t-1} - \eta_t h_t \notin \Psi\}}\|^4 \\ &\geq d_{\psi^*}^4 \times \mathbb{P}[\{\psi_{t-1} - \eta_t h_t \notin \Psi\}] \\ \implies \mathbb{P}[\{\psi_{t-1} - \eta_t h_t(\psi_{t-1}) \notin \Psi\}] &\leq \frac{1}{d_{\psi^*}^4} \mathbb{E} [\|\psi_t - \psi^*\|^4] \end{aligned}$$

where we noted $d_{\psi^*} = \inf_{\psi \in \partial\Psi} \|\psi - \psi^*\|$, which is positive by [A0](#) as $\psi^* \in \text{int}(\Psi)$.

Combining the two bounds, we thus obtain:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_i} \sqrt{\mathbb{E} [\|\Delta_i\|^2]} &\leq \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_i} \times \left(4\eta_i \left(\tau_{2p}^{2p} + 2C_\chi r_\Psi \tau_{4p}^{2p} \right)^{1/2p} \right) \times \frac{1}{d_{\psi^*}^{2/q}} \left(\mathbb{E} [\|\psi_i - \psi^*\|^4] \right)^{\frac{1}{2q}} \\ &= \frac{4 \left(\tau_{2p}^{2p} + 2C_\chi r_\Psi \tau_{4p}^{2p} \right)^{1/2p}}{d_{\psi^*}^{2/q}} \times \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E} [\|\psi_i - \psi^*\|^4] \right)^{1/2q} \end{aligned}$$

Recalling that

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n \left(\mathbb{E} [\|\psi_k - \psi^*\|^4] \right)^{1/2q} &\leqslant \frac{1}{n} \sum_{k=1}^n \left[\left(\frac{12C^{3/2}\tilde{\tau}_1^2}{\tilde{\mu}_m^{1/2}} \right)^{1/q} \frac{1}{k^{3\beta/2q}} + \left(\frac{26\tilde{\tau}_1^2 C}{\tilde{\mu}_m} \right)^{1/q} \frac{1}{k^{\beta}} \right] \\ &\quad + \frac{1}{n} \sum_{k=1}^n \exp \left(-\frac{1}{q} \times \left(\frac{\tilde{\mu}_m C}{16} k^{1-\beta} + 16\tilde{L}_1^2 C^2 \varphi_{1-2\beta}(k) + 24\tilde{L}_1^4 C^4 \right) \right) \\ &\quad \times \left(32\tilde{\tau}_1^4 C^4 + \frac{160\tilde{\tau}_1^4 C^3}{\tilde{\mu}_m} + \mathbb{E} \|\psi_0 - \psi^*\|^4 + \frac{20C\tilde{\tau}_1^2}{\tilde{\mu}_m} \delta_0 \right)^{1/2q} \\ &\leqslant \frac{1}{n} \left(\left(\frac{12C^{3/2}\tilde{\tau}_1^2}{\tilde{\mu}_m^{1/2}} \right)^{1/q} \times \varphi_{1-\frac{3\beta}{2q}}(n) + \left(\frac{26\tilde{\tau}_1^2 C}{\tilde{\mu}_m} \right)^{1/q} \times \varphi_{1-\frac{\beta}{q}}(n) \right) \\ &\quad + \frac{A_5}{n} \times \left(32\tilde{\tau}_1^4 C^4 + \frac{160\tilde{\tau}_1^4 C^3}{\tilde{\mu}_m} + \mathbb{E} \|\psi_0 - \psi^*\|^4 + \frac{20C\tilde{\tau}_1^2}{\tilde{\mu}_m} \delta_0 \right)^{1/2q} \end{aligned}$$

where $\tilde{\tau}_1$ and \tilde{L}_1 are defined in Lemma V.D.3, and

$$A_5 = \sum_{k=1}^n \exp \left(-\frac{1}{q} \times \left(\frac{\tilde{\mu}_m C}{16} k^{1-\beta} + 16 \tilde{L}_1^2 C^2 \varphi_{1-2\beta}(k) + 24 \tilde{L}_1^4 C^4 \right) \right),$$

which is finite if $\beta < 1$, and $O(n)$ otherwise, we have that

$$\frac{1}{n} \sum_i^n \left(\mathbb{E} \left[\left[\|\psi_n - \psi^*\|^4 \right] \right] \right)^{1/2q} = \mathcal{O}(n^{-\frac{\beta}{q}})$$

In particular, taking $(p, q) = (\frac{2\beta+1}{2\beta-1}, \frac{1}{2} + \beta)$, we have that

$$\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i}{\eta_i} \right\|^2 \right] = \mathcal{O}(n^{-\frac{2\beta}{2\beta+1}}) = o(n^{-1/2}).$$

for all $\beta \in (\frac{1}{2}, 1)$.

Controlling the first term The first term was handled in the case of standard SGD [169, Theorem 3], and the only condition needed to reuse their steps is that $(\psi_t)_{t \leq n}$ satisfies an upper bound of the same form as the one [169, Theorem 1] derived. This is precisely the nature of our bound of ψ_t established in Theorem V.D.1, with $\tilde{\mu}_m, \tilde{\sigma}_m, \tilde{L}$. Borrowing on their result, we have:

$$\begin{aligned} & \sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{\eta_i} (\psi_{i-1} - \psi_i) \right\|^2 \right]} \\ & \leq \frac{4\tilde{\sigma}_m \beta}{C^{1/2} n \tilde{\mu}_m} \varphi_{\beta/2}(n) + \frac{4\beta}{C n \tilde{\mu}_m^{1/2}} \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right)^{1/2} A_2 \\ & \quad + \frac{1}{n \tilde{\mu}_m^{1/2}} \left(\frac{1}{C} + 2\tilde{L} \right) \delta_0^{1/2} + \frac{2\tilde{L}}{n \tilde{\mu}_m^{1/2}} \frac{2C^{1/2} \tilde{\sigma}_m}{\tilde{\mu}_m^{1/2}} \varphi_{1-\beta}(n)^{1/2} \\ & \quad + \frac{4\tilde{L}}{n \tilde{\mu}_m^{1/2}} \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right)^{1/2} A_2^{1/2} \end{aligned}$$

where $\tilde{\mu}_m, \tilde{\sigma}_m$ and \tilde{L} are defined in Theorem V.D.1, and

$$A_2 = \sum_{k=1}^n e^{\frac{-\tilde{\mu}_m C}{16} k^{1-\beta} + 16 \tilde{L}_1^4 C^4 \varphi_{1-2\beta}(k)}$$

□

Lemma IV.D.7. *Under Assumptions A0, A1, A2, the online CD iterates produced by*

Algorithm 1 using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify

$$\frac{1}{n} \sum_{i=1}^n \sqrt{\mathbb{E} [\|\psi_i - \psi^*\|^4]} = \mathcal{O}(n^{-\beta}).$$

Proof. We proceed as in the proof of [169, Theorem 3], first establishing a recurrence for $\mathbb{E} [\|\psi_i - \psi^*\|^4]$, and then unrolling it. We have

$$\begin{aligned} & \mathbb{E} [\|\psi_n - \psi^*\|^4 \mid \mathcal{F}_{n-1}] \\ & \leq \|\psi_{n-1} - \psi^*\|^4 + 6\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 \mathbb{E} [\|h_n(\psi_{n-1}, X_t)\|^2 \mid \mathcal{F}_{n-1}] \\ & \quad + \eta_n^4 \mathbb{E} [\|h_n(\psi_{n-1}, X_t)\|^4 \mid \mathcal{F}_{n-1}] \\ & \quad - 4\eta_n \|\psi_{n-1} - \psi^*\|^2 \langle \psi_{n-1} - \psi^*, \mathbb{E} [h_n(\psi_{n-1}, X_t) \mid \mathcal{F}_{n-1}] \rangle \\ & \quad + 4\eta_n^3 \|\psi_{n-1} - \psi^*\| \mathbb{E} [\|h_n(\psi_{n-1}, X_t)\|^3 \mid \mathcal{F}_{n-1}]. \end{aligned}$$

The second and fourth terms will be controlled using results from our previous sections. For simplicity, we don't attempt to relate the moments of $\|h_n\|^4$ as precisely as before. Instead we use the fact that, for all $\psi \in \Psi$,

$$\mathbb{E} [\|h_n(\psi, X_t)\|^k] \leq 2^{k-1} \left(\mathbb{E} [\|\phi(K_\psi^m(X_n))\|^k] + \mathbb{E} [\|\phi(X_n)\|^k] \right).$$

Let us note $\tau = \left(\sup_{\psi \in \Psi} \mathbb{E} [\|\phi(X^\psi)\|^8] \right)^{1/8}$. We have, for $k \leq 4$, and $\psi \in \Psi$

$$\begin{aligned} \mathbb{E} [\|\phi(K_\psi^m(X_n))\|^k] & \leq \mathbb{E} [\|\phi(X^\psi)\|^k] \\ & \quad + C_\chi \left(\mathbb{E} \left[\left(\|\phi(X^\psi)\|^k - \mathbb{E} [\|\phi(X^\psi)\|^k] \right)^2 \right] \right)^{1/2} \|\psi - \psi^*\| \\ & \leq \tau^k + 2C_\chi \tau^k \|\psi - \psi^*\|. \end{aligned}$$

Here, we used the fact that $P_{\psi_{n-1}}^m$ is a contraction, and $(\mathbb{E} [\|\phi(X^\psi)\|^k])^{1/k}$ is an increasing function of k for all ψ . On the other hand, we simply have $\mathbb{E} [\|\phi(X_n)\|^k] \leq$

τ^k . Plugging this into the previous equation, applying it to $\psi = \psi_{n-1}$, we obtain

$$\begin{aligned} \mathbb{E} \left[\|\psi_n - \psi^*\|^4 \mid \mathcal{F}_{n-1} \right] &\leq \|\psi_{n-1} - \psi^*\|^4 \\ &\quad + 6\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 (4\tau^2 + 4C_\chi \tau^2 \|\psi_{n-1} - \psi^*\|) \\ &\quad + \eta_n^4 (16\tau^4 + 16C_\chi \tau^4 \|\psi_{n-1} - \psi^*\|) \\ &\quad - 4\eta_n \|\psi_{n-1} - \psi^*\|^2 \langle \psi_{n-1} - \psi^*, \mathbb{E}[h_n \mid \mathcal{F}_{n-1}] \rangle \\ &\quad + 4\eta_n^3 \|\psi_{n-1} - \psi^*\| (8\tau^3 + 8C_\chi \tau^3 \|\psi_{n-1} - \psi^*\|) \end{aligned}$$

To simplify the recursion, we use the four following inequalities:

$$\begin{aligned} \tau^2 \eta_n^2 \|\psi_{n-1} - \psi^*\|^3 &\leq \frac{1}{2} (\tau^2 \eta_n^2 (\|\psi_{n-1} - \psi^*\|^2 + \tau^2 \eta_n^2 \|\psi_{n-1} - \psi^*\|^4)) \\ \tau^4 \eta_n^4 \|\psi_{n-1} - \psi^*\| &\leq (\tau^4 \eta_n^4 + \frac{1}{4} \tau^4 \eta_n^4 \|\psi_{n-1} - \psi^*\|^4) \\ \eta_n^3 \tau^3 \|\psi_{n-1} - \psi^*\| &\leq \frac{1}{2} (\eta_n^2 \tau^2 \|\psi_{n-1} - \psi^*\|^2 + 16\eta_n^4 \tau^4) \\ \tau^3 \eta_n^3 \|\psi_{n-1} - \psi^*\|^2 &\leq \frac{1}{2} (\eta_n^4 \tau^4 + \|\psi_{n-1} - \psi^*\|^2 \eta_n^2 \tau^2) \end{aligned}$$

Injecting them in our recursion, we obtain:

$$\begin{aligned} \mathbb{E} \left[\|\psi_n - \psi^*\|^4 \mid \mathcal{F}_{n-1} \right] &\leq \|\psi_{n-1} - \psi^*\|^4 \\ &\quad + 12\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 \tau^2 \\ &\quad + 12C_\chi \tau^2 \eta_n^2 \|\psi_{n-1} - \psi^*\|^2 \\ &\quad + 12C_\chi \tau^2 \eta_n^2 \|\psi_{n-1} - \psi^*\|^4 \\ &\quad + 16\eta_n^4 \tau^4 \\ &\quad + 16C_\chi \eta_n^4 \tau^4 \\ &\quad + 4C_\chi \eta_n^4 \tau^4 \|\psi_{n-1} - \psi^*\|^4 \\ &\quad - 4\eta_n \tilde{\mu}_m \|\psi_{n-1} - \psi^*\|^4 \\ &\quad + 16\eta_n^2 \tau^2 \|\psi_{n-1} - \psi^*\|^2 \\ &\quad + 16\eta_n^4 \tau^4 \\ &\quad + 16\eta_n^4 C_\chi \tau^4 \\ &\quad + 16\eta_n^2 C_\chi \tau^2 \|\psi_{n-1} - \psi^*\|^2 , \end{aligned}$$

which, after further simplifications, yields

$$\begin{aligned}
& \mathbb{E}[\|\psi_n - \psi^*\|^4 \mid \mathcal{F}_{n-1}] \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12C_\chi \eta_n^2 \tau^2 + 4C_\chi \tau^4 \eta_n^4) \\
& \quad + \eta_n^2 \|\psi_{n-1} - \psi^*\|^2 (28(1+C_\chi)\tau^2) + 32\eta_n^4(\tau^4(1+C_\chi)) \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12(1+C_\chi)\eta_n^2 \tau^2 + 4(1+C_\chi)\tau^4 \eta_n^4) \\
& \quad + 28\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 ((1+C_\chi)\tau)^2 + 32\eta_n^4(1+C_\chi)\tau^4 \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12\eta_n^2((1+C_\chi)\tau)^2 + 4((1+C_\chi)\tau)^4 \eta_n^4) \\
& \quad + 28\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 ((1+C_\chi)\tau)^2 + 32\eta_n^4(1+C_\chi)\tau^4 \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12\eta_n^2(2(1+C_\chi)\tau)^2 + 16\eta_n^2(2(1+C_\chi)\tau)^3 \\
& \quad + 4(2(1+C_\chi)\tau)^4 \eta_n^4) + 20\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 (2(1+C_\chi)\tau)^2 + 16\eta_n^4(2(1+C_\chi)\tau)^4 \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12\eta_n^2(2(1+C_\chi)\tau + L)^2 + 16\eta_n^2(2(1+C_\chi)\tau + L)^3 \\
& \quad + 4(2(1+C_\chi)\tau + L)^4 \eta_n^4) + 20\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 (2(1+C_\chi)\tau)^2 \\
& \quad + 16\eta_n^4(2(1+C_\chi)\tau)^4 \\
& \leq \|\psi_{n-1} - \psi^*\|^4 (1 - 4\eta_n \tilde{\mu}_m + 12\eta_n^2 \tilde{L}_1^2 + 16\eta_n^2 \tilde{L}_1^3 + 4\tilde{L}_1^4 \eta_n^4) \\
& \quad + 20\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 \tilde{\tau}_1^2 + 16\eta_n^4 \tilde{\tau}_1^4
\end{aligned}$$

where we defined $\tilde{\tau}_1 := 2(1+C_\chi)\tau$ and $\tilde{L}_1 := 2(1+C_\chi)\tau + L$. This recursion is of the form of the one studied in [169, Equation 32] (note that by design, $\tilde{L}_1 \geq \tilde{\mu}_m$.) The steps performed to bound $\mathbb{E}[\|\psi_n - \psi^*\|^4]$ thus follow from their derivations, and we obtain:

$$\begin{aligned}
& \frac{1}{n} \sqrt{\sum_{i=1}^n \mathbb{E}[\|\psi_{i-1} - \psi^*\|^4]} \\
& \leq \frac{C\tilde{\tau}_1^2}{2n} \left(C^{1/2} \varphi_{1-3\beta/2}(n) + \tilde{\mu}_m^{-1/2} \varphi_{1-\beta}(n) \right) \\
& \quad + \frac{\sqrt{20}C^{1/2}\tilde{\tau}_1}{2n} A_1 \exp(24\tilde{L}_1^4 C^4) \left(\delta_0 + \frac{\tilde{\mu}_m \mathbb{E}[\|\psi_0 - \psi^*\|^4]}{20C\tilde{\tau}_1^2} + 2\tilde{\tau}_1^2 C^3 \tilde{\mu}_m + 8\tilde{\tau}_1^2 C^2 \right)^{1/2} \\
& = \mathcal{O}(n^{-\beta}),
\end{aligned}$$

where

$$A_1 = \sum_{k=1}^n e^{\frac{-\tilde{\mu}_m C}{16} k^{1-\beta} + 16\tilde{L}_1^4 C^4 \varphi_{1-2\beta}(k)}$$

and we have $A_1 < +\infty$ if $\beta < 1$, and $A_1 = O(n)$ otherwise. \square

Lemma IV.D.8. *For all $\psi \in \Psi$, we have:*

$$\begin{aligned} & \left\| \text{Cov} \left[\phi(K_\psi^m(X_n)) \right] - \text{Cov} [\phi(X^\psi)] \right\|_F \\ & \leq \alpha^m C_\chi (\bar{\tau}^{1/2} + 2 \|\log Z\|_{4,\infty} \sigma) \|\psi - \psi^*\| + \alpha^{2m} C_\chi^2 \sigma^2 \|\psi - \psi^*\|^2. \end{aligned} \quad (\text{IV.17})$$

where $\bar{\tau} := \sup_{\psi \in \Psi} \mathbb{E} \left[\|\phi(X^\psi) \phi^\top(X^\psi) - \mathbb{E}[\phi(X^\psi) \phi^\top(X^\psi)]\|_F^2 \right] < +\infty$ as well as $\|\log Z\|_{4,\infty} := \sup_{\psi \in \Psi} \|\mathbb{E}_\psi \phi\| \leq \sup_{\psi \in \Psi} \sum_{i=1}^d \partial_i^2 \log Z(\psi)^2$.

Proof. We have

$$\text{Cov} \left[\phi(K_\psi^m(X_n)) \right] = \mathbb{E} \left[P_\psi^m \left(\phi(X_n) \phi^\top(X_n) \right) \right] - \left(\mathbb{E}[P_\psi^m \phi(X_n)] \right) \left(\mathbb{E}[P_\psi^m \phi(X_n)]^\top \right)$$

Looking at the second moment first, we have

$$\begin{aligned} & \mathbb{E} \left[P_\psi^m \phi(X_n) \phi(X_n)^\top \right] - \mathbb{E} \left[\phi(X^\psi) \phi(X^\psi)^\top \right] \\ & = \underbrace{\mathbb{E} \left[P_\psi^m (\phi \phi^\top - \mathbb{E}[\phi(X^\psi) \phi(X^\psi)^\top])(X_n) \right]}_{\Delta_1}. \end{aligned} \quad (\text{IV.18})$$

Applying Lemma IV.C.7 to the \mathbb{R}^{d^2} -valued function f given by $f_{ij} := \phi_i \phi_j - \mathbb{E}[\phi_i(X^\psi) \phi_j(X^\psi)]$,

$$\begin{aligned} \|\Delta_1\|_F & \leq \|\psi - \psi^*\| \alpha^m C_\chi \sqrt{\mathbb{E}[\|(\phi \phi^\top - \mathbb{E}[(\phi \phi^\top)(X^\psi)])(X^\psi)\|_F^2]} \\ & \leq \bar{\tau}^{1/2} \alpha^m C_\chi \|\psi - \psi^*\|. \end{aligned}$$

We now investigate the first moment. We have

$$\begin{aligned} & \mathbb{E} \left[P_\psi^m \phi(X_n) \right] = \underbrace{\mathbb{E} \left[P_\psi^m \phi(X_n) \right] - \mathbb{E}[\phi(X^\psi)]}_{\Delta_{2,1}} + \mathbb{E}[\phi(X^\psi)] \\ & \implies \underbrace{(\mathbb{E} \left[P_\psi^m \phi(X_n) \right]) (\mathbb{E} \left[P_\psi^m \phi(X_n) \right])^\top}_{\Delta_2} - \mathbb{E}[\phi(X^\psi)] \mathbb{E}[\phi(X^\psi)^\top] = \Delta_{2,1} \Delta_{2,1}^T \\ & \quad + \Delta_{2,1} \mathbb{E}[\phi(X^\psi)^\top] + \mathbb{E}[\phi(X^\psi)] \Delta_{2,1}^\top. \end{aligned}$$

Thus, applying Lemma IV.C.7 on $\Delta_{2,1}$, we have:

$$\begin{aligned}\|\Delta_2\|_F &\leq \left\| \Delta_{2,1} \Delta_{2,1}^\top + \Delta_{2,1} \mathbb{E}[\phi(X^\psi)^\top] + \mathbb{E}[\phi(X_n^\psi)] \Delta_{2,1}^\top \right\|_F \\ &\leq \|\Delta_{2,1} \Delta_{2,1}^\top\|_F + 2 \|\Delta_{2,1}\| \|\mathbb{E}[\phi(X^\psi)]\| \\ &\leq \alpha^{2m} \sigma^2 C_\chi^2 \|\psi - \psi^*\|^2 + 2 \|\log Z\|_{4,\infty} \alpha^m C_\chi \sigma \|\psi - \psi^*\|,\end{aligned}$$

where we used that $\|\Delta_{2,1} \Delta_{2,1}^\top\|_F = \|\Delta_{2,1}\|^2$. We can now combine our two matrix moment bounds to obtain

$$\begin{aligned}\left\| \text{Cov}[\phi(K_\psi^m(X_n))] - \text{Cov}[\phi(X^\psi)] \right\|_F &= \|\Delta_1 + \Delta_2\|_F \\ &\leq \alpha^m C_\chi (\bar{\tau}^{1/2} + 2 \|\log Z\|_{4,\infty} \sigma) \|\psi - \psi^*\| + \alpha^{2m} C_\chi^2 \sigma^2 \|\psi - \psi^*\|^2.\end{aligned}$$

□

Lemma IV.D.9. *Under Assumptions A0, A1, A2 it holds that*

$$\begin{aligned}\sqrt{\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}(\psi_{i-1}) - h_i(\psi_{i-1})) \right\|^2 \right]} &\leq 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n}} \\ &+ \frac{\|\mathcal{I}(\psi^*)^{-2}\|_F^{1/2}}{n} \left((M + \alpha^m C_\chi (\bar{\tau}^{1/2} + 2 \|\log Z\|_{4,\infty} \sigma))^{1/2} \left(\sum_{i=1}^n \delta_{i-1}^{1/2} \right)^{1/2} \right. \\ &\left. + \alpha^{2m} C_\chi^2 \sigma^2 \left(\sum_{i=1}^n \delta_{i-1} \right)^{1/2} \right)\end{aligned}$$

where $M := \sup_{\psi \in \Psi} \|\nabla^3 \log Z(\psi)\|_{\text{op}(\|\cdot\|_F, \|\cdot\|_F)} < +\infty$.

Proof. Let us first note that

$$\begin{aligned}f''(\psi^*)^{-1} (\bar{h}(\psi_{n-1}) - h_n(\psi_{n-1})) &= f''(\psi^*)^{-1} \underbrace{(\phi(X_n) - \mathbb{E}[\phi(X_1)])}_{\Delta_{1,n}} \\ &+ f''(\psi^*)^{-1} \underbrace{\left(\phi(K_{\psi_{n-1}}^m(X_n)) - \mathbb{E} \left[\phi(K_{\psi_{n-1}}^m(X_n)) \mid \mathcal{F}_{n-1} \right] \right)}_{\Delta_{2,n}} \\ &= f''(\psi^*)^{-1} \Delta_{1,n} + f''(\psi^*)^{-1} \Delta_{2,n}.\end{aligned}$$

Noting Δ the l.h.s of Lemma IV.D.9, we have, summing over $[n]$, and using

Minkowski's inequality,

$$\Delta \leq \frac{1}{n} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n f''(\psi^\star)^{-1} \Delta_{1,i} \right\|^2} + \frac{1}{n} \sqrt{\mathbb{E} \left\| \sum_{i=1}^n f''(\psi^\star)^{-1} \Delta_{2,i} \right\|^2}$$

Note that this step was made possible because we are looking at the square-root of the variance, which is unlike the recursion in Lemma IV.3.1. This allows to separate the terms and use fewer intermediaries than in the proof of Lemma IV.3.1.

Since both $\Delta_{1,n}$ and $\Delta_{2,n}$ are martingale differences with respect to the filtration \mathcal{F}_{n-1} ,

the covariance terms vanish, and we have

$$\begin{aligned}
& \Delta \\
& \leq \frac{1}{n} \sqrt{\sum_{i=1}^n \mathbb{E} [\|f''(\psi^\star)^{-1} \Delta_{1,i}\|^2]} + \frac{1}{n} \sqrt{\sum_{i=1}^n \mathbb{E} [\|f''(\psi^\star)^{-1} \Delta_{2,i}\|^2]} \\
& \leq \frac{1}{n} \sqrt{\sum_{i=1}^n \text{tr}(f''(\psi^\star)^{-1} \mathbb{E} [\Delta_{1,i} \Delta_{1,i}^\top f''(\psi^\star)^{-1}])} \\
& \quad + \frac{1}{n} \sqrt{\sum_{i=1}^n \text{tr}(f''(\psi^\star)^{-1} \mathbb{E} [\Delta_{2,i} \Delta_{2,i}^\top f''(\psi^\star)^{-1}])} \\
& \leq \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} \\
& \quad + \frac{1}{n} \sqrt{\sum_{i=1}^n \text{tr}(f''(\psi^\star)^{-1} \mathbb{E} [\text{Cov} [\phi(K_{\psi_{i-1}}^m(X_i)) | \mathcal{F}_{i-1}] f''(\psi^\star)^{-1}])} \\
& \leq \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} \\
& \quad + \frac{1}{n} \left(\sum_{i=1}^n \text{tr}(f''(\psi^\star)^{-1} \mathbb{E} [(\text{Cov} [\phi(X_i)] + (\text{Cov} [\phi(X_i^{\psi_{i-1}}) | \psi_{i-1}] - \text{Cov} [\phi(X_i)])) + (\text{Cov} [\phi(K_{\psi_{i-1}}^m(X_i)) | \mathcal{F}_{i-1}] - \text{Cov} [\phi(X^{\psi_{i-1}}) | \mathcal{F}_{i-1}])] f''(\psi^\star)^{-1} \right)^{1/2} \\
& \leq 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} \\
& \quad + \frac{\sqrt{\text{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n} \left(\sum_{i=1}^n \mathbb{E} [\|(\text{Cov} [\phi(X^{\psi_{i-1}}) | \mathcal{F}_{i-1}] - \text{Cov} [\phi(X_i)])\|_F + \mathbb{E} [\|\text{Cov} [\phi(K_{\psi}^m(X_i)) | \psi_{i-1}] - \text{Cov} [\phi(X^{\psi_{i-1}}) | \mathcal{F}_{i-1}]\|_F] \right)^{1/2} \\
& \stackrel{(a)}{\leq} 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} \\
& \quad + \frac{\sqrt{\text{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n} \left((M + \alpha^m C_\chi (\bar{\tau}^{1/2} + 2 \|\log Z\|_{4,\infty} \sigma)^{1/2}) \times \sqrt{\sum_{i=1}^n \mathbb{E} [\|\psi_{i-1} - \psi^\star\|^2]} + \alpha^m C_\chi \sigma \sqrt{\sum_{i=1}^n \mathbb{E} [\|\psi_{i-1} - \psi^\star\|^2]} \right) \\
& \stackrel{(b)}{\leq} 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^\star)^{-1})}{n}} \\
& \quad + \frac{\sqrt{\text{tr}(\mathcal{I}(\psi^\star)^{-2})}}{n} \left((M + \alpha^m C_\chi (\bar{\tau}^{1/2} + 2 \|\log Z\|_{4,\infty} \sigma))^{1/2} \sqrt{\sum_{i=1}^n \delta_i^{1/2}} + \alpha^m C_\chi \sigma \sqrt{\sum_{i=1}^n \delta_i} \right)
\end{aligned}$$

In (a) we used Lemma IV.D.9, the cyclicity of the trace, $\text{tr}(A^\top B) \leq \|AB\|_F \leq \|A\|_F \|B\|_F$, and the fact that since $\text{Cov}_{p_{\psi_{i-1}}} [\phi(X_i)] = \nabla_\psi^2 \mathcal{L}(\psi_{i-1})$, by analyticity of \mathcal{L} ,

there exists a constant $M := \sup_{\psi \in \Psi} \|\nabla^3 \log Z(\psi)\|_{\text{op}(\|\cdot\|, \|\cdot\|_F)}$ such that

$$\|\nabla_\psi^2 \mathcal{L}(\psi_{i-1}) - \nabla_\psi^2 \mathcal{L}(\psi^*)\|_F \leq M \|\psi_{i-1} - \psi^*\|.$$

In (b) we used Jensen's inequality to get $\mathbb{E}[\|\psi_{i-1} - \psi^*\|] \leq \sqrt{\mathbb{E}[\|\psi_{i-1} - \psi^*\|^2]} = \delta_{i-1}^{1/2}$. \square

We are now ready to prove Theorem IV.3.3.

Proof of Theorem IV.3.3. It holds that:

$$\begin{aligned} & f''(\psi^*)(\psi_{n-1} - \psi^*) \\ &= f'(\psi_{n-1}) - f'(\psi^*) + (f''(\psi^*)(\psi_{n-1} - \psi^*) - f'(\psi_{n-1}) + f'(\psi^*)) \\ &= h_n(\psi_{n-1}) - f'(\psi^*) + (f''(\psi^*)(\psi_{n-1} - \psi^*) - f'(\psi_{n-1}) + f'(\psi^*)) \\ &\quad + (f'(\psi_{n-1}) - \bar{h}(\psi_{n-1})) + (\bar{h}(\psi_{n-1}) - h_n(\psi_{n-1})). \end{aligned}$$

Applying on both sides: (a) a summation over $i \in [n]$, (b) a multiplication by $f''(\psi^*)^{-1}$, (c) $\sqrt{\mathbb{E}[\|\cdot\|^2]}$, and using Minkowski's inequality on the r.h.s, we obtain

$$\begin{aligned} & \sqrt{\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n \psi_i - \psi^*\right\|^2\right]} \leq \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} h_i(\psi_{i-1})\right\|^2\right]}}_{(i)} \\ &+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (f''(\psi^*)(\psi_{i-1} - \psi^*) - f'(\psi_{i-1}))\right\|^2\right]}}_{(ii)} \\ &+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (f'(\psi_{i-1}) - \bar{h}(\psi_{i-1}))\right\|^2\right]}}_{(iii)} \\ &+ \underbrace{\sqrt{\mathbb{E}\left[\left\|\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}(\psi_{i-1}) - h_i(\psi_{i-1}))\right\|^2\right]}}_{(iv)}. \end{aligned}$$

(i) and (ii) have direct analogues in the proofs of prior work [169] on the convergence of (unbiased) SGD with Polyak-Ruppert averaging, and will be bounded similarly. (iii) captures the bias of the CD algorithm, while (iv) captures the variance.

Bounding (i) Using Lemma IV.D.6, we have (i) = $\mathcal{O}\left(n^{\max\left(\frac{\beta}{2}-1, -\frac{2\beta}{2\beta+1}\right)}\right)$.

Bounding (ii) Since $\log Z(\psi)$ is analytic, there exists some constant M' such that

$$\|f''(\psi^*)(\psi_{i-1} - \psi^*) - f'(\psi_{i-1})\| \leq M' \|\psi_{i-1} - \psi^*\|^2.$$

Thus, we have:

$$(ii) \leq \frac{M'}{n} \sqrt{\mathbb{E} \left[\left(\sum_{i=1}^n \|\psi_i - \psi^*\|^2 \right)^2 \right]} \leq \frac{M'}{n} \sum_{i=1}^n \sqrt{\mathbb{E} \left[\|\psi_i - \psi^*\|^4 \right]} = \mathcal{O}(n^{-\beta})$$

where the second-to-last inequality used Minkowski's inequality, and the last applied Lemma V.D.3.

Bounding (iii) By Minkowski's inequality, we have:

$$(iii) \leq \frac{1}{n} \sum_{i=1}^n \sqrt{\mathbb{E} \left[\|f'(\psi_{i-1}) - \bar{h}(\psi_{i-1})\|^2 \right]}$$

Moreover, using lemma IV.C.7, we have:

$$\begin{aligned} & \|f'(\psi_{i-1}) - \bar{h}(\psi_{i-1})\| \\ &= \left\| \mathbb{E} [\phi(X^{\psi_{i-1}}) | \mathcal{F}_{i-1}] - \mathbb{E} \left[P_{\psi_{i-1}}^m \phi(X_i) | \mathcal{F}_{i-1} \right] \right\| \\ &\leq \alpha^m \sqrt{\mathbb{E} \left[\left\| \phi(X^{\psi_{i-1}}) - \mathbb{E} [\phi(X^{\psi_{i-1}}) | \mathcal{F}_{i-1}] \right\|^2 \middle| \mathcal{F}_{i-1} \right]} C_\chi \|\psi_{i-1} - \psi^*\| \\ &\leq \alpha^m \sigma C_\chi \|\psi_{i-1} - \psi^*\|. \end{aligned}$$

We thus obtain

$$(iii) \leq \frac{\alpha^m C_\chi}{n} \sum_{i=1}^n (\mathbb{E} [\|\psi_{i-1} - \psi^*\|^2])^{1/2} = \frac{\alpha^m C_\chi}{n} \sum_{i=1}^n \delta_{i-1}^{1/2}.$$

Recalling that δ_i satisfies Theorem V.D.1, we have that $\delta_n^{1/2} = \mathcal{O}(n^{-\beta/2})$, and we thus have $\sum_{i=1}^n \delta_i^{1/2} = \mathcal{O}(n^{1-\frac{\beta}{2}})$. By squaring the result of Lemma IV.D.5,

we have $\sum_{i=1}^n \delta_i^{1/2} = \mathcal{O}(n^{-1-\frac{\beta}{2}})$, and thus, we obtain that (iii) = $\mathcal{O}(\alpha^m n^{-\beta/2}) = \mathcal{O}(n^{-(\frac{\beta}{2} + m \frac{|\log \alpha|}{\log n}))}$.

Bounding (iv) We have

$$\begin{aligned}
 & (iv) \\
 & \stackrel{(a)}{\leq} 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n}} + \frac{\|\mathcal{I}(\psi^*)^{-2}\|_{\text{F}}^{1/2}}{n} \left((M + \alpha^m C_\chi (\bar{\tau} + 2\|\log Z\|_{4,\infty} \sigma))^{1/2} \right. \\
 & \quad \times \sqrt{\sum_{i=1}^n \delta_{i-1}^{1/2}} + \alpha^m C_\chi \sigma \sqrt{\sum_{i=1}^n \delta_{i-1}} \Big) \\
 & \stackrel{(b)}{\leq} 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n}} + \mathcal{O}(n^{-\frac{1}{2}-\frac{\beta}{4}}).
 \end{aligned}$$

Where in (a), we used Lemma IV.D.9 and in (b), we used Lemma IV.D.5 and IV.D.4.

Final bound Putting everything together, we have that:

$$\begin{aligned}
 & \sqrt{\mathbb{E} \|\bar{\psi}_n - \psi^*\|^2} \\
 & \leq 2 \sqrt{\frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n}} + \mathcal{O}(n^{\max\left(-\left(\frac{1}{2} + \frac{\beta}{4}\right), -\beta, \frac{\beta}{2} - 1, -\frac{2\beta}{2\beta+1}, -\left(\frac{\beta}{2} + m \frac{|\log \alpha|}{\log n}\right)\right)}).
 \end{aligned}$$

If, furthermore, $m > \frac{(1-\beta)\log n}{2|\log \alpha|}$, we have

$$\max\left(-\left(\frac{1}{2} + \frac{\beta}{4}\right), -\beta, \frac{\beta}{2} - 1, -\frac{2\beta}{2\beta+1}, -\left(\frac{\beta}{2} + m \frac{|\log \alpha|}{\log n}\right)\right) < -\frac{1}{2},$$

which concludes the proof. \square

IV.E L_2 approximation by auxiliary gradient updates

In this section, we consider different gradient update schemes starting from some random initialization θ_{init} , and control the L_2 distance between the different updates and the deterministic target $\psi^* \in \Psi$.

Notation. Let θ^{init} be some Ψ -valued random initialization that is possibly correlated with X_1, \dots, X_n . We capture the effect of correlation through the following quantities:

For $\varepsilon > 0$ and $v > 2$, let

$$\vartheta_{n,m}^{\text{init}}(\varepsilon) := \mathbb{P}\left(\frac{\left\|\sum_{i=1}^n \mathbb{E}[\phi(K_{\theta^{\text{init}}}^m(X_i)) | X_i, \theta^{\text{init}}] - \mathbb{E}[\phi(K_{\theta^{\text{init}}}^m(X'_i)) | \theta^{\text{init}}]\right\|}{n} > \varepsilon\right),$$

and $\varepsilon_{n,m;v}^{\text{init}}(\varepsilon) := \sqrt{\varepsilon^2 + \kappa_{v,m}^2(\vartheta_{n,m}^{\text{init}}(\varepsilon))^{\frac{2}{v-2}}}.$

For notational clarity, we shall use $\theta_{m,B}$ to denote parameters arising from a one-step update, where the subscripts m, B represent performing the one-step update with length- m Markov chains and with batch size B . This is to be distinguished from ψ_t elsewhere in the text, which denotes the parameter from the actual multi-step CD algorithm and the subscript t denotes the t -th CD iterate.

Gradient update schemes. We consider five different updates. Let X'_1 be an i.i.d. copy of X_1 drawn independently of all other random variables. The SGD-with-replacement update is given by

$$\theta_{m,B}^{\text{SGDw}} := F_{m,B}^{\text{SGDw}}(\theta^{\text{init}}), \text{ where } F_{m,B}^{\text{SGDw}}(\psi) := \psi - \frac{\eta}{B} \sum_{i \in S^w} \left(\phi(X_i) - \phi(K_{i;\psi}^m(X_i)) \right)$$

and S^w is a uniformly drawn size- B subset of $[n]$. The SGD-without-replacement update, after renormalizing the learning rate, is given by the N -fold function composition

$$\theta_{m,B}^{\text{SGDo}} := F_{m,B;N}^{\text{SGDo}} \circ \dots \circ F_{m,B;1}^{\text{SGDo}}(\theta^{\text{init}}),$$

where $F_{m,B;j}^{\text{SGDo}}(\psi) := \psi - \frac{\eta}{NB} \sum_{i \in S_j^o} \left(\phi(X_i) - \phi(K_{i;\psi}^m(X_i)) \right)$ for each $j \in [N]$,

and S_j^o 's are disjoint size- B random subsets of $[n]$, defined by $(S_1^o, \dots, S_N^o) = \pi([n])$ for a uniformly drawn element π of the permutation group on n objects. The full-batch gradient update is given by

$$\theta_m^{\text{GD}} := F_m^{\text{GD}}(\theta^{\text{init}}), \text{ where } F_m^{\text{GD}}(\psi) := \psi - \frac{\eta}{n} \sum_{i \leq n} \left(\phi(X_i) - \phi(K_{i;\psi}^m(X_i)) \right).$$

The full-batch gradient update with an infinite-length Markov chain is given by

$$\theta_\infty^{\text{GD}} := F_\infty^{\text{GD}}(\theta^{\text{init}}), \quad \text{where} \quad F_\infty^{\text{GD}}(\psi) := \psi - \frac{\eta}{n} \sum_{i \leq n} (\phi(X_i) - \phi(X_1^\psi)).$$

The population gradient update with an infinite-length Markov chain is given by

$$\theta^{\text{pop}} := f^{\text{pop}}(\theta^{\text{init}}), \quad \text{where} \quad f^{\text{pop}}(\psi) := \psi - \eta \mathbb{E}[\phi(X_1) - \phi(X_1^\psi)],$$

where we use the lowercase f to emphasize that f^{pop} is a deterministic function.

The forthcoming results are summarized below:

$$\theta_{m,B}^{\text{SGDw}} \stackrel{\text{Lemma IV.E.1}}{\approx} \theta_m^{\text{GD}} \stackrel{\text{Lemma IV.E.2}}{\approx} \theta_\infty^{\text{GD}} \stackrel{\text{Lemma IV.E.3}}{\approx} \theta^{\text{pop}} \stackrel{\text{Lemma IV.E.4}}{\approx} \psi^*.$$

Figure IV.E.1: Overview of approximation results between different updates

Lemma IV.E.1. Let \mathcal{F}_n be the sigma algebra generated by $\{X_i, K_{i;\theta^{\text{init}}}^m(X_i) \mid 1 \leq i \leq n\}$.

Then

$$\mathbb{E}[\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}} \mid \theta^{\text{init}}, \mathcal{F}_n] = 0 \quad \text{almost surely}.$$

Moreover, under **A0** and **A6**, we have

$$\mathbb{E}\|\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}\|^2 \leq \frac{4\eta^2(\sigma^2 + \kappa_{v,m}^2)}{B} \mathbb{I}_{\{B < n\}}.$$

Proof of Lemma IV.E.1. Write $A := (A_1, \dots, A_n)$, where

$$A_i := (\phi(X_i) - \phi(K_{i;\theta^{\text{init}}}^m(X_i))) - \mathbb{E}[\phi(X_1) - \phi(K_{1;\theta^{\text{init}}}^m(X_1))].$$

Since S^w is uniformly drawn from all size- B subsets of $[n]$ and independently of all other variables, we have that almost surely

$$\mathbb{E}[\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}} \mid \theta^{\text{init}}, \mathcal{F}_n] = \mathbb{E}\left[\frac{\eta}{B} \sum_{i \in S^w} A_i - \frac{\eta}{n} \sum_{i \leq n} A_i \mid A\right] = 0.$$

To prove the remaining bound, we note that the above relation implies $\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}$

is zero-mean. By the law of total variance, we have

$$\begin{aligned} \mathbb{E}\|\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}\|^2 &= \text{TrCov}[\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}] \\ &= \text{Tr}\mathbb{E}\text{Cov}[\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}} \mid \theta^{\text{init}}, \mathcal{F}_n] \\ &= \eta^2 \text{Tr}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{B}\sum_{i \in S^w} A_i\right)\left(\frac{1}{B}\sum_{i \in S^w} A_i\right)^\top \middle| A\right] - \left(\frac{1}{n}\sum_{i \leq n} A_i\right)\left(\frac{1}{n}\sum_{i \leq n} A_i\right)^\top\right]. \end{aligned}$$

To compute the covariance, recall that S^w is a uniformly drawn size- B subset of $[n]$ with $B = n/N$. Let $\mathcal{P}_N([n])$ be the collection of all partitions of $[n]$ into N size- B subsets. We can generate S^w by the following two-step process:

- (i) Uniformly draw a partition $P' = (P'_1, \dots, P'_N)$ from $\mathcal{P}_N([n])$;
- (ii) Uniformly sample an index K from $[N]$ and set $S^w = P'_K$.

Then we have, almost surely

$$\begin{aligned} &\mathbb{E}\left[\left(\frac{1}{B}\sum_{i \in S^w} A_i\right)\left(\frac{1}{B}\sum_{i \in S^w} A_i\right)^\top \middle| A\right] - \left(\frac{1}{n}\sum_{i \leq n} A_i\right)\left(\frac{1}{n}\sum_{i \leq n} A_i\right)^\top \\ &= \frac{1}{|\mathcal{P}_N([n])|} \sum_{P' \in \mathcal{P}_N([n])} \left(\frac{1}{N} \sum_{k \leq N} \frac{1}{B^2} \sum_{i,j \in P'_k} A_i A_j^\top - \frac{1}{n^2} \sum_{i,j \leq n} A_i A_j^\top \right) \\ &= \frac{1}{|\mathcal{P}_N([n])|} \sum_{P' \in \mathcal{P}_N([n])} \left(\frac{1}{NB^2} \sum_{k \leq N} \sum_{i,j \in P'_k} A_i A_j^\top - \frac{1}{N^2 B^2} \sum_{k,l \leq N} \sum_{i \in P'_k, j \in P'_l} A_i A_j^\top \right) \\ &= \frac{1}{|\mathcal{P}_N([n])|} \sum_{P' \in \mathcal{P}_N([n])} \left(\frac{N-1}{N^2 B^2} \sum_{k \leq N} \sum_{i,j \in P'_k} A_i A_j^\top - \frac{1}{N^2 B^2} \sum_{k \neq l} \sum_{i \in P'_k, j \in P'_l} A_i A_j^\top \right). \end{aligned}$$

By noting that A_i 's are exchangeable, we obtain

$$\begin{aligned} \mathbb{E}\|\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}\|^2 &= \eta^2 \text{Tr}\mathbb{E}\left[\frac{N-1}{NB} A_1 A_1^\top + \frac{(N-1)(B-1)}{NB} A_1 A_2^\top - \frac{N-1}{N} A_1 A_2^\top\right] \\ &= \eta^2 \text{Tr}\mathbb{E}\left[\frac{N-1}{NB} A_1 A_1^\top - \frac{N-1}{NB} A_1 A_2^\top\right] \\ &= \frac{\eta^2(N-1)}{NB} (\mathbb{E}\|A_1\|^2 - \mathbb{E}\langle A_1, A_2 \rangle) \\ &\stackrel{(a)}{\leq} \frac{2\eta^2}{B} \mathbb{E}\|A_1\|^2 \\ &= \frac{2\eta^2}{B} \mathbb{E}\left\|(\phi(X_1) - \mathbb{E}[\phi(X_1)]) - \left(\phi(K_{i;\theta^{\text{init}}}(X_1)) - \mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X_1))]\right)\right\|^2 \\ &\leq \frac{4\eta^2}{B} \left(\text{TrCov}[\phi(X_1)] + \text{TrCov}\left[\phi(K_{i;\theta^{\text{init}}}(X_1))\right]\right) \\ &\stackrel{(b)}{\leq} \frac{4\eta^2(\sigma^2 + \kappa_{v,m}^2)}{B}. \end{aligned}$$

In (a), we have used a Cauchy-Schwarz inequality; in (b), we have used **A0** and **A6**. Finally, we note that if $B = n$, $\theta_{m,B}^{\text{SGDw}} = \theta_m^{\text{GD}}$ almost surely, which implies the desired bound. \square

Lemma IV.E.2. Denote \mathcal{A}_n as the sigma algebra generated by $\theta^{\text{init}}, X_1, \dots, X_n$. Under **A0**, **A1**, **A2** and **A6**, we have that for any $\varepsilon > 0$ and $v > 2$,

$$\begin{aligned}\mathbb{E} \|\mathbb{E}[\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} | \mathcal{A}_n]\|^2 &\leq \eta^2 \left(\alpha^m \sigma C_\chi \sqrt{\mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2} + \varepsilon_{n,m;v}^{\text{init}}(\varepsilon) \right)^2, \\ \mathbb{E} \|\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}}\|^2 &\leq \eta^2 \left(\left(\alpha^m \sigma C_\chi \sqrt{\mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2} + \varepsilon_{n,m;v}^{\text{init}}(\varepsilon) \right)^2 + \frac{\kappa_{v,m}^2 + \sigma^2}{n} \right).\end{aligned}$$

Proof of Lemma IV.E.2. The main challenge arises from the possible correlation between θ^{init} and X_1, \dots, X_n . First note that for any $\varepsilon > 0$, $v > 2$ and a real-valued random variable Y , by Hölder's inequality, we have

$$\begin{aligned}\mathbb{E}[Y^2] &= \mathbb{E}[Y^2 \mathbb{I}_{\{Y \leq \varepsilon\}} + Y^2 \mathbb{I}_{\{Y > \varepsilon\}}] \\ &\leq \varepsilon^2 + \mathbb{E}[Y^2 \mathbb{I}_{\{Y > \varepsilon\}}] \leq \varepsilon^2 + (\mathbb{E}[Y^v])^{2/v} \mathbb{P}(Y > \varepsilon)^{(v-2)/v}.\end{aligned}\quad (\text{IV.19})$$

Also note the useful inequality that for two real-valued random vectors (possibly correlated) V_1, V_2 , we have

$$\begin{aligned}\mathbb{E}[\|V_1 + V_2\|^2] &\leq \mathbb{E}[(\|V_1\| + \|V_2\|)^2] \leq \mathbb{E}\|V_1\|^2 + 2\sqrt{(\mathbb{E}\|V_1\|^2)(\mathbb{E}\|V_2\|^2)} + \mathbb{E}\|V_2\|^2 \\ &= (\sqrt{\mathbb{E}\|V_1\|^2} + \sqrt{\mathbb{E}\|V_2\|^2})^2.\end{aligned}\quad (\text{IV.20})$$

Now to control the first quantity of interest, by using a triangle inequality, we have

$$\begin{aligned}\mathbb{E} \|\mathbb{E}[\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} | \mathcal{A}_n]\|^2 &= \mathbb{E} \left\| \mathbb{E} \left[\frac{\eta}{n} \sum_{i \leq n} (\phi(K_{i;\theta^{\text{init}}}(X_i)) - \phi(X_i^{\theta^{\text{init}}})) \mid \mathcal{A}_n \right] \right\|^2 \\ &\stackrel{(\text{IV.20})}{\leq} \eta^2 \left(\sqrt{\mathbb{E}[\Delta_1^2]} + \sqrt{\mathbb{E}[\Delta_2^2]} \right)^2,\end{aligned}$$

where

$$\Delta_1 := \left\| \mathbb{E} \left[\frac{1}{n} \sum_{i \leq n} (\phi(K_{i;\theta^{\text{init}}}(X_i)) - \mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X'_i)) \mid \theta^{\text{init}}]) \mid \mathcal{A}_n \right] \right\|$$

$$\begin{aligned}
&= \left\| \frac{1}{n} \sum_{i \leq n} (\mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X_i)) \mid \theta^{\text{init}}, X_i] - \mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X'_i)) \mid \theta^{\text{init}}]) \right\|, \\
\Delta_2 &:= \left\| \mathbb{E}\left[\frac{1}{n} \sum_{i \leq n} (\mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X'_i)) \mid \theta^{\text{init}}] - \phi(X_i^{\theta^{\text{init}}})) \mid \mathcal{A}_n \right] \right\| \\
&= \left\| \mathbb{E}[\phi(K_{1;\theta^{\text{init}}}(X'_1)) - \phi(X_1^{\theta^{\text{init}}}) \mid \theta^{\text{init}}] \right\|,
\end{aligned}$$

and X'_1 is an i.i.d. copy of X_1 and in particular independent of θ^{init} . Δ_1 is controlled via (IV.19):

$$\begin{aligned}
\mathbb{E}[\Delta_1^2] &\leq \varepsilon^2 + (\mathbb{E}[\Delta_1^\nu])^{2/\nu} \mathbb{P}(\Delta_1 > \varepsilon)^{\nu/(\nu-2)} \\
&\stackrel{(a)}{\leq} \varepsilon^2 + \left(\mathbb{E} \left\| \phi(K_{1;\theta^{\text{init}}}(X_1)) - \mathbb{E}[\phi(K_{1;\theta^{\text{init}}}(X'_1)) \mid \theta^{\text{init}}] \right\|^{\nu} \right)^{2/\nu} \\
&\quad \times \mathbb{P} \left(\frac{\|\sum_{i \leq n} (\mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X_i)) \mid \theta^{\text{init}}, X_i] - \mathbb{E}[\phi(K_{i;\theta^{\text{init}}}(X'_i)) \mid \theta^{\text{init}}])\|}{n} > \varepsilon \right)^{\frac{\nu-2}{\nu}} \\
&\stackrel{(b)}{\leq} \varepsilon^2 + \kappa_{\nu;m}^2 (\vartheta_{n,m}^{\text{init}}(\varepsilon))^{\frac{\nu-2}{\nu}} \\
&= (\varepsilon_{n,m;\nu}^{\text{init}}(\varepsilon))^2. \tag{IV.21}
\end{aligned}$$

In (a), we have plugged in the definition of Δ_1 and applied a Jensen's inequality with respect to the empirical average; in (b), we have used A6 to bound the ν -th moment term as

$$\begin{aligned}
&\left(\mathbb{E} \left\| \phi(K_{1;\theta^{\text{init}}}(X_1)) - \mathbb{E}[\phi(K_{1;\theta^{\text{init}}}(X'_1)) \mid \theta^{\text{init}}] \right\|^{\nu} \right)^{1/\nu} \\
&= \left(\mathbb{E} \left[\mathbb{E} \left[\left\| \phi(K_{1;\theta^{\text{init}}}(X_1)) - \mathbb{E}[\phi(K_{1;\theta^{\text{init}}}(X'_1)) \mid \theta^{\text{init}}] \right\|^{\nu} \mid \theta^{\text{init}} \right] \right] \right)^{1/\nu} \\
&\leq \sup_{\psi \in \Psi} \left(\mathbb{E} \left[\left\| \phi(K_{1;\psi}(X_1)) - \mathbb{E}[\phi(K_{1;\psi}(X'_1))] \right\|^{\nu} \right] \right)^{1/\nu} \\
&= \sup_{\psi \in \Psi} \left(\mathbb{E} \left[\left\| \phi(K_{1;\psi}(X_1)) - \mathbb{E}[\phi(K_{1;\psi}(X_1))] \right\|^{\nu} \right] \right)^{1/\nu} \leq \kappa_{\nu;m}
\end{aligned}$$

and recalled the definitions of $\vartheta_{n,m}^{\text{init}}$ and $\varepsilon_{n,m;\nu}^{\text{init}}$. On the other hand,

$$\begin{aligned}
\mathbb{E}[\Delta_2^2] &= \mathbb{E} \left\| \int_{\mathbb{R}^d} \phi(x) (K_{1;\theta^{\text{init}}}^m p_{\psi^*})(x) dx - \mathbb{E}[\phi(X_1^{\theta^{\text{init}}}) \mid \theta^{\text{init}}] \right\|^2 \\
&\stackrel{(a)}{=} \mathbb{E} \left\| \int_{\mathbb{R}^d} (K_{1;\theta^{\text{init}}}^m \phi)(x) p_{\psi^*}(x) dx - \mathbb{E}_{p_{\theta^{\text{init}}}} [\phi] \right\|^2 \\
&\stackrel{(b)}{=} \mathbb{E} \left\| \int_{\mathbb{R}^d} (K_{1;\theta^{\text{init}}}^m (\phi - \mathbb{E}_{p_{\theta^{\text{init}}}} [\phi]))(x) \times p_{\psi^*}(x) dx \right\|^2
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{=} \mathbb{E} \left\| \int_{\mathbb{R}^d} (K_{1;\theta^{\text{init}}}^m(\phi - \mathbb{E}_{p_{\theta^{\text{init}}}}[\phi]))(x) \times (p_{\psi^*}(x) - p_{\theta^{\text{init}}}(x)) dx \right\|^2 \\
&\stackrel{(d)}{=} \sum_{l=1}^d \mathbb{E} \left(\int_{\mathbb{R}^d} (K_{t1;\theta^{\text{init}}}^m(\phi - \mathbb{E}_{p_{\theta^{\text{init}}}}[\phi]))(x)^\top e_l \times p_{\theta^{\text{init}}}(x) \times \frac{p_{\psi^*}(x) - p_{\theta^{\text{init}}}(x)}{p_{\theta^{\text{init}}}(x)} dx \right)^2 \\
&\stackrel{(e)}{\leq} \sum_{l=1}^d \mathbb{E} \left[\left(\int_{\mathbb{R}^d} ((K_{t1;\theta^{\text{init}}}^m(\phi - \mathbb{E}_{p_{\theta^{\text{init}}}}[\phi]))(x)^\top e_l)^2 p_{\theta^{\text{init}}}(x) dx \right. \right. \\
&\quad \times \left. \left. \int_{\mathbb{R}^d} \left(\frac{p_{\psi^*}(x) - p_{\theta^{\text{init}}}(x)}{p_{\theta^{\text{init}}}(x)} \right)^2 p_{\theta^{\text{init}}}(x) dx \right)^2 \right] \\
&\stackrel{(f)}{\leq} \sum_{l=1}^d \mathbb{E} \left(\alpha^{2m} \text{TrCov}[\phi(X_1^{\theta^{\text{init}}}) \mid \theta^{\text{init}}] \chi^2(p_{\psi^*}, p_{\theta^{\text{init}}}) \right)^2 \\
&\stackrel{(g)}{\leq} \alpha^{2m} \sigma^2 C_\chi^2 \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2.
\end{aligned}$$

In (a), we have used that $(Kf)(x) = \int K(x, y) f(y) dy$; in (b), we have used that the Markov operator leaves the constant function invariant; in (c), we used that $K_{1;\theta^{\text{init}}}^m$ leaves $p_{\theta^{\text{init}}}$ invariant; in (d), we denoted $(e_l)_{l \leq d}$ as the standard basis vectors of \mathbb{R}^d and multiplied and divided by $p_{\theta^{\text{init}}}(x)$; in (e), we have used a Cauchy-Schwarz inequality; in (f), we have used the definition of the spectral gap α in A2; in (g), we have used A0 and A1. Combining the bounds gives the first inequality that

$$\mathbb{E} \|\mathbb{E}[\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} \mid \mathcal{A}_n]\|^2 \leq \eta^2 \left(\alpha^m \sigma C_\chi \sqrt{\mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2} + \varepsilon_{n,m;v}^{\text{init}}(\varepsilon) \right)^2.$$

Now we handle the second quantity by conditioning on θ^{init} and perform a bias-variance decomposition:

$$\begin{aligned}
\mathbb{E} \|\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}}\|^2 &= \eta^2 \mathbb{E} \left[\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i \leq n} (\phi(K_{i;\theta^{\text{init}}}^m(X_i)) - \phi(X_i^{\theta^{\text{init}}})) \right\|^2 \mid \mathcal{A}_n \right] \right] \\
&= \eta^2 (Q_B + Q_V),
\end{aligned}$$

where

$$\begin{aligned}
Q_B &:= \mathbb{E} \left\| \mathbb{E} \left[\frac{1}{n} \sum_{i \leq n} (\phi(K_{i;\theta^{\text{init}}}^m(X_i)) - \phi(X_i^{\theta^{\text{init}}})) \mid \mathcal{A}_n \right] \right\|^2, \\
Q_V &:= \mathbb{E} \left[\text{Tr} \left(\text{Cov} \left[\frac{1}{n} \sum_{i \leq n} \phi(K_{i;\theta^{\text{init}}}^m(X_i)) \mid \mathcal{A}_n \right] + \text{Cov} \left[\frac{1}{n} \sum_{i \leq n} \phi(X_i^{\theta^{\text{init}}}) \mid \theta^{\text{init}} \right] \right) \right].
\end{aligned}$$

Note that the covariance terms separate because $X_i^{\theta^{\text{init}}}$ is independent of $K_{i;\theta^{\text{init}}}^m(X_i)$ conditioning on θ^{init} . $\eta^2 Q_B$ is exactly the quantity controlled above, so it suffices to

bound the variance term Q_V . By explicitly computing the second covariance term while noting that $X_1^{\theta^{\text{init}}}, \dots, X_n^{\theta^{\text{init}}}$ are conditionally i.i.d. given θ^{init} and $K_{i;\theta^{\text{init}}}^m(X_i)$'s are conditionally independent across $1 \leq i \leq n$ given \mathcal{A}_n , we have

$$\begin{aligned} Q_V &= \frac{\sum_{i \leq n} \mathbb{E}[\text{TrCov}[\phi(K_{i;\theta^{\text{init}}}^m(X_i)) | \theta^{\text{init}}, X_i]]}{n^2} + \frac{\mathbb{E}[\text{TrCov}[\phi(X_1^{\theta^{\text{init}}}) | \theta^{\text{init}}]]}{n} \\ &\stackrel{(a)}{=} \frac{\mathbb{E}[\text{TrCov}[\phi(K_{1;\theta^{\text{init}}}^m(X_1)) | \theta^{\text{init}}, X_1]]}{n} + \frac{\mathbb{E}[\text{TrCov}[\phi(X_1^{\theta^{\text{init}}}) | \theta^{\text{init}}]]}{n} \\ &\leq \frac{\kappa_{v;m}^2 + \sigma^2}{n}, \end{aligned}$$

where we have used [A3](#), [A6](#) and [A0](#) in the last line. Combining the bounds, we obtain that

$$\begin{aligned} \mathbb{E} \|\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}}\|^2 &= \eta^2(Q_B + Q_V) \\ &\leq \eta^2 \left(\left(\alpha^m \sigma C_\chi \sqrt{\mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2} + \varepsilon_{n,m;v}^{\text{init}}(\varepsilon) \right)^2 + \frac{\kappa_{v;m}^2 + \sigma^2}{n} \right). \end{aligned}$$

□

Lemma IV.E.3. *Under [A0](#), $\mathbb{E} \|\theta_\infty^{\text{GD}} - \theta^{\text{pop}}\|^2 \leq \frac{4\eta^2\sigma^2}{n}$.*

Proof of Lemma IV.E.3. Since both F_∞^{GD} and f^{pop} involve infinite-length Markov chains, the initializations do not matter, and we can decouple the stochasticity of X_i and $K_{i;\theta^{\text{init}}}$. In particular,

$$\begin{aligned} \mathbb{E} \|\theta_\infty^{\text{GD}} - \theta^{\text{pop}}\|^2 &= \eta^2 \mathbb{E} \left\| \frac{1}{n} \sum_{i \leq n} (\phi(X_i) - \phi(X_i^{\theta^{\text{init}}})) - \mathbb{E}[\phi(X_1) - \phi(X_1^{\theta^{\text{init}}})] \right\|^2 \\ &\leq \eta^2 (Q'_1 + 2\sqrt{Q'_1 Q'_2} + Q'_2), \end{aligned}$$

where

$$\begin{aligned} Q'_1 &:= \mathbb{E} \left\| \frac{1}{n} \sum_{i \leq n} (\phi(X_i) - \mathbb{E}[\phi(X_1)]) \right\|^2 = \frac{\text{TrCov}[\phi(X_1)]}{n} = \frac{\text{Tr}\nabla_\theta^2 \log Z(\psi^*)}{n} \leq \frac{\sigma^2}{n}, \\ Q'_2 &:= \mathbb{E} \left\| \frac{1}{n} \sum_{i \leq n} (\phi(X_i^{\theta^{\text{init}}}) - \mathbb{E}[\phi(X_1^{\theta^{\text{init}}})]) \right\|^2 \\ &= \frac{\mathbb{E}[\text{TrCov}[\phi(X_1^{\theta^{\text{init}}}) | \theta^{\text{init}}]]}{n} = \frac{\mathbb{E}[\text{Tr}\nabla_\theta^2 \log Z(\theta^{\text{init}})]}{n} \leq \frac{\sigma^2}{n}. \end{aligned}$$

In the computations above, we have used the relation $\nabla_{\theta}^2 \log Z(\theta) = \text{Cov}_{X \sim p_{\theta}}[\phi(X)]$ and the assumption $\sup_{\theta \in \Psi} \text{tr}(\nabla_{\theta}^2 \log Z(\theta)) = \sigma^2$ from A0. This implies the desired bound. \square

Lemma IV.E.4. *Under A0, $\mathbb{E} \|\theta^{\text{pop}} - \psi^*\|^2 \leq (1 - 2\mu\eta + L^2\eta^2) \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2$.*

Proof of Lemma IV.E.4. Recall that

$$f^{\text{pop}}(\theta') = \theta - \eta \mathbb{E}[\phi(X_1) - \phi(X_1^{\theta'})] = \theta - \eta (\nabla_{\psi} \log Z(\psi^*) - \nabla_{\psi} \log Z(\theta')).$$

By construction, f^{pop} is deterministic and $f^{\text{pop}}(\psi^*) = \psi^*$. By plugging in the recursions and expanding the square, we get that

$$\begin{aligned} \mathbb{E} \|\theta^{\text{pop}} - \psi^*\|^2 &= \mathbb{E} \|f^{\text{pop}}(\theta^{\text{init}}) - f^{\text{pop}}(\psi^*)\|^2 \\ &= \mathbb{E} \|(\theta^{\text{init}} - \psi^*) - \eta (\nabla_{\psi} \log Z(\theta^{\text{init}}) - \nabla_{\psi} \log Z(\psi^*))\|^2 \\ &= \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2 - 2\eta \mathbb{E} [\langle \theta^{\text{init}} - \psi^*, \nabla_{\psi} \log Z(\theta^{\text{init}}) - \nabla_{\psi} \log Z(\psi^*) \rangle] \\ &\quad + \eta^2 \mathbb{E} \|\nabla_{\psi} \log Z(\theta^{\text{init}}) - \nabla_{\psi} \log Z(\psi^*)\|^2 \\ &\leq \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2 - 2\mu\eta \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2 + L^2\eta^2 \mathbb{E} \|\theta^{\text{init}} - \psi^*\|^2. \end{aligned}$$

In the last line, we have recalled $\inf_{\psi \in \Psi} \lambda_{\min}(\nabla_{\psi}^2 \log Z(\psi)) = \mu$ as well as $\sup_{\theta \in \Psi} \lambda_{\max}(\nabla_{\psi}^2 \log Z(\psi)) = L$ by A0 and applied Lemma IV.C.8. Combining the coefficients gives the desired statement. \square

IV.F Proofs for offline SGD

We prove Theorem IV.B.1 (which directly implies Theorem IV.4.3) and Theorem IV.B.2 in this section. The key ingredient of both proofs is Lemma IV.F.1 below, which provides an iterative error bound for the SGD-with-replacement scheme by combining different approximation bounds in Section IV.E. Throughout this section, we denote $\delta_{t,j}^{\text{SGDw}} := \mathbb{E} \|\psi_{t,j}^{\text{SGDw}} - \psi^*\|^2$.

Lemma IV.F.1. Under A0, A1, A2, A3 and A6, we have that for $1 \leq j \leq N - 1$,

$$\begin{aligned} & \sqrt{\delta_{t,j}^{\text{SGDw}}} \\ & \leq \left(1 - \eta_t \left(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2} \eta_t\right)\right) \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \eta_t \left(\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v,m}}{\sqrt{B}}\right), \end{aligned}$$

where $1 - \eta_t \left(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2} \eta_t\right) > 0$.

Proof of Lemma IV.F.1. We first remark that in view of Lemma IV.C.4, the projection step in Algorithm 2 does not increase $\delta_{t,j}^{\text{SGDw}}$, so it suffices to bound $\delta_{t,j}^{\text{SGDw}}$ as if projection is not performed on $\psi_{t,j}^{\text{SGDw}}$. To apply the results from Section IV.E, we identify $\theta^{\text{init}} = \psi_{t,j-1}^{\text{SGDw}}$ and $\eta = \eta_t$, which allows us to write $\psi_{t,j}^{\text{SGDw}} = \theta_{m,B}^{\text{SGDw}}$. This also implies $\mathbb{E}\|\theta^{\text{init}} - \psi^*\|^2 = \delta_{t,j-1}^{\text{SGDw}}$ and $\varepsilon_{n,m,t;v}^{\text{init}}(\varepsilon) \leq \varepsilon_{v;n,m,t}^{\text{SGDw}}(\varepsilon)$. By adding and subtracting the auxiliary gradient updates followed by expanding the square, we obtain

$$\begin{aligned} & \delta_{t,j}^{\text{SGDw}} \\ &= \mathbb{E}\|\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}} + \theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} + \theta_\infty^{\text{GD}} - \theta^{\text{pop}} + \theta^{\text{pop}} - \psi^*\|^2 \\ &\stackrel{(a)}{=} \mathbb{E}\|\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}\|^2 + \mathbb{E}\|\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}}\|^2 + \mathbb{E}\|\theta_\infty^{\text{GD}} - \theta^{\text{pop}}\|^2 + \mathbb{E}\|\theta^{\text{pop}} - \psi^*\|^2 \\ &\quad + 2\mathbb{E}\langle \theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}, \theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} \rangle + 2\mathbb{E}\langle \theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}}, \theta_\infty^{\text{GD}} - \theta^{\text{pop}} \rangle \\ &\quad + 2\mathbb{E}\langle \mathbb{E}[\theta_{m,B}^{\text{SGDw}} - \theta_m^{\text{GD}} | \theta^{\text{init}}, \mathcal{F}_n], \theta^{\text{pop}} - \psi^* \rangle + 2\mathbb{E}\langle \theta_m^{\text{GD}} - \theta_\infty^{\text{GD}}, \theta_\infty^{\text{GD}} - \theta^{\text{pop}} \rangle \\ &\quad + 2\mathbb{E}\langle \mathbb{E}[\theta_m^{\text{GD}} - \theta_\infty^{\text{GD}} | \mathcal{A}_n], \theta^{\text{pop}} - \psi^* \rangle + 2\mathbb{E}\langle \theta_\infty^{\text{GD}} - \theta^{\text{pop}}, \theta^{\text{pop}} - \psi^* \rangle \\ &\stackrel{(b)}{\leq} \frac{4\eta_t^2(\sigma^2 + \kappa_{v,m}^2)}{\sqrt{B}} \mathbb{I}_{\{B < n\}} + \eta_t^2 \left(\left(\alpha^m \sigma C_\chi \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) \right)^2 + \frac{\kappa_{v,m}^2 + \sigma^2}{n} \right) \\ &\quad + \frac{4\eta_t^2\sigma^2}{n} + (1 - 2\mu\eta_t + L^2\eta_t^2)\delta_{t,j-1}^{\text{SGDw}} \\ &\quad + 2\frac{2\eta_t\sqrt{\sigma^2 + \kappa_{v,m}^2}}{\sqrt{B}} \mathbb{I}_{\{B < n\}} \eta_t \left(\alpha^m \sigma C_\chi \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{\sqrt{\kappa_{v,m}^2 + \sigma^2}}{\sqrt{n}} \right) \\ &\quad + 2\frac{2\eta_t\sqrt{\sigma^2 + \kappa_{v,m}^2}}{\sqrt{B}} \mathbb{I}_{\{B < n\}} \frac{2\eta_t\sigma}{\sqrt{n}} + 0 \\ &\quad + 2\eta_t \left(\alpha^m \sigma C_\chi \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{\sqrt{\kappa_{v,m}^2 + \sigma^2}}{\sqrt{n}} \right) \frac{2\eta_t\sigma}{\sqrt{n}} \\ &\quad + 2\eta_t \left(\alpha^m \sigma C_\chi \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) \right) \sqrt{(1 - 2\mu\eta_t + L^2\eta_t^2)\delta_{t,j-1}^{\text{SGDw}}} \\ &\quad + 2\frac{2\eta_t\sigma}{\sqrt{n}} \sqrt{(1 - 2\mu\eta_t + L^2\eta_t^2)\delta_{t,j-1}^{\text{SGDw}}} \end{aligned}$$

$$\begin{aligned}
&\stackrel{(c)}{\leq} \left(1 - 2\mu\eta_t + L\eta_t^2 + \eta_t^2\alpha^{2m}\sigma^2C_\chi^2 + 2\eta_t\alpha^m\sigma C_\chi\sqrt{1 - 2\mu\eta_t + L^2\eta_t^2} \right) \times \delta_{t,j-1}^{\text{SGDw}} \\
&+ 2 \left(\eta_t^2\alpha^m\sigma C_\chi \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{2\eta_t^2(\sqrt{\sigma^2 + \kappa_{v;m}^2} + \sigma)\alpha^m\sigma C_\chi}{\sqrt{B}} \right. \\
&\quad \left. + \eta_t\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon)\sqrt{1 - 2\mu\eta_t + L_2\eta_t^2} + \frac{2\eta_t\sigma\sqrt{1 - 2\mu\eta_t + L^2\eta_t^2}}{\sqrt{n}} \right) \times \sqrt{\delta_{t,j-1}^{\text{SGDw}}} \\
&+ \eta_t^2 \left(\frac{4(\sigma^2 + \kappa_{v;m}^2)}{B} + (\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon))^2 + \frac{5\kappa_{v;m}^2 + 9\sigma^2}{B} + \frac{4\sqrt{\sigma^2 + \kappa_{v;m}^2}}{\sqrt{B}} \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) \right. \\
&\quad \left. + \frac{8\sigma\sqrt{\sigma^2 + \kappa_{v;m}^2}}{B} + \frac{4\sigma\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon)}{\sqrt{B}} \right) \\
&=: (A_1) \times \delta_{t,j-1}^{\text{SGDw}} + (A_2) \times \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + (A_3).
\end{aligned}$$

In (a), we have expanded the square, used \mathcal{F}_n defined in Lemma IV.E.1 and \mathcal{A}_n defined in Lemma IV.E.2, and noted that $\theta^{\text{pop}} - \psi^*$ is almost surely constant given θ^{init} ; in (b), we have applied Lemmas IV.E.1, IV.E.2, IV.E.3 and IV.E.4 under A0, A1, A2, A3 and A6, and used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$; in (c), we have grouped the terms by powers of $\delta_{t,j-1}^{\text{SGDw}}$, bounded the indicator function from above by 1 and used $B \leq n$ to replace n in the denominator by B . While the computation above is complicated, we remark that the key steps are taking the conditional expectations for the cross-terms in (b), which gives us a tighter bound than directly applying a triangle inequality of the form $\mathbb{E}\|Y_1 + Y_2\|^2 \leq (\sqrt{\mathbb{E}\|Y_1\|^2} + \sqrt{\mathbb{E}\|Y_2\|^2})$. To further simplify the bounds, we seek to bound each coefficient by a square, which yields

$$\begin{aligned}
(A_1) &= \left(\eta_t\alpha^m\sigma C_\chi + \sqrt{1 - 2\mu\eta_t + L\eta_t^2} \right)^2, \\
(A_2) &\leq 2 \left(\eta_t\alpha^m\sigma C_\chi + \sqrt{1 - 2\mu\eta_t + L\eta_t^2} \right) \times \eta_t \left(\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{4\sigma + 2\kappa_{v;m}}{\sqrt{B}} \right), \\
(A_3) &\leq \eta_t^2 \left((\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon))^2 + \frac{8\sigma^2 + 4\kappa_{v;m}^2}{\sqrt{B}} \varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{21\sigma^2 + 17\kappa_{v;m}^2}{B} \right) \\
&\leq \eta_t^2 \left(\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v;m}}{\sqrt{B}} \right)^2.
\end{aligned}$$

This implies

$$\begin{aligned}
&\delta_{t,j}^{\text{SGDw}} \\
&\leq (A_1) \times \delta_{t,j-1}^{\text{SGDw}} + (A_2) \times \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + (A_3)
\end{aligned}$$

$$\leq \left(\left(\eta_t \alpha^m \sigma C_\chi + \sqrt{1 - 2\mu \eta_t + L \eta_t^2} \right) \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \eta_t \left(\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v;m}}{\sqrt{B}} \right) \right)^2.$$

Now note that since $\mu \leq L$ by the definitions in A0, $2\mu \eta_t - L^2 \eta_t^2 \leq 2L\eta_t - L^2 \eta_t^2 \leq 1$.

By using $\sqrt{1-x} \leq 1 - \frac{x}{2}$ for all $x \leq 1$, we get that

$$\sqrt{1 - 2\mu \eta_t + L^2 \eta_t^2} \leq 1 - \mu \eta_t + \frac{L^2}{2} \eta_t^2.$$

Substituting this into the earlier bound and taking a square-root, we obtain the desired bound that

$$\begin{aligned} & \sqrt{\delta_{t,j}^{\text{SGDw}}} \\ & \leq \left(1 - \eta_t \left(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2} \eta_t \right) \right) \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + \eta_t \left(\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v;m}}{\sqrt{B}} \right). \end{aligned}$$

Moreover, since $L \geq \mu$ by definition from A0, we have

$$1 - \eta_t \left(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2} \eta_t \right) = \left(\frac{L}{\sqrt{2}} \eta_t - 1 \right)^2 + (\sqrt{2}L - \mu) \eta_t + \alpha^m \sigma C_\chi \eta_t > 0,$$

which finishes the proof. \square

IV.F.1 Proof of Theorem IV.B.1

Under A0, A1, A2, A3 and A6, Lemma IV.F.1 implies

$$\sqrt{\delta_{t,j}^{\text{SGDw}}} \leq \left(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} \right) \sqrt{\delta_{t,j-1}^{\text{SGDw}}} + C t^{-\beta} \sigma_{n,T}^{\text{SGDw}}$$

for $1 \leq t \leq T$ and $1 \leq j \leq N$, where we have used $\tilde{\mu}_m = \mu - \alpha^m \sigma C_\chi$, $\eta_t = C t^{-\beta}$ and

$$\varepsilon_{n,m,t;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v;m}}{\sqrt{B}} \leq \varepsilon_{n,m,T;v}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{v;m}}{\sqrt{B}} = \sigma_{n,T}^{\text{SGDw}}.$$

By an induction on $j = 1, \dots, N$, we have

$$\begin{aligned} \sqrt{\delta_{t,N}^{\text{SGDw}}} & \leq \left(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} \right)^N \sqrt{\delta_{t,0}^{\text{SGDw}}} \\ & + C \sigma_{n,T}^{\text{SGDw}} t^{-\beta} \sum_{j=1}^N \left(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} \right)^{j-1}. \end{aligned}$$

By noting that $\delta_{t,0}^{\text{SGDw}} = \delta_{t-1,N}^{\text{SGDw}}$ almost surely for $t \geq 2$ and using another induction on $t = 1, \dots, T$,

$$\sqrt{\delta_{T,N}^{\text{SGDw}}} \leq Q_0 \sqrt{\delta_{0,0}^{\text{SGDw}}} + C \sigma_{n,T}^{\text{SGDw}} \sum_{t=1}^T Q_t A_t t^{-\beta}, \quad (\text{IV.22})$$

where

$$Q_t := \prod_{s=t+1}^T \left(1 - \tilde{\mu}_m C s^{-\beta} + \frac{L^2 C^2}{2} s^{-2\beta} \right)^N, A_t := \sum_{j=1}^N \left(1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} \right)^{j-1}.$$

Note that by Lemma IV.F.1, we also have that $1 - \tilde{\mu}_m C t^{-\beta} + \frac{L^2 C^2}{2} t^{-2\beta} > 0$. This allows us to apply Lemma IV.C.3 to obtain

$$\kappa_0 \leq \exp \left(1 - N \tilde{\mu}_m C \varphi_{1-\beta}(T+1) + \frac{NL^2C^2}{2} \varphi_{1-2\beta}(T+1) \right) = E_1^{T,N}.$$

To control $\sum_{t=1}^T Q_t A_t t^{-\beta}$, recall that $\beta \in [0, 1]$, $\tilde{\mu}_m > 0$, $\frac{L^2 C^2}{2} > 0$, and that

$$E_2^{T,N} = \exp \left(- \frac{N \tilde{\mu}_m C}{2} \varphi_{1-\beta}(T+1) + 2NL^2C^2 \varphi_{1-2\beta}(T+1) \right).$$

We can now apply Lemma IV.C.3: If $\beta \notin \{\frac{1}{2}, 1\}$, then

$$\sum_{t=1}^T Q_t A_t t^{-\beta} \leq \frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)} \frac{N}{(T+1)^\beta}} + \frac{3^\beta (1 + \tilde{\mu}_m C)^{N-1} (T+2)^\beta}{L^2 C^2} E_2^{T,N}.$$

If $\beta = \frac{1}{2}$, i.e. $2\beta = 1$, we have

$$\sum_{t=1}^T Q_t A_t t^{-\beta} \leq \frac{4}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C N}{(T+1)^{1/2}}} + 2N(1 + \tilde{\mu}_m C)^{N-1} \varphi_{\frac{1}{2}-L^2 C^2 N}(T+1) E_2^{T,N}.$$

If $\beta = 1$, we get that

$$\sum_{t=1}^T Q_t A_t t^{-\beta} \leq \frac{4}{\tilde{\mu}_m C} + \frac{3N(1 + \frac{L^2 C^2}{2})^{N-1} e^{2L^2 C^2 N} \log(T+1)}{(T+1)^{(\tilde{\mu}_m C N)/2}}$$

Substituting the bounds into (IV.22), we get the desired bound that $\sqrt{\delta_{T,N}^{\text{SGDw}}}$ is upper bounded by

$$\left\{ \begin{array}{l} E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDw}}} + C\sigma_{n,T}^{\text{SGDw}} \left(\frac{4e^{\frac{\tilde{\mu}_m CN}{(T+1)^{1/2}}}}{\tilde{\mu}_m C} + 2N(1+\tilde{\mu}_m C)^{N-1} \varphi_{\frac{1}{2}-L^2C^2N}(T+1) E_2^{T,N} \right) \\ \quad \text{for } \beta = \frac{1}{2}, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDw}}} + C\sigma_{n,T}^{\text{SGDw}} \left(\frac{4}{\tilde{\mu}_m C} + \frac{3N(1+\frac{L^2C^2}{2})^{N-1} e^{2L^2C^2N} \log(T+1)}{(T+1)^{(\tilde{\mu}_m CN)/2}} \right) \text{ for } \beta = 1, \\ E_1^{T,N} \sqrt{\delta_{0,0}^{\text{SGDw}}} + C\sigma_{n,T}^{\text{SGDw}} \left(\frac{2^{2\beta+1}}{\tilde{\mu}_m C} e^{\frac{\tilde{\mu}_m C}{2(1-\beta)} \frac{N}{(T+1)^\beta}} + \frac{3^\beta (1+\tilde{\mu}_m C)^{N-1} (T+2)^\beta}{L^2C^2} E_2^{T,N} \right) \\ \quad \text{otherwise.} \end{array} \right.$$

□

IV.F.2 Proof of Theorem IV.B.2

Recall that $\delta_{t,j}^{\text{SGD}\text{Do}} := \mathbb{E} \|\psi_{t,j}^{\text{SGD}\text{Do}} - \psi^*\|^2$. To apply the results from Section IV.E to $\psi_{t,j}^{\text{SGD}\text{Do}}$, we condition on S_t^o , which in particular fixes $S_{t,j}^o$, the last size- B subset of $[n]$ chosen. We then identify $\theta^{\text{init}} = \psi_{t,j-1}^{\text{SGD}\text{Do}}$, $\eta = \eta_t$ and the dataset used as $\mathcal{D}_{t,j}^o := (X_i : i \in S_{t,j}^o)$, which allows us to identify $\psi_{t,j}^{\text{SGD}\text{Do}} = \theta_m^{\text{GD}}$, the full-batch gradient descent update using $\mathcal{D}_{t,j}^o$. Meanwhile, we note that almost surely $\theta_m^{\text{GD}} = \theta_{m,B}^{\text{SGDw}}$, the SGD-with-replacement iterate that uses the full dataset $\mathcal{D}_{t,j}^o$. Observe that the proof of Lemma IV.F.1 holds with $\delta_{t,j-1}^{\text{SGD}\text{Do}}$ replaced by any random initialization θ^{init} possibly correlated with X_1, \dots, X_n , which allows us to obtain

$$\begin{aligned} & \sqrt{\delta_{t,j}^{\text{SGD}\text{Do}}} \\ & \leq \left(1 - \eta_t \left(\mu - \alpha^m \sigma C_\chi - \frac{L^2}{2} \eta_t \right) \right) \sqrt{\delta_{t,j-1}^{\text{SGD}\text{Do}}} + \eta_t \left(\varepsilon_{n,m,t;\nu}^{\text{SGDw}}(\varepsilon) + \frac{5\sigma + 5\kappa_{\nu;m}}{\sqrt{B}} \right). \end{aligned}$$

Since the error recursion for $\delta_{t,j}^{\text{SGD}\text{Do}}$ is identical to that of $\delta_{t,j}^{\text{SGDw}}$ in Lemma IV.F.1, the proof of Theorem IV.4.3 follows directly, thereby yielding an identical result for $\delta_{T,N}^{\text{SGD}\text{Do}}$ as with $\delta_{T,N}^{\text{SGDw}}$ in Theorem IV.4.3. □

IV.G Proofs for tail probability bounds in offline SGD

We present the proofs for results in Section IV.B.3 that control the tail probability terms $\vartheta_{v;n,m,T}^{\text{SGDw}}$ and $\varepsilon_{v;n,m,T}^{\text{SGDw}}$.

Proof of Lemma IV.B.3. For $\delta > 0$, let N_δ be the δ -covering number of Ψ , which satisfies $N_\delta \leq (r_\Psi(1 + 2/\delta))^p$ (Example 5.8, [262]). Note also that by the Jensen's inequality applied to $\mathbb{E}[\cdot | X_1]$ and Assumption A5, there exist some $\sigma_m, \zeta_m > 0$ such that, for any $z \in \mathbb{R}^p$ with $\|z\| \leq \zeta_m$,

$$\mathbb{E}[e^{z^\top (\mathbb{E}[\phi(K_{\psi^*}^m(X_1)) | X_1] - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \leq \mathbb{E}[e^{z^\top (\phi(K_{\psi^*}^m(X_1)) - \mathbb{E}[\phi(K_{\psi^*}^m(X_1))])}] \leq e^{\sigma_m^2 \|z\|^2 / 2}.$$

Meanwhile, under A4, recycling the proof of Lemma 3.1 of [121] shows that if $C_m \delta / \sqrt{p} < \sigma_m^2 \zeta_m$,

$$\begin{aligned} \mathbb{P}\left(\sup_{\psi \in \Psi} \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\phi(K_\psi^m(X_i)) | X_i] - \mathbb{E}[\phi(K_\psi^m(X_i))]\right\| > 3C_m \delta\right) \\ \leq 2N_\delta p \exp\left(-\frac{nC_m^2 \delta^2}{2p\sigma_m^2}\right) \\ \leq 2(r_\Psi)^p \exp\left(p \log(1 + 2\delta^{-1}) - \frac{nC_m^2 \delta^2}{2p\sigma_m^2}\right). \end{aligned}$$

Since the probability above is an upper bound to $\vartheta_{n,m,T}^{\text{SGDw}}(3C_m \delta)$, we get that if $C_m \delta / \sqrt{p} < \sigma_m^2 \zeta_m$,

$$\begin{aligned} \left(\varepsilon_{v;n,m,T}^{\text{SGDw}}(3C_m \delta)\right)^2 &= 9C_m^2 \delta^2 + \kappa_{v;m}^2 \left(\vartheta_{n,m,T}^{\text{SGDw}}\left(\frac{C_m \delta}{\sqrt{p}}\right)\right)^{\frac{v-2}{v}} \\ &\leq 9C_m^2 \delta^2 + \kappa_{v;m}^2 2^{\frac{2(v-2)}{v}} (r_\Psi)^{\frac{(v-2)p}{v}} \exp\left(\frac{(v-2)p}{v} \log(1 + 2\delta^{-1}) - \frac{n(v-2)C_m^2 \delta^2}{2vp\sigma_m^2}\right). \end{aligned}$$

Recall that by assumption, $\frac{\log n}{n} < \frac{\sigma_m^2 \zeta_m^2}{p+v-2}$. We now choose

$$\delta = \frac{\sigma_m}{C_m} \sqrt{p \left(p + \frac{2v}{v-2}\right) \times \frac{\log n}{n}} = \frac{\sigma_m}{C_m} \sqrt{p \times \frac{(v-2)p+2v}{v-2} \times \frac{\log n}{n}},$$

which implies

$$\inf_{\varepsilon > 0} \left(\varepsilon_{v;n,m,T}^{\text{SGDw}}(\varepsilon)\right)^2 \leq \left(\varepsilon_{v;n,m,T}^{\text{SGDw}}(3C_m \delta)\right)^2$$

$$\begin{aligned}
&\leq \frac{9\sigma_m^2 p((v-2)p+2v)\log n}{(v-2)n} + \kappa_{v;m}^2 2^{\frac{v-2}{v}} (r_\Psi)^{\frac{(v-2)p}{v}} (1+2\delta^{-1})^{\frac{(v-2)p}{v}} \\
&\quad \times \exp\left(-\frac{(v-2)p+2v}{2v}\log n\right) \\
&\leq \frac{9\sigma_m^2 p((v-2)p+2v)\log n}{(v-2)n} \\
&\quad + \kappa_{v;m}^2 2^{\frac{v-2}{v}} (r_\Psi)^{\frac{(v-2)p}{v}} \left(1 + \frac{2C_m(v-2)^{1/2}}{\sigma_m p^{1/2}((v-2)p+2v)^{1/2}} \frac{\sqrt{n}}{\sqrt{\log n}}\right)^{\frac{(v-2)p}{v}} n^{-\frac{(v-2)p+2v}{2v}} \\
&\leq \frac{9\sigma_m^2 p((v-2)p+2v)\log n}{(v-2)n} \\
&\quad + \kappa_{v;m}^2 2^{\frac{v-2}{v}} (r_\Psi)^{\frac{(v-2)p}{v}} \left(1 + \frac{2C_m(v-2)^{1/2}}{\sigma_m p^{1/2}((v-2)p+2v)^{1/2}}\right)^{\frac{(v-2)p}{v}} n^{\frac{(v-2)p}{2v} - \frac{(v-2)p+2v}{2v}} \\
&\leq \left(\frac{9\sigma_m^2 p((v-2)p+2v)}{v-2} + \kappa_{v;m}^2 2^{\frac{v-2}{v}} (r_\Psi)^{\frac{(v-2)p}{v}} \right. \\
&\quad \left. \left(1 + \frac{2C_m(v-2)^{1/2}}{\sigma_m p^{1/2}((v-2)p+2v)^{1/2}}\right)^{\frac{(v-2)p}{v}}\right) \frac{\log n}{n}.
\end{aligned}$$

Taking a square root across and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem IV.B.1. \square

Proof of Lemma IV.B.4. Let $\delta > 0$ and N_δ be defined as in the proof of Lemma IV.B.3, with $N_\delta \leq (r_\Psi(1+2/\delta))^p$. Let $(\psi_l)_{l=1}^{N_\delta}$ be the centers of the covering δ -balls. The covering-ball argument of the proof of Lemma 3.1 of [121] shows that

$$\begin{aligned}
&\vartheta_{n,m,T}^{\text{SGDw}}(3C_m\delta) \\
&\leq \mathbb{P}\left(\sup_{\psi \in \Psi} \left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\phi(K_\psi^m(X_i)) | X_i] - \mathbb{E}[\phi(K_\psi^m(X_i))] \right\| > 3C_m\delta\right) \\
&\leq \sum_{l=1}^{N_\delta} \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\phi(K_{\psi_l}^m(X_i)) | X_i] - \mathbb{E}[\phi(K_{\psi_l}^m(X_i))] \right\| \geq C_m\delta\right).
\end{aligned}$$

By a Markov's inequality followed by the Burkholder's inequality applied to an average of i.i.d. summands (see e.g. [56] for $v > 2$), there exists a constant $C_v > 0$ depending only on v such that

$$\begin{aligned}
\vartheta_{n,m,T}^{\text{SGDw}}(3C_m\delta) &\leq \sum_{l=1}^{N_\delta} \frac{\mathbb{E}\left\| \sum_{i=1}^n \mathbb{E}[\phi(K_{\psi_l}^m(X_i)) | X_i] - \mathbb{E}[\phi(K_{\psi_l}^m(X_i))] \right\|^v}{n^v C_m^v \delta^v} \\
&\leq \sum_{l=1}^{N_\delta} \frac{\mathbb{E}\left\| \mathbb{E}[\phi(K_{\psi_l}^m(X_1)) | X_1] - \mathbb{E}[\phi(K_{\psi_l}^m(X_1))] \right\|^v}{n^{v/2} C_m^v \delta^v}
\end{aligned}$$

$$\leq \frac{N_\delta \kappa_{v;m}^v}{n^{v/2} C_m^v \delta^v} \leq \frac{(r_\Psi)^p (1+2\delta^{-1})^p \kappa_{v;m}^v}{n^{v/2} C_m^v \delta^v},$$

where we have used assumption A3 in the last line. This implies

$$\begin{aligned} \left(\epsilon_{v;n,m,T}^{\text{SGDw}}(3C_m \delta) \right)^2 &= 9C_m^2 \delta^2 + \kappa_{v;m}^2 \left(\vartheta_{n,m,T}^{\text{SGDw}}(3C_m \delta) \right)^{(v-2)/v} \\ &\leq 9C_m^2 \delta^2 + \kappa_{v;m}^v (r_\Psi)^{\frac{(v-2)p}{v}} C_m^{-(v-2)} \times \frac{(1+2\delta^{-1})^{\frac{(v-2)p}{v}}}{n^{\frac{v-2}{2}} \delta^{v-2}}. \end{aligned}$$

Choosing $\delta = n^{-\frac{(v-2)v}{2(v^2+(v-2)p)}} \leq 1$, we get that

$$\begin{aligned} \inf_{\epsilon>0} \left(\epsilon_{v;n,m,T}^{\text{SGDw}}(\epsilon) \right)^2 &\leq \left(\epsilon_{v;n,m,T}^{\text{SGDw}}(3C_m \delta) \right)^2 \\ &\leq 9C_m^2 n^{-\frac{(v-2)v}{v^2+(v-2)p}} + \kappa_{v;m}^v (r_\Psi)^{\frac{(v-2)p}{v}} C_m^{-(v-2)} 3^{\frac{(v-2)p}{v}} n^{-\frac{(v-2)v}{v^2+(v-2)p}} \\ &= (9C_m^2 + \kappa_{v;m}^v (r_\Psi)^{\frac{(v-2)p}{v}} C_m^{-(v-2)} 3^{\frac{(v-2)p}{v}}) n^{-\frac{(v-2)v}{v^2+(v-2)p}}. \end{aligned}$$

Taking a square root across and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem IV.B.1. \square

Proof of Lemma IV.B.5. By a Markov's inequality and a Jensen's inequality with respect to the empirical average, we have

$$\begin{aligned} \vartheta_{n,m,T}^{\text{SGDw}}(\epsilon) &= \sup_{t \in [T], j \in [N]} \frac{\sum_{i=1}^n \mathbb{E} \left\| \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X_i) \right) \middle| X_i, \psi_{t-1,j}^{\text{SGDw}} \right] - \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X'_1) \right) \middle| \psi_{t-1,j}^{\text{SGDw}} \right] \right\|}{n\epsilon} \\ &=: \sup_{t \in [T], j \in [N]} \frac{\sum_{i=1}^n A_{tji}}{n\epsilon}. \end{aligned}$$

Meanwhile, by a triangle inequality and the ergodicity assumption, we have

$$\begin{aligned} A_{tji} &\leq \mathbb{E} \left\| \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X_i) \right) \middle| X_i, \psi_{t-1,j}^{\text{SGDw}} \right] - \mathbb{E} \left[\phi \left(X_1^{\psi_{t-1,j}^{\text{SGDw}}} \right) \middle| \psi_{t-1,j}^{\text{SGDw}} \right] \right\| \\ &\quad + \mathbb{E} \left\| \mathbb{E} \left[\phi \left(X_1^{\psi_{t-1,j}^{\text{SGDw}}} \right) \middle| \psi_{t-1,j}^{\text{SGDw}} \right] - \mathbb{E} \left[\phi \left(K_{\psi_{t-1,j}^{\text{SGDw}}}^m(X'_1) \right) \middle| \psi_{t-1,j}^{\text{SGDw}} \right] \right\| \end{aligned}$$

$$\leq 2\tilde{C}_K \tilde{\alpha}^m.$$

This implies $\vartheta_{n,m,T}^{\text{SGDw}}(\varepsilon) \leq 2\tilde{C}_K \tilde{\alpha}^m \varepsilon^{-1}$, and therefore

$$(\varepsilon_{v;n,m,T}^{\text{SGDw}}(\varepsilon))^2 \leq \varepsilon^2 + 2^{\frac{v-2}{v}} \kappa_{v;m}^2(\tilde{C}_K)^{\frac{v-2}{v}} \tilde{\alpha}^{\frac{(v-2)m}{v}} \varepsilon^{-\frac{v-2}{v}}.$$

Choosing $\varepsilon = \tilde{\alpha}^{(v-2)m/(3v-2)}$ gives

$$\inf_{\varepsilon > 0} (\varepsilon_{v;n,m,T}^{\text{SGDw}}(\varepsilon))^2 \leq (1 + 2^{\frac{v-2}{v}} \kappa_{v;m}^2(\tilde{C}_K)^{\frac{v-2}{v}}) \tilde{\alpha}^{\frac{2(v-2)m}{3v-2}}.$$

Taking a square root across and using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b > 0$ gives the desired bound. The limiting result follows by substituting this bound into Theorem IV.B.1. \square

CHAPTER V

Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference

This Chapter is based on the following work:

Pierre Glaser, Michael Arbel, Arnaud Doucet, and Arthur Gretton. Maximum likelihood learning of energy-based models for simulation-based inference. A previous version of this work is available at <https://arxiv.org/abs/2210.147562025>, 2025

Models are the linchpin of modern-day science [70]. They describe in a symbolic manner the rules that dictate the evolution of objects of interest. While often incomplete and inexact, models are able to explain observed data in a parsimonious manner, improving our understanding of scientific phenomena and their interactions [209, 193]. Scientific models often include in their description parameters whose values are unknown; such “free parameters” are typically set in order to maximize the match between “real” data collected from the physical system of interest, and “synthetic” data obtained from the model of that system. In the common case when the relationship between the parameter and the observed data is non-deterministic,

this task corresponds to statistical inference. Inference is a cornerstone procedure of the scientific process. In its common form however, it requires a description of the models in terms of a probability density function of the data given the parameters, also known as the likelihood. For many models, several sources of intractability prevent access to this likelihood, making traditional inference methods inapplicable. Fortunately, for a significant part of models with intractable likelihoods, it is possible to draw, or *simulate* data following the rules of the model given a set of values for the parameters. *Simulation-Based Inference* (SBI) methods [47] are specifically designed to perform inference in the setting of a simulator with an intractable likelihood. When a prior on the free parameters is specified, these methods repeatedly generate synthetic data using the simulator to build an estimate of the posterior, that either can be used for any observed data (resulting in a so-called *amortized* inference procedure) or one that is *targeted* for a specific observation. While the accuracy of inference increases as more simulations are run, so does computational cost, especially when the simulator is expensive, which is common in many applications, such as particle physics [47] or neuroscience [82]. It is thus crucial to design methods with high simulation efficiency, i.e. that require as few simulations as possible to reach a given level of inference accuracy. Early SBI variants, known as Approximate Bayesian Computation (ABC) [18], which use sampling algorithms on a plug-in estimator of the posterior distribution in order to perform inference, typically exhibit poor simulation efficiency in high dimensions, due to the failure of the plug-in estimator to capture accurately the true posterior distribution. More recently, a new class of methods was introduced [272, 190, 101, 87], which departs from the ABC paradigm by sampling from a *learned* model of the posterior using a neural network fitted using the simulated data. Multiple variants of these methods were proposed, and often distinguished based on how tractably inference can be performed. The sample efficiency—e.g. the posterior accuracy as a function of the number of simulation—of these methods was empirically studied in [160], and shown to be drastically better than ABC. The gains are particularly important when comparing ABC to *targeted* SBI methods, implemented in a multi-round procedure

that refines the model around the observed data, by sequentially simulating data points that are closer to the observed ones [87, 190, 101]. However, while highly variable, the difference in performance across the various neural SBI methods was not explained in [160], leaving little guidance on which one to use in practice, or whether the efficiency of these methods could be improved. For expensive simulators, days of computation are still necessary to obtain accurate posterior estimates [78].

Contributions. In this work, we investigate the main SBI methods from the point of view of sample efficiency, a crucial factor when the simulator is expensive. We discuss how these methods may contain efficiency bottlenecks, which manifest themselves in practice. With that in mind, we introduce a new SBI Method, termed *Unnormalized Likelihood Neural Estimation* (UNLE), which we design to address these bottlenecks. UNLE learns a Conditional Energy Based Model (CEBM) of the simulator by Maximum Likelihood and performs MCMC on the posterior estimate obtained after invoking Bayes’ Rule. In doing so, we introduce a new method to train CEBMs that extends the contrastive divergence algorithm, an approximate maximum likelihood training method for standard EBMs, and prove its consistency. Additionally, we introduce a variational inference method that can improve the tractability of UNLE’s posterior. We demonstrate the properties of UNLE on an array of synthetic benchmark models [160], and apply UNLE to a neuroscience model of the crab *Cancer borealis*, where we improve the posterior accuracy over prior state-of-the-art while needing only a fraction of the simulations required by the most efficient previous method [78].

V.1 Neural SBI and its simulation efficiency

V.1.1 Background

Given a physical system from which recordings $x_o \in \mathcal{X}$ may be obtained, and a model of that system with free parameters $\theta \in \Theta$ taking the form of a likelihood $p(x \in \cdot | \theta = \diamond)$, Simulation-Based Inference (SBI) refers to the task of inferring which values of the free parameters θ were most susceptible to have generated a given recording x_o when the model’s likelihood $p(x \in \cdot | \theta = \diamond)$ is not available, but

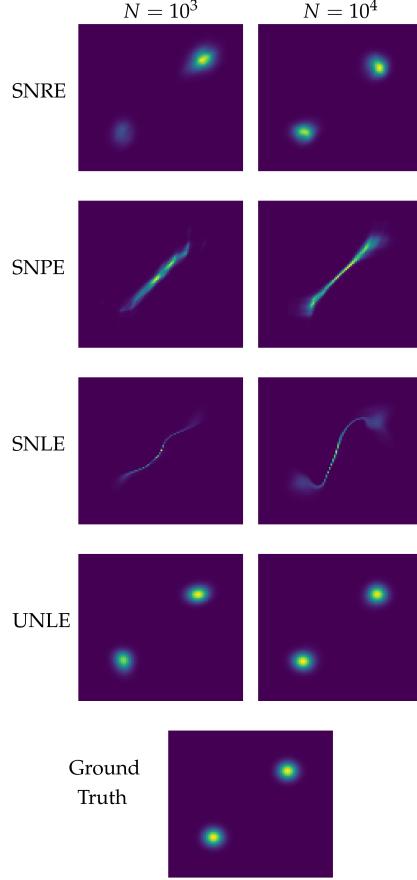


Figure V.0.1: Performance of NRE, NPE, NLE and UNLE, all trained using a simulator with a bimodal likelihood $p(x|\theta)$ and a Gaussian prior $p(\theta)$, using 1000 (left) and 10000 (right) simulations.

draws from $p(x \in \cdot | \theta)$ can be obtained for any $\theta \in \Theta$. When a prior distribution $p(\theta \in \cdot)$ is assumed on θ , SBI methods reduce to obtaining a tractable approximation of posterior

$$p(\theta \in \cdot | x) \propto p(\theta \in \cdot) \overbrace{\frac{dp(x \in \cdot | \theta)}{dp(x \in \cdot)}}^{\text{unknown}}(x) \quad (\text{V.1})$$

of the joint system $p((x, \theta) \in \cdot) := p(\theta \in \cdot)p(x \in \cdot | \theta)$ formed by the prior and the model. Although $p(x \in \cdot | \theta)$ is unknown, making standard Bayesian inference inapplicable, SBI methods have proven that that *approximate* Bayesian inference can be performed using independent parameter-simulation pairs $\mathcal{D}_n := \{\theta_i, X_i\}_{i=1}^n$ sampled from $p_\pi((x, \theta) \in \cdot) := \pi(\theta \in \cdot)p(x \in \cdot | \theta)$, which we call the simulated distribution. Here, π is a *proposal* distribution which SBI methods often do not require to be the prior for reasons that will become clear later. To perform this task, SBI methods

first learn a neural approximation $q_\psi(\theta \in \cdot | x = \diamond)$ of the posterior density (with neural network weights ψ) from such pairs. Practical inference operations (such as visualization, or computing expectations under q_ψ) can then be performed by sampling from q_ψ - directly if the q_ψ allows it, or using MCMC if the approximation is unnormalized. In this paper, we focus on the 3 main SBI variants benchmarked in [160], which have been successfully used in domain sciences: Neural Likelihood Estimation (NLE), Neural Ratio Estimation (NPE), and Neural Posterior Estimation (NPE).

Neural Likelihood Estimation [190] learns an approximation of the likelihood in the form of a (conditional) normalizing flow trained by maximizing the objective $\mathcal{L}_{\text{NLE}}(\psi) := \mathbb{E}_{p_\pi((x, \theta) \in \cdot)} \log \frac{dq_\psi}{d\mu}(x | \theta)$ and then applies Bayes' rule to obtain its learned unnormalized posterior estimate $q_\psi(\theta \in \cdot | x = \diamond) \propto p(\theta \in \cdot) \frac{dq_\psi}{d\mu}(x | \theta)$, which can be approximately sampled from using MCMC.

Neural Ratio Estimation [101] fits a model $r_\psi(x, \theta) := e^{-E_\psi(x, \theta)}$ of the likelihood-to-marginal ratio $\frac{dp(x \in \cdot | \theta)}{dp_\pi(x \in \cdot)}(x)$ of the simulated distribution p_π , by learning to distinguish between pairs (θ, x) from the simulated distribution $p_\pi((x, \theta) \in \cdot)$ and pairs $\{X_i, \theta_j\}_{1 \leq i \neq j \leq n}$ from the product of marginals $p_\pi(x \in \cdot) \pi(\theta \in \cdot)$. NRE then combines this ratio with the prior to obtain its posterior estimate $q_\psi(\theta \in \cdot | x = \diamond) \propto p(\theta \in \cdot) r_\psi(x, \theta)$, which is normalized only if $\pi(\theta \in \cdot)$ is the prior.

Neural Posterior Estimation [87] solves a modified, multi-class version of the binary classification problem of NRE, in order to directly learn a model $r_\psi(x, \theta)$ of $\frac{dp(x \in \cdot | \theta)}{dp(x \in \cdot)}$ instead of $\frac{dp(x \in \cdot | \theta)}{dp_\pi(x \in \cdot)}$. Using the identity $\frac{dp(x \in \cdot | \theta)}{dp(x \in \cdot)} = \frac{dp(\theta \in \cdot | x)}{dp(\theta \in \cdot)}$, NPE uses a ratio estimator of the form $r_\psi(x, \theta) := \frac{dq_\psi(\theta \in \cdot | x)}{dp(\theta \in \cdot)}$ where $q_\psi(\theta \in \cdot | x = \diamond)$ a conditional normalizing flow, and sets its learned q_ψ to be its posterior estimate. NPE thus produces a flow-based posterior estimate q_ψ that can be tractably sampled from regardless of the proposal π used.

V.1.2 The simulation efficiency of neural SBI methods

As discussed previously, an empirical comparison of the simulation efficiency of the main neural SBI methods already exist [160]. This work revealed a significant variability in the simulation efficiency of these methods depending on the simulator at hand, but did not provide explanations as of why. In the following section, we explain this variability by analyzing the *optimization* error and the *statistical* error of these methods, as well as how they link to two critical hyperparameters of an SBI method: the objective function used to fit the model, and the parametric family (or *model class*) in which this model is fit.

V.1.2.1 Sources of (in)accuracy in statistical estimation

All SBI methods discussed above are fit through *Empirical Risk Minimization* (ERM): Given the training data \mathcal{D}_n , they proceed by producing $\hat{\psi}_n$ an approximate minimizer of a random estimate (the “empirical” risk) $\mathcal{L}_n(\mathcal{D}_n, \psi)$ with minimizer $\bar{\psi}_n$ of a loss function (the “population risk”) $\mathcal{L}(p_\pi, \psi)$ which—assuming there exists some ψ^* such that $q_{\psi^*}(\theta \in \cdot | x = \diamond) = p(\theta \in \cdot | x)$ —is uniquely minimized at ψ^* . In that context, one should strive for accurate SBI methods, e.g. methods such that $\hat{\psi}_n$ is close (in, e.g. mean-squared error) to ψ^* . A mature and complete statistical analysis would involve quantifying this mean-squared error as a function of n and the other problem parameters. However, this is a difficult task, for only partial results exist in the literature. This quantity can however be analyzed through two important and independent proxies: its (asymptotic) statistical error, and its optimization error.

The statistical error of SBI methods Below, we propose to quantify the statistical error of SBI by a key quantity in the field of statistical estimation: the asymptotic variance \mathcal{V} its empirical minimizer $\bar{\psi}_n$, defined as:

$$\mathcal{V} := \lim_{n \rightarrow \infty} n \times \text{Cov} [\bar{\psi}_n - \psi^*].$$

By targeting the $\bar{\psi}_n$ (which is unknown in practice) instead of the approximate minimizer $\hat{\psi}_n$, this quantity isolates the “statistical” performance of the method, by not taking into account the error caused by using an inexact optimization procedure.

Such a quantity has been studied many times to analyze the statistical error of modern machine learning methods [92, 136, 151, 37, 244], and was empirically shown to be strongly correlated with the actual mean-squared error of $\hat{\psi}_n$. The impact of $\mathcal{L}(p_\pi, \psi)$ and $\mathcal{L}_n(\mathcal{D}_n, \psi)$ on this quantity can be understood by studying the case where $\mathcal{L}_n(\mathcal{D}_n, \psi) = \sum_{i=1}^n \ell((X_i, \theta_i), \psi)$ for some ℓ , as in NLE. In that case, $\mathcal{V} = H^{-1}GH^{-1}$, where $H := \mathbf{H}_\psi \mathcal{L}(p_\pi, \psi^*)$ and $G := \text{Cov} \nabla_\psi \ell((X_i, \theta_i), \psi^*)$. This quantity cannot be lower than a quantity called the Fisher Information Matrix of the model, the best possible asymptotic variance for any consistent estimator of the model [256]. Importantly, this lower bound is matched by Maximum Likelihood Estimators (MLE). *Thus, from a statistical perspective, using non-MLE estimators can yield to a significantly suboptimal accuracy.*

The optimization error of SBI methods The optimization error, defined as $\mathbb{E} \|\hat{\psi}_n - \bar{\psi}_n\|^2$, represents the error due to minimizing L_N approximately, and not exactly. The optimization error depends on the optimization algorithm used (typically variants of gradient descent). Its performance usually strongly depends on the *smoothness* of the problem, and, as we will see, can itself strongly depend on the model class used.

V.1.2.2 Impact of contrastive losses on the statistical error of SBI methods

NRE and NPE are not maximum likelihood estimators: instead of maximizing their log-likelihood, they use a contrastive loss function $\mathcal{L}_n(\mathcal{D}_n, \psi)$, e.g. minimize a classification loss between samples from the joint distribution and samples from the product of marginals. The asymptotic variance of estimators using contrastive losses was analyzed in 92, 151, 244. In particular, 151, 244 provides instances where this variance grows *exponentially* with dimension. The reason behind it is that the asymptotic variance of such contrastive methods depends on the ratio between the joint distribution and the noise distribution. Loosely speaking, this variance will be large, even possibly for few dimensions, when classifying between the joint samples and the marginal samples can be done accurately without having to learn very well the ratio $\frac{dp(x \in \cdot | \theta)}{dp_\pi(x \in \cdot)}$. This situation happens when the product of marginal distribution assigns low probability to high-density regions of the joint distribution, a case which

is likely to happen in practice. Moreover, while this results formally only holds in the case of NRE, we expect that NPE, which also relies on contrastive estimation, will exhibit similar issues. NPE offers a particularly interesting test bed for efficiency comparison, since it can use Maximum Likelihood in lieu of contrastive estimation otherwise. We confirm the possible deficits in simulation efficiency of contrastive methods by plotting the posterior obtained by running the NRE, as well as both the contrastive variant and the likelihood-based variant of NPE on the two moons model of [160], and which are plotted in Figure V.1.1.

Figure V.1.1: Amortized Posterior estimate of NRE and NPE trained on the two moons model for various number of samples. The middle column represents the posterior of NPE obtained by maximizing its contrastive objective, while the right column represents the posterior of NRE obtained by maximizing its likelihood-based objective, which converges faster to the true posterior.

V.1.2.3 Impact of normalizing flows on the optimization error of SBI methods

Above, we have seen that NRE and NPE could suffer from a large statistical error, due to their contrastive loss. NLE on the other hand uses a Maximum Likelihood loss, and does not suffer from this issue; indeed, as mentioned above, Maximum Likelihood Estimators have the smallest possible asymptotic variance among all consistent estimators of the model. On the other hand, NLE may have a large optimization error: this comes from the fact that NLE (as well as NPE) uses conditional versions of *Normalizing Flows* (NF) to model the likelihood. NFs [207] are models that transform a simple tractable distribution μ using a (parameterized) invertible transformation g_ψ , yielding a model $q_\psi(x \in \cdot | \theta = \diamond)$ of the form $(g_\psi)_\# \mu$. The conditional variants of NFs used by NLE and NPE allow the pushforward map g_ψ to depend on the conditioned variable θ [189]. On the one hand, can be tractably sampled from, and have tractable densities, the latter being useful for MLE-based training. On the other hand, such pushforward models are prone to poor conditioning, shown in the following lemma:

Lemma V.1.1. *Let $\mathcal{X} = \mathbb{R}^d$, and Θ be an arbitrary measurable set. Assume that there exists a set B with $\pi(\theta \in \cdot)(B) > 0$, such that, for all $\theta \in B$, the conditional probability of X given θ admits a regular version of the form $\lambda \mathcal{N}(m_1, \sigma^2 I) + (1 - \lambda) \mathcal{N}(m_2, \sigma^2 I)$, where $\lambda \in (0, 1)$, $m_1, m_2 \in \mathcal{X}$, $\sigma > 0$, I is the identity matrix, and $\mathcal{N}(m, \Sigma)$ denotes the multivariate normal distribution with mean m and covariance matrix Σ . Then there exists no regular version q of the conditional probability of $p(x \in \cdot | \theta = \diamond)$ of the form $p(x \in \cdot | \theta = \diamond) = (g_\psi)_\# \mu$ for some invertible Lipschitz function $g : \mathcal{X} \times \Theta \rightarrow \Theta$ such that*

$$\sup_{\theta \in B} \text{Lip}_x(g_\psi(\cdot, \theta)) < \sigma \exp \left[\|m_1 - m_2\|^2 / (8\sigma^2) - \Phi^{-1}(\lambda)^2 / 2 \right]$$

where $\varphi(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{x^2}{2}}$ and $\Phi(u) = \int_{-\infty}^u \varphi(x) dx$

The proof, given in Appendix V.C, is a natural extension of arguments developed to prove similar results in the unconditional case [215]. Lemma V.1.1 shows that when the likelihood is of the form of a mixture of Gaussians in some region B of the parameter space (which can be arbitrarily small, so long as it has positive measure), any pushforward model $(g_\psi(\cdot, \theta))_\# \mu$ will have to be ill-conditioned in the sense that its Lipschitz constant will blow up exponentially with the distance between the means. For NFs whose parameter norm grows polynomially with the Lipschitz constant of the pushforward map g_ψ , this implies that the empirical loss minimizer will be located in exponentially far regions of the parameter space, thereby complicating the optimization procedure, and increasing the optimization error.

V.2 Unnormalized Neural Likelihood Estimation

In this section, we introduce a novel method for performing Simulation-Based Inference, which we call Unnormalized Neural Likelihood Estimation (UNLE), and which we design to limit the two sources of inaccuracy of existing SBI methods discussed in the previous section. UNLE fits a flexible, conditional Energy-Based Model [147, 228]

$$q_\psi(x \in \cdot | \theta = \diamond) = \frac{e^{-E_\psi(x, \theta)}}{Z(\theta, \psi)} \mu(x \in \cdot),$$

$$Z(\theta, \psi) = \int e^{-E_\psi(x, \theta)} \mu(dx) \quad (\text{V.2})$$

to the unknown likelihood by approximately maximizing its empirical *average conditional log-likelihood*

$$\begin{aligned} \mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi) &:= \frac{1}{n} \sum_{i=1}^n \log \frac{dq_\psi}{d\mu}(X_i | \theta_i) \\ &= \frac{1}{n} \sum_{i=1}^n -E_\psi(X_i, \theta_i) - \log Z(\theta_i, \psi) \end{aligned} \quad (\text{V.3})$$

on the simulated dataset \mathcal{D}_n , and across a parameter set Ψ . Typically, E_ψ can be set to be a neural network with weights ψ , and $\mu(x \in \cdot)$ is a reference distribution, such as Gaussian measure, or even a possibly non-finite measure (such as the Lebesgue measure) if one wishes to learn the tail behavior of the likelihood within E_ψ . Finally, $Z(\theta, \psi)$ is an unknown normalizing function which, as we will see, does not need to be computed to perform either learning or inference. We propose to maximize 271 this loss approximately using an algorithm which adapts the well-known Contrastive Divergence algorithm [105, 121, 247], used to fit EBMs, to the conditional setting. Once this model is trained, UNLE invokes Bayes' rule to obtain its unnormalized posterior estimate:

$$q_\psi(\theta \in \cdot | x = \diamond) \propto p(\theta \in \cdot) \frac{dq_\psi}{d\mu}(x | \theta) = p(\theta \in \cdot) \frac{e^{-E_\psi(x, \theta)}}{Z(\theta, \psi)} \quad (\text{V.4})$$

This posterior, which contains an unknown term $Z(\theta, \psi)$ that depends on θ , can be approximately sampled from using known adaptations of a special class of MCMC algorithms called auxiliary variable MCMC [168, 175]. The combined use of a conditional EBM, which are density estimators that do not suffer from the poor conditioning issues of normalizing flows (at the cost of an intractable normalizing function) and of a maximum-likelihood objective, is intended to maximize its sample efficiency: the use of a maximum-likelihood objective ensures that the finite-sample posterior estimate has a low statistical error, while the use of a Conditional EBM

Table V.2.1: Comparison of the main SBI methods in their modeling choice and their objective function.

	NRE [101]	NPE [88]	NLE [190]	SMNLE [188]	UNLE (ours)
Direct Density Model	✓	✗	✗	✓	✓
Fisher-Efficient Estimator	✗	✗	✓	✗	✓

as its modelling components alleviates the risk of poor conditioning that comes with normalizing flows; we later show that this efficiency manifest itself in practice on both synthetic and real-world simulators. UNLE comes in two variants: one for amortized inference, and one for targeted (or *sequential* inference), the latter following the multi-round logic explained in Section V.1.1. Following standard terminological practices, we call them UNLE and SUNLE respectively, and detail their logic in Algorithm 3 and Algorithm 4. The function `ApproxSample` refers to any auxiliary variable MCMC algorithm that can sample from UNLE’s posterior.

Algorithm 3 UNLE

input: prior $p(\theta \in \cdot)$, simulator G , budget n , initial CEBM parameters ψ_0 , observation x_o , num. posterior samples p
sample $\{\theta_i \sim p(\theta \in \cdot)\}_{i=1}^n, \mathcal{D}_n \leftarrow \{X_i \sim G(\theta_i, \cdot)\}_{i=1}^n$
 $\hat{\psi}_n \leftarrow := \text{maximize_cebm_log_l}(\mathcal{D}_n, \psi_0)$
Set $q_\psi(\theta \in \cdot | x = \diamond) \propto p(\theta \in \cdot) \frac{dq_\psi}{d\mu}(x | \theta)$
 $\{\theta_i^{\text{post}}\}_{i=1}^p \leftarrow \text{ApproxSample}(q_\psi(\theta \in \cdot | x = \diamond), x_o)$
return $\{\theta_i^{\text{post}}\}_{i=1}^p, q_\psi(\theta \in \cdot | x = \diamond)$

Algorithm 4 Sequential-UNLE

Input: prior $p(\theta \in \cdot)$, simulator G , budget n , num. rounds R initial CEBM parameters ψ_0 , observation x_o
 $\pi^1 \leftarrow p, n_R \leftarrow \lceil n/R \rceil$
 $\{\theta_i^1 \sim \pi^1\}_{i=1}^{n_R}, \mathcal{D}_n \leftarrow \emptyset$

for $i = 1, \dots, n_R$ **do**
 $\hat{\psi}_n^i := \text{maximize_cebm_log_l}(\mathcal{D}_n, \hat{\psi}_n^{i-1})$
 $\pi^i \propto p(\theta \in \cdot) \frac{dq_\psi}{d\mu}(x | \theta)$
 n_R or $p \leftarrow p$ if $i = R$ else n_R
 $\{\theta_i^r\}_{i=1}^p \leftarrow \text{ApproxSample}(\pi^i, x_o)$
end for
return $\{\theta_i^R\}_{i=1}^p, q_\psi^R \propto p(\theta \in \cdot) \frac{dq_\psi}{d\mu}(x | \theta)$

We now detail the two steps of SUNLE: (i) the training of its CEBM likelihood model $q_{\psi^*}(x|\theta)$, and (ii) the sampling from its posterior.

V.2.1 Learning Conditional Energy-Based Models via maximum-likelihood

In optimizing $\mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi)$ when the model class is a CEBM, one is faced with a substantial challenge: the *gradient* of a CEBM’s log-likelihood contains the gradient of the log-normalizing function $\nabla_\psi \log Z(\theta, \psi)$

$$\nabla_\psi \log \frac{dq_\psi}{d\mu}(x|\theta) = -\nabla_\psi E_\psi(x, \theta) - \underbrace{\nabla_\psi \log Z(\theta, \psi)}_{\text{intractable}}, \quad (\text{V.5})$$

which is intractable. This intractability propagates into the gradient of $\mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi)$

$$\begin{aligned} \nabla_\psi \mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi) = & -\frac{1}{n} \sum_{i=1}^n \left(\nabla_\psi E_\psi(X_i, \theta_i) \right. \\ & \left. + \underbrace{\nabla_\psi \log Z(\theta_i, \psi)}_{\text{intractable}} \right), \quad (\text{V.6}) \end{aligned}$$

and consequently, it is not possible to use gradient-based methods in their simplest form to optimize it. Introspecting further the intractable terms however, it is possible to show that $\nabla_\psi \log Z(\theta, \psi)$ can be re-written in the form of an (intractable) expectation over the distribution $q_\psi(x \in \cdot | \theta = \diamond)$

$$\nabla_\psi \log Z(\theta, \psi) = \mathbb{E}_{x \sim q_\psi(x \in \cdot | \theta = \diamond)} \nabla_\psi E_\psi(x, \theta) \quad (\text{V.7})$$

This identity is well-known in the *unconditional* case (e.g. without the presence of a conditioning variable θ) and can be extended naturally to CEBMs. Motivated by this identity, we introduce an algorithm to maximize $\mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi)$.

This algorithm, described in Algorithm 5 proceeds at each epoch k and minibatch iteration i by approximating the B intractable terms present in the minibatch version of Equation (V.6) using, for each term, a particle approximation \hat{q}_j obtained by running an MCMC algorithm (represented by the procedure `MCMCAprox` in Algo-

rithm Algorithm 5) with target $q_\psi(x \in \cdot | \theta_j)$. The algorithm then combines these B approximations to obtain a minibatch gradient estimate for $\mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi)$ and produces the new parameter iterate $\psi_{k,i}$ by taking a gradient ascent step with step size $\gamma_{k,i}$

Theoretical Guarantees: Near-Optimal Accuracy Algorithm 5 designed to inherit the statistical efficiency of maximum-likelihood learning, is can be theoretically guaranteed, in certain cases, to achieve its purpose: in particular, we show below that assuming $p(x \in \cdot | \theta) = q_{\psi^*}(x \in \cdot | \theta)$ for some $\psi^* \in \Psi$, for conditional exponential families, a specific class of CEBMs (and with compact parameter space Ψ), using the online averaged variant of Algorithm 5 (e.g. Algorithm 5 with $B, K = 1$ and $B_{1,i} = \{i\}$ and averaging turned on) combined with a simple MCMC procedure MCMCAprox that returns, for each j , the last iterate of a single m -step MCMC chain with 341 target $q_{\psi^*}(x \in \cdot | \theta_j)$ and initial position X_j , and under mixing assumptions on the kernel and moment assumptions on the CEBM, the parameter $\hat{\psi}_h$ returned by the algorithm will have a near-optimal mean squared error to the true parameter ψ^* , differing from the Cramér-Rao lower bound by higher order terms, and a universal constant of 2. We refer to Section V.D.1 for a precise statement of the Theorem, and to Section V.D.2 for its full proof.

Algorithm 5 `maximize_cebm_log_1(D, ψ₀)`

Input: training data D_n , Initial EBM parameters $ψ₀$ learning rate schedule $γ_{k,i}$, batch size B , batching schedule $B_{k,i}$ number of epochs K average
 $\{\hat{q}_i \leftarrow δ_{X_i}\}_{i=1}^n, m \leftarrow \lceil n/B \rceil \quad ψ_{1,0} \leftarrow ψ₀$
for $k = 1, \dots, K - 1$ **do**
 for $i = 1, \dots, m$ **do**
 $\{\hat{q}_j \leftarrow \text{MCMCAprox}(q_{ψ_{k,i-1}}(x \in \cdot | θ_j))\}_{j \in B_{k,i}}$
 $G_{k,i} \leftarrow \frac{1}{B} \sum_{j \in B_{k,i}} \left(-\nabla_ψ E_{ψ_{k,i-1}}(X_j, θ_j) \right.$
 $\left. - \mathbb{E}_{\tilde{X}_j \sim \hat{q}_j} \nabla_ψ E_{ψ_{k,i-1}}(\tilde{X}_j, θ_j) \right)$
 $ψ_{k,i} \leftarrow ψ_{k,i-1} + γ_{k,i} G_{k,i}$
 end for
 end for
 $ψ_{k+1,0} \leftarrow ψ_{k,m}$
end for
if average **then**
 $\hat{ψ}_n \leftarrow \frac{1}{Km} \sum_{k=1}^K \sum_{i=1}^m ψ_{k,i}$
else
 $\hat{ψ}_n \leftarrow ψ_{K,m}$
end if

Theorem V.2.1 (informal). Consider CEBMs with energies of the form $E_ψ(x, θ) = ⟨ψ, φ(x, θ)⟩$ for some function $φ$, with compact parameter space $Ψ$, and satisfying mild moment and smoothness assumptions (see Assumptions A8, A9 in Section V.D.1 for details). Consider the MCMCAprox function, which, for any $q_ψ(x \in \cdot | θ_j)$, returns

$$\text{MCMCAprox}(q_ψ(x \in \cdot | θ_j)) := δ_{\tilde{X}_j}, \quad \tilde{X}_j \sim k_{ψ,θ_j}^m(\cdot, X_j)$$

where $k_{ψ,θ_j}^m(\cdot, X_j)$ is the distribution of an m -step MCMC kernel with invariant distribution $q_ψ(x \in \cdot | θ_j)$ and initial value X_j , and $k_{ψ,θ_j}^m$ satisfies certain mixing assumptions (see Assumption A10 in Section V.D.1). Consider the sequence of iterates returned by Algorithm 5 with any initialization $ψ₀ ∈ Ψ$, $B = 1, K = 1, B_{1,i} = \{i\}$, averaging turned on, a step-size schedule of the form $γ_{1,t} = Ct^{-γ}$ for some $C > 0$ and $γ ∈ (\frac{2}{3}, 1)$, the MCMCAprox function defined above with a number of MCMC steps $m(n)$ (defined in the full version of the theorem in Section V.D.1, and which depends on the mixing properties of the kernel $k_{ψ,θ}$). Then the parameter $\hat{ψ}_n$ returned by the algorithm satisfies

$$\mathbb{E} [\|\hat{\psi}_n - \psi^*\|^2] \leq 2 \frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n} + o(n^{-1}),$$

where $\mathcal{I}(\psi^*) := \mathbb{E}_{\theta \sim \pi} \text{Cov}_{X \sim q_{\psi}(x \in \cdot | \theta)} [\phi(X, \theta) | \theta]$ is the Fisher information matrix of the model at ψ^* . Consequently, we have that

$$\limsup_{n \rightarrow \infty} n \times \mathbb{E} [\|\hat{\psi}_n - \psi^*\|^2] \leq 2 \times \text{tr}(\mathcal{I}(\psi^*)^{-1})$$

Theorem V.2.1 provides bounds on the accuracy of the parameter returned by Algorithm 5 that hold for any number of samples n . This is an improvement compared to traditional analyses of statistical estimators that usually only provide guarantees in the limit $n \rightarrow \infty$, and is more in line with modern analyses such as the ones of [185, 229]. Its guarantees are strong: it ensures that the mean squared error of $\hat{\psi}_n$ is, up to higher order terms, at most twice the trace of the inverse of the Fisher information matrix $\mathcal{I}(\psi^*)$, the theoretical lower bound on the mean squared error for any unbiased estimator of ψ^* . From a technical standpoint, this result is achieved using 3 main features: First, using energies of the form $\langle \psi, \phi(x, \theta) \rangle$, which makes the objective function $\mathcal{L}_{\text{CLL},n}(\mathcal{D}_n, \psi)$ concave in ψ , allowing fast convergence and near optimal accuracy of averaged (unbiased) stochastic gradient ascent methods [169] with decreasing step sizes. Second, using training data points $X_j \sim p(x \in \cdot | \theta_j) = q_{\psi^*}(x \in \cdot | \theta_j)$ as the initialization of the MCMC algorithm used to approximate $q_{\psi_{1,t}}(x \in \cdot | \theta_j)$, ensures that the bias of the MCMC approximation decreases as the optimization proceeds and $\psi_{1,t}$ approaches ψ^* . Third, the use of enough MCMC steps $m(n)$ to ensure that the bias of the MCMC approximation can be controlled. The proof builds on the recently established results [247], which obtained similar guarantees in the unconditional case. One difference is that the mixing condition on the Markov Kernel of MCMCAprox, which was imposed to be uniform across all $\psi \in \Psi$ in [247], are not assumed to be uniform across $\theta \in \Theta$: in particular, our algorithm will be consistent even when the mixing time of MCMCAprox can get arbitrarily large for certain θ , as long as the proportion of such θ is small enough. Our results show that, depending on the distribution of a quantity θ , defined in Equation 27, and behaving analogously to the mixing times of the MCMC kernel

used at iteration t in MCMCAprox (e.g. the number of MCMC steps required to obtain a small bias in the approximation of $q_\psi(x \in \cdot | \theta)$, convergence will be obtained for a number MCMC steps either (i) logarithmic in n (if the mixing times distribution is bounded) (ii) almost logarithmic¹ (if the mixing times distribution has sub-Gaussian tails), (iii) polynomial (if the mixing times distribution has heavier sub-exponential tails), or (iv) exponential (if the mixing times distribution only has very heavy tails, with only a finite number of moment defined). The slightly more restricted allowed step size schedules $\beta \in (2/3, 1)$ compared to the unconditional case ($\beta \in (\frac{1}{2}, 1)$) comes from the need to handle projections onto the compact set Ψ , combined with the fact that certain moment conditions which automatically hold in the unconditional case (e.g. when θ is absent) do not hold anymore in the conditional case. The main limitation of this result is that it currently only holds for a specific form of energy functions, which are linear as a function of the parameter ψ . We believe that these results could be extended to more general CEBMs by leveraging more advanced techniques from the non-convex optimization literature [203].

V.2.2 Posterior Sampling

UNLE's posterior candidates q_ψ are posteriors obtained by applying Bayes rule using the prior p and its likelihood. Naive Metropolis-Hastings (MH) based MCMC with a proposal $q(\theta \in \cdot)$ cannot be performed, as computing the acceptance probability:

$$\alpha(\theta, \theta') = \min \left(1, \frac{e^{-E_\psi(x, \theta')} p(\theta') q(\theta) Z(\theta, \psi)}{e^{-E_\psi(x, \theta)} p(\theta) q(\theta') Z(\theta', \psi)} \right)$$

requires access to the intractable normalizing constant $Z(\theta, \psi)$ and $Z(\theta', \psi)$. Posterior distributions with intractable likelihoods, often called *doubly intractable* posteriors, are well-known in the Bayesian Inference literature, and a wide set of methods have been proposed to sample from them [191]. In this work, we use one class of such methods, called *auxiliary-variable* MCMC [168, 175]. Auxiliary Variable MCMC are MCMC algorithms that draw approximate samples from a doubly intractable posterior $q_\psi(\theta \in \cdot | x = \diamond)$ by performing “regular” MCMC sampling targeting an

¹Slower than any polynomial

augmented distribution $q_\psi((\theta, Y) \in \cdot | x)$ whose marginal in θ equals the posterior $q_\psi(\theta \in \cdot | x = \diamond)$: approximate posterior samples are obtained by selecting the θ component of the augmented samples returned by the MCMC algorithm while throwing away the auxiliary part. Impressively, and unlike alternative sampling methods for doubly intractable posteriors, auxiliary variable MCMC algorithms target the correct posterior distribution q_ψ . Such methods usually involve a sample from the likelihood $q_\psi(x \in \cdot | \theta)$ at every iteration to compute the acceptance probability of the proposed augmented sample. In cases when such samples cannot be obtained exactly, it is possible to replace them by approximate samples obtained by running an MCMC targeting the likelihood [188, 66, 5]. As UNLE’s likelihood $q_\psi(x \in \cdot | \theta)$ falls in this category, our sampling approach uses a similar strategy.

V.2.3 Handling Invalid Simulations

In problems where the dimension of the simulations is very high, SBI methods often learn posterior models $q_\psi(\theta \in \cdot | s(x))$ of parameters given lower-dimensional summary statistics $s(x)$ of the simulations, often curated by field experts. In such cases, simulations that fall outside the domain of these summary statistics have to be discarded. As discussed in [78], this procedure can bias the posterior estimate of synthetic likelihood methods. [78] proposed a method which debiases posteriors known up to a normalizing constant by its density with a correction term $c(\theta)$. However, it is not directly obvious how this correction term should be used for doubly intractable posteriors $q_\psi(\theta \in \cdot | s(x))$, where a clear distinction is made between the prior $p(\theta)$ and the likelihood $q_\psi(s(x) \in \cdot | \theta)$ in the algorithms. We elucidate this question in the following proposition.

Proposition V.2.1. *Let $s : \mathcal{X}_V \subset \mathcal{X} \mapsto \mathbb{R}^{d_s}$ be a summary-statistics function, and \mathcal{D}_n be n parameter-simulation pairs. Let $q_\psi(x \in \cdot | \theta = \diamond)$ be the likelihood model learned on the summary statistics by discarding the invalid simulations using $\{\theta_i, s(X_i)\}_{\{i \in (1, \dots, n), X_i \in \mathcal{X}_V\}}$, and let $\tilde{p}(x \in \cdot | \theta)$ be the conditional distribution of $s(X_i)$ given θ_i . Assume that $\tilde{p}(x \in \cdot | \theta) = q_\psi(s(x) \in \cdot | \theta)$ for some ψ . Then any*

auxiliary variable MCMC method on the corrected posterior $\tilde{q}_\psi(\theta \in \cdot | s(x))$

$$\tilde{q}_\psi(\theta \in \cdot | s(x)) := \tilde{p}(x \in \cdot | \theta) q_\psi(s(x) \in \cdot | \theta)$$

formed by the corrected prior $p(\theta | x \in \mathcal{X}_V)$ and the learned likelihood model $q_\psi(s(x) \in \cdot | \theta)$ has $p(\theta | s(x))$ as its stationary distribution.

The proof is given in Section V.A. We use this correction in our experiments to debias UNLE’s posterior estimate for learned on a real world neuroscience model.

V.3 Variational Inference Methods for UNLE

From the point of view of simulator efficiency, UNLE’s method has three major advantages: it uses a likelihood-based objective, an energy-based model, and does not require knowing the form of the proposal π used to draw the parameters θ_i . On the other hand, it comes with a reduced ease of posterior manipulation when compared to other methods. Indeed, posterior sampling requires approximate, auxiliary variable MCMC techniques to sample from its posterior. These methods are slower than standard MCMC methods, since every MCMC step require running a separate MCMC chain targeting the likelihood. Additionally, even posterior conditionals $q_\psi(\theta_{[i]} | x, \theta_{[-i]})$ or pairwise conditionals $q_\psi(\theta_{[i]}, \theta_{[j]} | x, \theta_{[-(i,j)]})$ cannot be visualized, unlike other methods like (S)NLE or (S)NPE, or (S)NRE. In this section, we propose two variational inference methods, respectively the amortized and the sequential case, that can be used to increase the tractability of UNLE’s posterior, at the cost of performing additional approximations. In this section, we first describe these two approaches, discuss the impact of these approximations, and provide guidance on when to use them.

V.3.1 Variational Amortized UNLE

In this section, we introduce our first variational inference method, dedicated to the amortized inference setting. This method, which we call Variational (Amortized) UNLE, assumes that the form of the proposal distribution π is known - this assumption is realistic in the amortized inference setting where the proposal distribution is

often set to be the prior distribution $p(\theta \in \cdot)$. Variational (Amortized) UNLE builds on this fact by directly fitting an alternative – tilted – energy-based model of this joint distribution, given by:

$$\begin{aligned} q_{\psi, \pi}((x, \theta) \in \cdot) &:= \pi(\theta \in \cdot) \mu(x \in \cdot) \frac{e^{-E_\psi(x, \theta)}}{Z_\pi(\psi)}, \\ Z_\pi(\psi) &= \int e^{-E_\psi(x, \theta)} \pi(d\theta) \mu(dx), \end{aligned} \quad (\text{V.8})$$

This model is an *unconditional* (joint) Energy-Based Model: as such, we use the standard (unconditional) approximate Maximum Likelihood learning technique for EBMs, such as contrastive divergence, to train it. Under appropriate assumptions, theoretical guarantees for this approach can be found in [247], matching the ones of Theorem V.2.1. Like UNLE’s joint model $\pi(\theta \in \cdot)q_\psi(x \in \cdot | \theta)$, the conditional distribution $q_{\psi, \pi}(x \in \cdot | \theta)$ of this joint model equals (up to a normalizing constant) the likelihood density $q_\psi(x \in \cdot | \theta)$ given by Equation (V.2). However, unlike UNLE’s joint model, its marginal distribution $q_{\psi, \pi}(\theta \in \cdot)$ equals:

$$q_{\psi, \pi}(\theta \in \cdot) = \int q_{\psi, \pi}(dx, \theta) = \pi(\theta \in \cdot) f_{\psi, \pi}(\theta)$$

where $f_{\psi, \pi}(\theta) := Z(\theta, \psi)/Z_\pi(\psi)$ is an intractable function of θ . This marginal depends on ψ , and does not equal *a priori* the marginal of the training data, namely the proposal distribution $p(x \in \cdot)$. Thus, unlike UNLE’s model, AUNLE’s model will have to learn to match π during training. While this property might make this model seem undesirable at first glance, the normalizing function $Z(\theta, \psi)$ of the parameter ψ^* learned by this model satisfies a property that significantly improves posterior sampling efficiency, as described in the following proposition:

Proposition V.3.1. *Assume that there exists a ψ^* such that $q_{\psi^*, \pi}((x, \theta) \in \cdot) = p_\pi((x, \theta) \in \cdot)$. Then any maximizer ψ of the AUNLE’s population objective satisfies*

$$\begin{aligned} q_{\psi^*, \pi}(x \in \cdot | \theta) &= p(x \in \cdot | \theta) \text{ and } Z(\theta, \psi) = Z_\pi(\psi^*), \\ \forall \theta \in \Theta. \quad (\text{V.9}) \end{aligned}$$

Proof. Building on the above paragraph's observation, we have that $q_{\psi,\pi}((x, \theta) \in \cdot) = f_{\psi,\pi}(\theta)\pi(\theta \in \cdot)q_{\psi}(x \in \cdot | \theta)$. At the optimum, the joint EBM $q_{\psi^*,\pi}((x, \theta) \in \cdot)$ will equal the training joint distribution $\pi(\theta \in \cdot)p(x \in \cdot | \theta = \diamond)$. Integrating out x on both sides of the equality yields $f_{\psi^*,\pi}(\theta)\pi(\theta \in \cdot) = \pi(\theta)$, proving the desired result. \square

Proposition V.3.1 shows that AUNLE indeed can learn the correct likelihood using a CEBM model stemming from the joint model $q_{\psi,\pi}((x, \theta) \in \cdot)$ *tilting* the proposal π with $f_{\psi,\pi}(\theta)$. The presence of π in the joint model is instead here to avoid the need for the optimal joint model to model π , which could be cumbersome if $\pi(\theta \in \cdot)$ is complicated. Proposition V.3.1 further shows that AUNLE's optimal likelihood model will have a normalizing function $Z(\theta, \psi^*)$ constant (or *uniform*) in θ , reducing AUNLE's posterior to a standard unnormalized posterior

$$q_{\psi^*,\pi}(\theta \in \cdot | x) = p(\theta \in \cdot) \frac{e^{-E_{\psi^*}(x, \theta)}}{Z_{\pi}(\psi^*)}. \quad (\text{V.10})$$

This property above will be approximately verified by the maximizers of AUNLE's empirical log-likelihood. Building on this observation, AUNLE sets its final posterior estimate to be:

$$\hat{q}_{\hat{\psi}_n,\pi}(\theta \in \cdot | x) \leftarrow p(\theta \in \cdot) \frac{e^{-E_{\hat{\psi}_n}(x, \theta)}}{Z_{\pi}(\psi^*)} \approx q_{\hat{\psi}_n,\pi}(\theta \in \cdot | x). \quad (\text{V.11})$$

AUNLE then performs inference using classical MCMC algorithms targeting $\hat{q}_{\hat{\psi}_n,\pi}$. The standard nature of AUNLE's posterior contrasts with the posterior of SMNLE [188], and allows to expand the range of inference methods applicable to it. In particular, unlike UNLE's posterior, AUNLE's posterior can be sampled from using importance (re-)sampling methods such as Sequential Monte Carlo, methods that can have inference accuracy superior to MCMC the posterior has multiple modes. The whole algorithm for AUNLE is summarized in Algorithm 6. The routing `maximize_ebm_log_1` stands for any approximate Maximum Likelihood learning method for EBMs, such as contrastive divergence [105].

Sources of Approximation in AUNLE Variational Inference Methods are methods that approximate a (less tractable) distribution with a more tractable one, at the cost of possibly reducing the accuracy of inference when there is a mismatch between the family to which the approximation belongs and the approximated distribution. So far, we have shown that AUNLE can improve inference tractability. But to what extent AUNLE's posterior is an approximation of UNLE's posterior? How does AUNLE's model class compare to UNLE's one, and in what sense is AUNLE's posterior an approximation of UNLE's one? UNLE and AUNLE share the same model class: as discussed before, for a given parameter ψ , the likelihood $q_{\psi,\pi}(x \in \cdot | \theta)$ of AUNLE's model is the same as UNLE's likelihood $q_\psi(x \in \cdot | \theta)$. However, for a given target likelihood $p(x \in \cdot | \theta)$, any parameter ψ such that $q_\psi(x \in \cdot | \theta) = g(\theta)p(x \in \cdot | \theta)/Z$ for some arbitrary function $g(\theta)$ and constant Z , will maximize UNLE's population objective. On the other hand, AUNLE's objective is maximized only when $q_{\psi,\pi}(x | \theta) = p(x | \theta)/Z$ for some constant Z . Thus, AUNLE's maximizers constitute a subset of UNLE's maximizers. In certain cases, these maximizers can be poorly conditioned, making them harder to estimate from samples. This observation gives substance to fact that AUNLE's posterior is an approximation of UNLE's posterior. We expect this to be the case when there exists a $g(\theta)$ such that $g(\theta)p(x \in \cdot | \theta)$ is better conditioned than $p(x \in \cdot | \theta)$. Interestingly, this exact situation arises in one synthetic experiment present in a well known SBI benchmark suite, which we discuss in the experiment section.

Algorithm 6 Amortized-UNLE

Input: proposal $\pi(\theta \in \cdot)$, simulator G , budget n , initial EBM parameters ψ_0
 $\mathcal{D}_n \leftarrow \{\theta_i \sim \pi(\theta \in \cdot), X_i \sim G(\theta_i, \cdot)\}$
 //See, e.g. [247]
 $\hat{\psi}_n \leftarrow \text{maximize_ebm_log_l}(\mathcal{D}_n, \psi_0)$
 $\hat{q}_{\hat{\psi}_n, \pi}(\theta \in \cdot | x) \leftarrow p(\theta \in \cdot) e^{-E_{\hat{\psi}_n}(x, \theta)} / f_{\hat{\psi}_n, \pi}(\theta)$
 $\{\theta_i^{\text{post}}\}_{i=1}^p \leftarrow \text{ApproxSample}(q_{\hat{\psi}_n, \pi}(\theta \in \cdot | x), x_o)$
Return: $\{\theta_i^{\text{post}}\}_{i=1}^p, \hat{q}_{\hat{\psi}_n, \pi}(\theta \in \cdot | x)$

V.3.2 Double Variational Inference

AUNLE is restricted to cases where the training parameters are drawn from a proposal π with a known analytical form. This precludes a naive adaptation of the multi-round version of UNLE used for sequential inference described in Algorithm 4. Indeed, the marginal distribution of the training parameters at round r resembles a mixture of all the proposals used for the round $1, \dots, r - 1$, e.g. a mixture of intractable posteriors. In this section, we propose an alternative, general-purpose approach for approximating a doubly-intractable posterior with a standard unnormalized posterior. This approach, which we call *Double Variational Inference* (DVI), proceeds by approximating the log-normalizing function $\log Z(\theta, \psi)$ of a UNLE posterior $q_\psi(\theta \in \cdot | x)$ of the form of Equation (V.4) using a neural network function $LZ_\eta(\theta)$ (η are the network's weights), and returning an approximate posterior obtained by plugging-in this approximation in the target posterior. DVI thus seeks an approximation of the form

$$\begin{aligned} q_{\psi, \eta}(\theta \in \cdot | x) &\propto p(\theta \in \cdot) e^{-E_\psi(x, \theta) - LZ_\eta(\theta)} \\ &\approx p(\theta \in \cdot) e^{-E_\psi(x, \theta) - \log Z(\theta, \psi)} (\propto q_\psi(\theta \in \cdot | x)). \end{aligned} \tag{V.12}$$

It remains to obtain a good approximation $LZ_{\eta_s}(\theta)$ the log-normalizing function. To do so, we first show that the function $\theta \mapsto \log Z(\theta, \psi)$ solves a specific minimization problem, as described in the following proposition.

Proposition V.3.2. *Assume that $\frac{dq_\psi}{d\mu}(x | \theta)$ is differentiable w.r.t θ , and let \mathcal{F} be the space of 1-differentiable real-valued functions on Θ . Let $v(\theta \in \cdot)$ be any distribution with full support on Θ , and let $f^* \in \mathcal{F}$. Then f^* is a solution of:*

$$\min_{f \in \mathcal{F}} \mathbb{E}_{q_\psi(x \in \cdot | \theta) v} \ell(x, \theta; f), \tag{V.13}$$

(where $\ell(x, \theta; f) := \|\nabla_\theta(f(\theta) + E_\psi(x, \theta))\|^2$) if and only if $f^* = \log Z(\theta, \psi) + C$, for some constant C .

The proof of Proposition V.3.2, given in Section V.E.1, relies on two key elements: the property of CEBM given in Equation (V.6), an extension of the well-known property

of EBMs given in Equation (V.7), and the characterization of conditional expectations as minimizers of least-squares losses. Importantly, Proposition V.3.2's loss function minimized by $\log Z(\cdot, \psi)$ takes the form of a sample average on $q_\psi(x \in \cdot | \theta) v(\theta \in \cdot)$. DVI, summarized in Algorithm 7, leverages this fact and produces its approximation $LZ_{\eta_s}(\theta)$ of $\log Z(\theta, \psi)$ by first obtaining samples $\{x_i, \theta_i\}_{i=1}^s \sim v(\theta \in \cdot) q_\psi(x \in \cdot | \theta)$ and minimizing the finite-sample analogue of Equation (V.13). The parameter η_s of its approximation LZ_η are given by:

$$\eta_s = \arg \min_{\eta} \mathcal{L}_{LZ,s}(\eta) = \arg \min_{\eta} \frac{1}{s} \sum_{i=1}^s \ell(x_i, \theta_i, LZ_\eta) \quad (\text{V.14})$$

The training samples $\{x_i, \theta_i\}_{i=1}^s$, are computed in parallel by sampling θ_i from the proposal v , and sampling $x_i | \theta_i$ using MCMC chains targeting $q_\psi(x \in \cdot | \theta_i)$ for each $i = 1 \dots s$. The following proposition bounds the difference between the DVI-approximated posterior, and UNLE's doubly intractable posterior.

Proposition V.3.3. *Let $q_\psi(\theta \in \cdot | x)$ be a posterior of the form of Eqaution V.4. Let $\mathcal{F}_H := \{LZ_\eta, \eta \in H\}$ be the class of log-normalizing function approximations, and assume that $q_\psi(\theta \in \cdot | x)$ takes the form $q_{\psi,\eta}(\theta \in \cdot | x) \propto q_{\psi,\eta^*}(\theta \in \cdot | x)$ for some $\eta^* \in H$.*

$$\mathcal{R}_s := \mathbb{E}_{\{\theta_i, x_i\}_{i=1}^s, \{\varepsilon_i\}_{i=1}^s} \sup_{f \in \mathcal{F}_H} \left| \frac{1}{s} \sum_{i=1}^s \varepsilon_i \ell(x_i, \theta_i, f) \right|$$

be the Rademacher complexity of \mathcal{F}_H with respect to the loss ℓ . Then, we have

$$\begin{aligned} & \mathbb{E}_{\theta \sim v} \mathbb{E}_{x \sim q_\psi(x \in \cdot | \theta)} \\ & \left\| \nabla_\theta \log \frac{dq_{\psi,\eta_s}(\theta | x)}{dp(\theta)} - \nabla_\theta \log \frac{dq_{\psi,\eta^*}(\theta | x)}{dp(\theta)} \right\|^2 \leq 2\mathcal{R}_s \end{aligned} \quad (\text{V.15})$$

Consequently, if $\lim_{s \rightarrow \infty} \mathcal{R}_s = 0$, we have

$$\begin{aligned} & \lim_{s \rightarrow \infty} \mathbb{E}_{\theta \sim v} \mathbb{E}_{x \sim q_\psi(x \in \cdot | \theta)} \\ & \left\| \nabla_\theta \log \frac{dq_{\psi,\eta_s}(\theta | x)}{dp(\theta)} - \nabla_\theta \log \frac{dq_{\psi,\eta^*}(\theta | x)}{dp(\theta)} \right\|^2 = 0 \end{aligned} \quad (\text{V.16})$$

The proof of Proposition 3.3 is given in Appendix V.E.2. Note that when $v(\theta \in \cdot)$ is

set to the prior distribution $p(\theta \in \cdot)$ Proposition V.3.3 guarantees that the expected Fisher divergence [117] between the DVI-approximated posterior and the target posterior goes to zero as the number of samples n goes to infinity, provided that the Rademacher complexity of the model class \mathcal{F}_H goes to zero.

Discussion DVI shares similarities with [174], which runs an approximation of the unadjusted langevin algorithm (ULA) to sample from the doubly-intractable posterior of undirected graphical models by computing sample-averages of $\nabla_\theta \log Z(\theta, \psi)$. Instead DVI estimates both $\log Z(\theta, \psi)$ and $\nabla_\theta \log Z(\theta, \psi)$, making it possible to use metropolis-hasting corrections of the ULA, in order to target the correct posterior distribution. Moreover, DVI computes these estimates ahead of time, in an *amortized* manner, e.g. for every $\theta \in \Theta$. Moreover, while we applied DVI to the case of likelihood given by Conditional EBMs, DVI can be applied to any model with intractable likelihood. Finally, note that like AUNLE, DVI satisfies the two defining properties of a variational inference method. It decreases the complexity of the inference procedure at the cost of performing an additional approximation. Unlike AUNLE however, DVI introduces a new set of parameters and a new learning procedure. The difficulty of this learning problem increases with dimension of the parameter space Θ , which is the dimension of $\log Z(\cdot, \psi)$'s domain. We thus expect DVI to perform better when Θ is of low dimensionality. Sampling from

Algorithm 7 DVI(\mathcal{D}, ψ, η)

Input: proposal $v(\theta \in \cdot)$, UNLE posterior $q_\psi(\theta \in \cdot | x = \diamond)$, compute budget s
 $\mathcal{E} \leftarrow \{\theta_i \sim v, x_i \sim \text{ApproxSample}(q_\psi(x \in \cdot | \theta_i))\}_{i=1}^s$
// Minimize using, e.g. Stochastic Gradient Descent
 $\eta_s \leftarrow \arg \min_{\eta \in H} \frac{1}{s} \sum_{i=1}^s \ell(x_i, \theta_i, LZ_\eta)$
Return $q_{\psi, \eta_s}(\theta \in \cdot | x) \propto p(\theta \in \cdot) e^{-E_\psi(x, \theta) - LZ_{\eta_s}(\theta)}$

doubly intractable posteriors is a well-studied problem. One the one end of the solution spectrum lie auxiliary-variable samplers, used for instance by UNLE. On the other end lie method that seek to approximate quantities involving the posterior's log-normalizer that would be needed to perform standard MCMC, which DVI falls into.

V.4 Experiments

In this section, we study the performance of UNLE and its variational approximations in three different settings: a toy model that highlights the failure modes of other synthetic likelihood methods, a series of benchmark datasets for SBI, and a real life neuroscience model.

Experimental details AUNLE and SUNLE are implemented using jax [71]. We approximate expectations of AUNLE’s joint EBM using 1000 independent MCMC chains with a Langevin kernel parameterised by a step size σ , that automatically update their step size to maintain an acceptance rate of 0.5 during a per-iteration warmup period, before freezing the chain and computing a final particle approximation. The particle approximations are persisted across iterations [251, 62] to reduce the risk of learning a “short run” EBM [182, 273] that would not approximate the true likelihood correctly (see Section V.F.3 for a detailed discussion). All experiments are averaged across 5 random seeds (and additionally 10 different observations x_o for benchmark problems). We provide all code² needed to reproduce the experiments of the paper. Training and inference are computed using a single RTX5000 GPU. For benchmark models, a single round of EBM training takes around 2 minutes on a GPU (see Section V.F.5).

V.4.1 A toy model with a multi-modal likelihood

First, we illustrate the issues that other SBI methods can face when applied to model certain distributions using a simulator with a bimodal likelihood. On such likelihoods, pushforward models approximating perfectly this distribution have a Lipschitz constant lower-bounded by a quantity proportional to $\exp(\sigma \|m_1 - m_2\|^2)$ where m_1 and m_2 are the locations of the two modes [215]. This property is likely to preclude flow-based models to efficiently fit the data distribution with few training samples, as gradient descent may find better conditioned, lower-norm models explaining the training data equally well. Moreover, in such cases, the asymptotic variance of the parameters accounting for the mode proportions learned by score-matching estimators is known to be significantly worse than the one learned from

²<https://github.com/pierreglaser/sunle>

maximum likelihood [136]. Figure V.0.1 indeed shows the likelihood models learned by NLE on this simulator, which exhibit the pathologies mentioned above: the score-matched likelihood only recovers a single mode of the likelihood, while the flow-based likelihood learned by NLE only becomes precise when the number of training samples is very large, at which point most of the more imprecise, better conditioned models have lower log-likelihood than the optima (but ill conditioned) one. Finally, we see that NRE and NPE, which use noise-contrastive estimation to learn the likelihood, require a significant number of training samples to learn the posterior with high precision. In contrast, UNLE estimates both the likelihood and the posterior accurately, even with a small (1000) number of training samples.

V.4.2 Results on SBI Benchmark Datasets

While very visible in the previous example, the gain in sample efficiency of UNLE can extend to other kind of likelihood models. To prove this, we next study the performance of AUNLE and SUNLE on 4 SBI benchmark datasets with well-defined likelihood and varying dimensionality and structure [160]:

SLCP: A toy SBI model introduced by [190] with a unimodal Gaussian likelihood. The dependence of $p(x \in \cdot | \theta)$ on θ is nonlinear, yielding a complex posterior.

The Lotka-Volterra Model [159]: An ecological model describing the evolution of the populations of two interacting species, usually referred to as predators and prey.
Two Moons: A famous 2-d toy model with posteriors comprised of two moon-shaped regions, and yet not solved completely by SBI methods.

Gaussian Linear Uniform: A simple Gaussian generative model, with a 10-dimensional parameter space.

These models encompass a variety of posterior structures (see Section V.F.2 for posterior pairplots): the two-moons and SLCP posteriors are multimodal, include cutoffs, and exhibit sharp and narrow regions of high density, while posteriors of the Lotka-Volterra model place mass on a very small region of the prior support. We compare the performance of AUNLE and SUNLE with NLE and its sequential analogue SNLE, respectively: NLE and SNLE represent the gold standard of current synthetic likelihood methods, and perform particularly well on benchmark problems

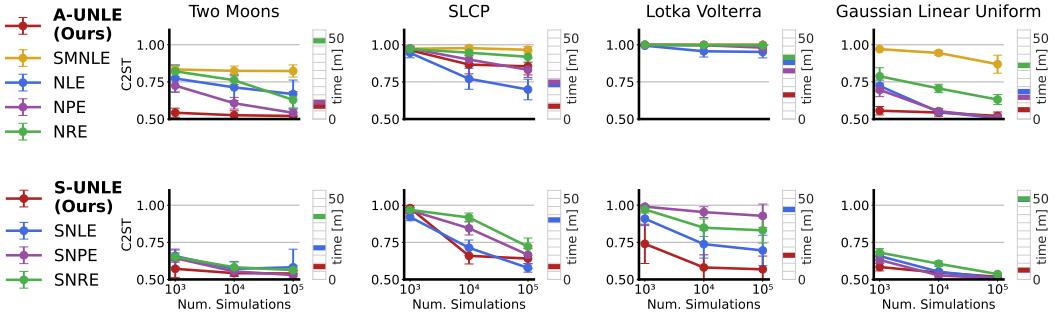


Figure V.4.1: Performance of AUNLE (resp. SUNLE) compared with NLE and SMNLE (resp. SNLE), using the Classifier Accuracy Metric [160] (lower is better). Runtime, in minutes, is also displayed for all methods . AUNLE and SUNLE exhibit robust performance across a wide array of problems. Additional details on the experimental setup can be found in Section V.F.6.

[160]. We use the same set of hyperparameters for all models, and use a 4-layer MLP with 50 hidden units and swish activations for the energy function. Results are shown in Figure V.4.1. All experiments used the DVI method to obtain posterior samples at each round.

While some fluctuations exist depending on the task considered, these results show that the performance of AUNLE (and SUNLE when targeted inference is necessary) is on par with that of (S)NLE, thus demonstrating that a generic method involving Energy-Based models can be trained robustly, without extensive hyperparameter tuning. Interestingly, the model where UNLE has the greatest advantage over NLE is Two Moons, which is the benchmark that exhibits a likelihood with the most complex geometry; in comparison, the three remaining benchmarks have simple normal (or log-normal) likelihood, which are unimodal distributions for which normalizing flows are particularly well suited. This point underlines the benefits of using EBMs to fit challenging densities.

Interestingly, we notice that in the case of SLCP, SUNLE performs as well as SNLE, while AUNLE performs worse than NLE. The reason is that the likelihood of the SLCP simulator is non-smooth, and diverges to $+\infty$ at $\theta_{3,4} = (0,0)$. The (Z, θ) -uniformity of AUNLE’s optimal likelihoods $q_{\psi^*}(x \in \cdot | \theta)$ makes its optimal energies $E_{\psi^*}(x, \theta)$ non-smooth in that case, and thus hard to estimate. In contrast, SUNLE, whose optimal likelihoods are not (Z, θ) -uniform, admits smooth optimal energies

for that problem, which are easier to estimate.

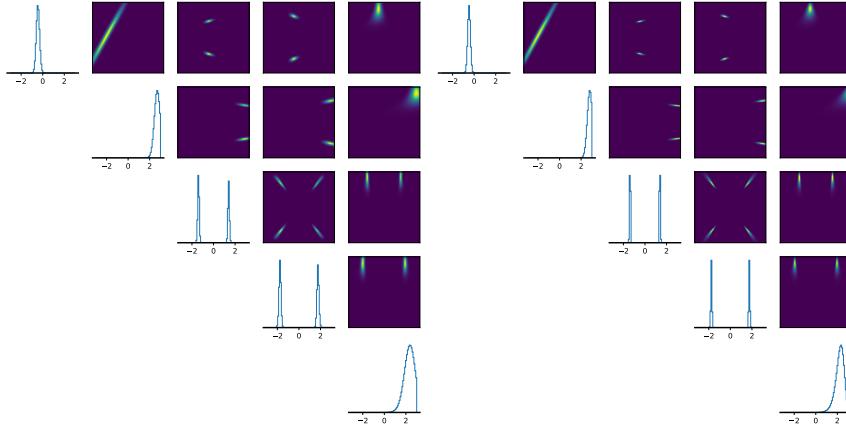


Figure V.4.2: Left: conditional pairplots $q_\psi(\theta_{[i]}, \theta_{[j]} | x, \theta_{[-(i,j)]})$ of SUNLE+DVI’s posterior estimate. Right: ground-truth conditional pairplots.

V.4.3 Using SUNLE in a Real World neuroscience model

We investigate further the performance of SUNLE by running its inference procedure on a simulator model of a pyloric network located in stomatogastric ganglion (STG) of the crab *Cancer borealis* given an observed an neuronal recording [94]. This model simulates 3 neurons, whose behaviors are governed by synapses and membrane conductances that act as simulator parameters θ of dimension 31. The simulated observations are composed of 15 summary statistics of the voltage traces produced by neurons of this network [197, 198]. The small volume of physiologically plausible regions of the parameter space Θ , coupled with the nonlinearity and high computational cost of running the model, make it a particular challenge for computational neuroscientists to fit to data (i.e., to characterize the regions of high probability of the posterior on θ). Indeed, fewer than 1% of draws from the prior on θ result in neural traces with well-defined summary statistics. Amortized SBI methods require tens of millions of samples for this problem; currently, the most sample-efficient targeted inference method is a variant of SNLE called SNVI [78] which uses 30 rounds, each simulating 10000 samples. We perform targeted inference on this model using SUNLE with a MLP of 9 layers and 300 hidden units per layers for the energy E_ψ . To maximize performance, and keeping in mind the high

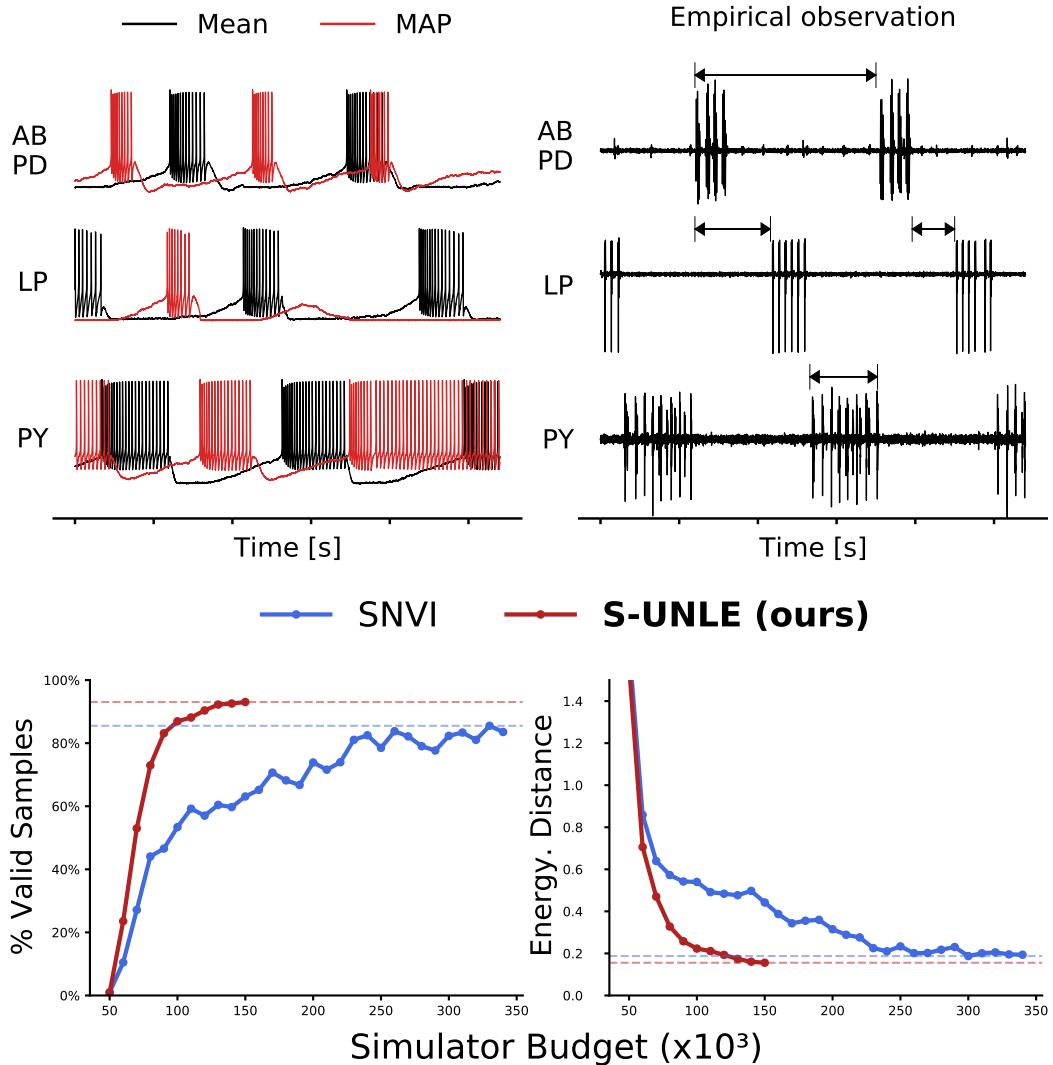


Figure V.4.3: Inference with SUNLE on a model of the pyloric network. Top-left: simulations obtained by using the final posterior mean and maximum a posteriori (MAP) as a parameter. Top-right: the empirical observation x_o : arrows indicate the summary statistics. Bottom-left: fraction of simulated observations with well-defined summary statistics (higher is better) at each round for SNVI and SUNLE, with dashed lines indicating the maximum fraction for each method. Bottom-right: performance of the posterior using the Energy Distance.

dimensionality of θ , we use doubly intractable MCMC instead of DVI to draw new proposal parameters across rounds. All inference and training steps are initialized using the previously available MCMC chains and EBM parameters. We report in Figure V.4.3 the evolution of the rate of simulated observations with valid summary statistics, - a metric indicative of posterior quality - as well as the Energy-Scoring Rule [80] of SUNLE and SNVI's posteriors across rounds. The synthetic observation simulated using SUNLE's posterior mean closely matches the empirical observation (Figure V.4.3, Left vs Center). As shown in Figure V.4.3, SUNLE matches the performance of SNVI in only 5 rounds, reducing by 6 the simulation budget of SNVI to achieve a comparable inference quality. After 10 rounds, SUNLE's posterior significantly exceeds the performance of SNVI in terms of number of valid samples obtained by taking the final posterior samples as parameters. The total procedure takes only 3 hours (half of which is spent simulating samples), *10 times less than SNVI*.

Conclusion The expanding range of applications of SBI poses new challenges to the way SBI algorithms model data. In this work, we presented SBI methods that use an expressive Energy-Based Model as their inference engine, fitted using Maximum Likelihood. We demonstrated promising performance on synthetic benchmarks and on a real-world neuroscience model. In future work, we hope to see applications of this method to other fields where EBMs have been proven successful, such as physics [184] or protein modelling [119].

Appendix

Supplementary Material for the paper *Maximum Likelihood Learning of Energy-Based Models for Simulation-Based Inference*

The supplementary materials include the following:

- Appendix [V.A](#) discusses details regarding the filtering of invalid simulations by UNLE, proving Proposition [V.2.1](#)
- Appendix [V.B](#) provides additional details on the coverage analysis of UNLE
- Appendix [V.C](#) proves Lemma [V.1.1](#)
- Appendix [V.D](#) provides a detailed statement of Theorem [V.2.1](#) and its proof.
- Appendix [V.E](#) provides the proof of Proposition [V.3.2](#) and [V.3.3](#)
- Appendix [V.F](#) provides additional experimental details.

V.A Filtering out invalid simulations

Prior to training its likelihood model, UNLE filters out parameter-simulation pairs that exhibit some properties that are problematic for training: in all settings UNLE discards pairs where the simulations that have very large numerical values compared to the average, as such simulations can hurt the stability of the training procedure. Additionally, for simulators that further process the raw simulation outputs into summary statistics (like the simulator of the pyloric networks), UNLE discards pairs where the simulation does not have a well-defined summary statistic. As noted in [78], discarding “invalid” pairs from the training set of synthetic likelihood methods based on the values of their simulations introduces a bias in the final posterior estimate, which will not learn the true posterior any more even as the number of training samples increases. [78] show that the posterior estimate $q(\theta|x_o)$ can be debiased by multiplying it with a correction factor $c(\theta)$ that can be estimated from the training data, and depends on the filtering procedure. However, [78]’s analysis does not discuss how this correction factor interacts with doubly-intractable MCMC methods used by UNLE. In this section, we study the impact of filtering out invalid simulations on the posterior estimate of AUNLE and SUNLE. For SUNLE, we show that while SUNLE’s posterior requires a filtering correction step. We also show how to incorporate this correction factor within doubly-intractable inference methods. Additionally, we show that AUNLE’s posterior estimate does not require any debiasing. Our analysis relies on a probabilistic view of the process of filtering out invalid simulations, which differs from the one presented in [78, 161, 24]. This view is, to the best of our knowledge, new, and allows to considerably simplify the analysis of the filtering correction factor.

V.A.1 A probabilistic view of invalid simulations filtering

In this section, we propose a simple probabilistic view to model the process of filtering out invalid simulations. Without loss of generality, let us assume that $S(x) = x$. Let us now denote by V the random variable defined as:

$$V := \mathbb{I}_{\{X \in \mathcal{X}_V\}}, \quad X \sim p(x|\theta), \quad \theta \sim \pi(\theta).$$

The graphical structure of the triplet (X, θ, v) is depicted below. Given n i.i.d pairs

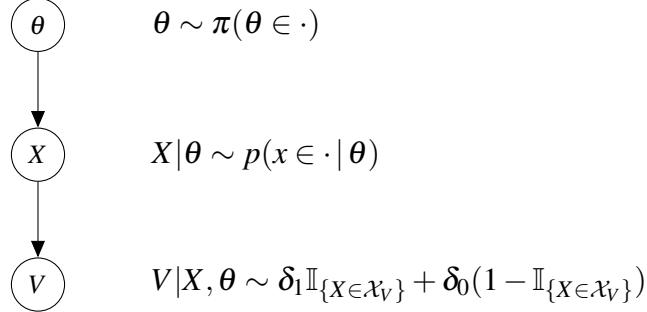


Figure V.A.1: Graphical model of the triplet (X, θ, V)

of samples $\{X_i, \theta_i\}_{i=1}^n$, let us denote by $\{X_i^v, \theta_i^v\}_{i=1}^{n_v}$ the subset of pairs with valid simulations, which form the final training set of UNLE. The training data marginally satisfies

$$\mathbb{P}[(X_i^v, \theta_i^v) \in \cdot] = \mathbb{P}[(X, \theta) \in \cdot | V = 1]$$

with associated likelihood $p_\pi((x, \theta) \in \cdot | V = 1)$. Importantly, the density of (X_i^v, θ_i^v) is proportional to the one of (X_i, θ_i) on \mathcal{X}_V ; in fact, so are the conditional distribution of $X_i^v | \theta_i^v$ and of $\theta_i^v | X_i^v$. Therefore, to evaluate $p(\theta \in \cdot | x)$ up to a normalizing constant from the filtered data, one can compute

$$\begin{aligned} p(\theta \in \cdot | x) &\propto p(\theta \in \cdot | x, V = 1) \\ &= \frac{dp(x \in \cdot | \theta, V = 1)}{dp(x \in \cdot | V = 1)}(x) \times p(\theta \in \cdot | V = 1) \\ &\propto \frac{dp_\pi(x \in \cdot | \theta, V = 1)}{dp_\pi(x \in \cdot | V = 1)}(x) \times p(\theta \in \cdot | V = 1) \end{aligned}$$

The first term will be estimated (up to a normalizing constant) by $q_\psi(x \in \cdot | \theta = \diamond)$, while the second term can be further decomposed as:

$$p(\theta \in \cdot | V = 1) = \frac{dp(V = 1 | \theta)}{dp(V = 1)} p(\theta \in \cdot) \propto p(V = 1 | \theta) p(\theta \in \cdot).$$

Now $p(V = 1 | \theta)$ can be estimated using the samples $\{\theta_i, V_i\}_{i=1}^n$ using an estimator $c_\zeta(\theta) \approx p(V = 1 | \theta)$, and we will have $c_\zeta(\theta) \rightarrow p(V = 1 | \theta)$ as $n \rightarrow \infty$. In practice, we parametrize the log-odds of the classifier by a neural network, and model the

classifier by composing the neural network with a sigmoid function. The final posterior estimate of SUNLE is thus given by:

$$\tilde{q}_\psi(\theta \in \cdot | s(x)) := q_\psi(\theta \in \cdot | s(x)) \times c_\zeta(\theta) \propto p(\theta \in \cdot | x)$$

This analysis is consistent with the analysis of [78]. However, it offers additional insights on the role of this correction term when the posterior is used in doubly-intractable inference methods, as we discuss below.

Incorporating correction terms into auxiliary variable methods The formalism of invalid simulations above allows us to characterize the target posterior distribution as the one of the joint system $p((x, \theta) \in \cdot | V = 1)$, with well characterized likelihood $p(x \in \cdot | \theta, V = 1)$, and prior $p(\theta \in \cdot | V = 1)$. This characterization is key to allow us to perform inference using doubly intratable methods [175, 168]. Indeed, such methods do not target standard unnormalized distributions, but posteriors that decomposes as the product of a likelihood $q(x|\theta) = f(x, \theta)/Z(\theta)$ with an unknown normalizing function $Z(\theta)$ and a prior $q(\theta)$. Implementing such methods by setting $q(x|\theta)$ to $p(x|\theta, V = 1)$ and $q(\theta)$ to $p(\theta|V = 1)$ thus guarantees correct posterior inference.

Incorporating correction terms in AUNLE As its objective function differs from the conditional log-likelihood objective of UNLE. As a consequence, AUNLE is not subject to neither the analysis of [78], nor the conclusions of our analysis made for UNLE. Instead, AUNLE learns a model of the filtered joint distribution $p(x, \theta|V = 1)$. Moreover, AUNLE parametrizes this joint distribution by:

$$p((x, \theta) \in \cdot) \approx \pi(\theta \in \cdot) \mu(x \in \cdot) e^{-E_\psi(x, \theta)} / Z_\pi(\psi).$$

Now, assuming a ψ^* such that $\pi(\theta \in \cdot) \mu(x \in \cdot) e^{-E_{\psi^*}(x, \theta)} / Z_\pi(\psi^*) \propto p_\pi((x, \theta) \in \cdot | V = 1)$, we have

$$e^{-E_{\psi^*}(x, \theta)} = \frac{dp_\pi(\theta \in \cdot | V = 1)}{d\pi(\theta \in \cdot)}(\theta) \frac{dp(x \in \cdot | \theta, V = 1)}{d\mu(x \in \cdot)} \propto p(V = 1 | \theta) \frac{dp(x \in \cdot | \theta, V = 1)}{d\mu(x \in \cdot)}$$

By identifying the two r.h.s of the equalities above, we have

$$\begin{aligned}
p(\theta \in \cdot | x) &\propto p(\theta \in \cdot | x, V = 1) \quad \text{on } \mathcal{X}_V \\
&= p(\theta \in \cdot | V = 1) \frac{dp(x \in \cdot | \theta, V = 1)}{dp(x \in \cdot | V = 1)}(x) \\
&\propto p(\theta \in \cdot) p(V = 1 | \theta) \frac{dp(x \in \cdot | \theta, V = 1)}{dp(x \in \cdot | V = 1)}(x) \\
&\propto p(\theta \in \cdot) \mu(x \in \cdot) e^{-E_{\psi^*}(x, \theta)}
\end{aligned}$$

Meaning that learned energy parameters are the only parameters needed for inference, and that no correction term is needed.

V.B Coverage Study

SBI posteriors provide an estimate of the inherent variation within the set of all parameters susceptible to have generated the observed data. While ideally, this estimated variation would match the ground truth, in practical settings, access to a finite amount of simulator budget will result in the presence of a gap. In particular, posteriors can fail to *cover* (e.g. assign adequate probability mass) some high probability region of a multi-modal posterior. Such posteriors will conceal scientists from credible alternative explanations of the observed data, and decrease the quality of the scientific discovery process. Diagnosing and addressing the overconfidence of SBI posteriors has been an important focal point of the SBI literature in recent years [103, 53, 54].

In this section, we perform a coverage analysis of the different variants of UNLE. We adopt the approach of [104] by computing the coverage probability of UNLE's posterior for varying confidence levels for a set benchmark problems. We show that UNLE exhibits similar level of calibration as popular SBI methods.

Background on coverage analysis We first briefly recall the relevant quantities appearing the coverage analysis of [104]. The definition of coverage involves the notion of *highest density region* (HDR), which we recall below.

Definition V.B.1 (Highest density region [116]). Below, we assume that the base

density $\mu(x \in \cdot)$ is the Lebesgue measure on \mathcal{X} . The highest density region of a probabilistic model $q_\psi(\theta \in \cdot | x = \diamond)$ with density $q_\psi(\theta | x)$ defined [see, e.g., 116] by

$$\Theta_{q_\psi(\theta \in \cdot | x)}(1 - \alpha) := \{\theta : q_\psi(\theta | x) \geq c_{q_\psi(\theta \in \cdot | x)}(1 - \alpha)\}$$

where

$$c_{q_\psi(\theta \in \cdot | x)}(1 - \alpha) := \sup\{c : \int \mathbb{I}_{\{q_\psi(\theta | x) \in [c, \infty)\}} q_\psi(d\theta | x) \geq 1 - \alpha\} \quad (\text{V.17})$$

The notion of HDR is used to define the central quantity of coverage analysis, the expected coverage probability.

Definition V.B.2 (Expected coverage probability of a Bayesian model [104]). Let $q_\psi(\theta | x)$ be the density of a density-based posterior estimate of some posterior $p(\theta \in \cdot | x)$ for all x . Then the expected coverage probability of the $1 - \alpha$ highest density region of the posterior model $q_\psi(\theta \in \cdot | x = \diamond)$ is defined as:

$$\mathbb{E}_{\mathbb{P}(x)} \left[\mathbb{P}(\theta \in \Theta_{q_\psi(\theta \in \cdot | x)}(1 - \alpha)) \right] \left(= \mathbb{E}_{(\theta, x) \sim \mathbb{P}(\theta, x)} \mathbb{I}_{\{\Theta_{q_\psi(\theta \in \cdot | x)}(1 - \alpha)\}} \right) \quad (\text{V.18})$$

This quantity represents the probability (on expectation over x) that a sample from the true posterior lies within the highest-density region of the estimated posterior. Plotting the expected coverage probability as a function of α yields the *coverage curve* of the posterior estimate, which describes fully the coverage properties of the posterior estimate. Unfortunately, exact values of expected coverage probabilities are not available analytically, and need to be approximate. The recipe used by [104] to estimate them goes as follow:

- Sample n pairs $\{(\theta_i, X_i)\}_{i=1}^{n_c}$ of parameter-observation from joint distribution $p((x, \theta) \in \cdot)$ formed by the prior and the simulator.
- Pick an α -level
- If the method is sequential, train a posterior estimate $q_\psi(\theta \in \cdot | X_i)$ for every sampled observation X_i . If the method is amortized, train a single amortized

posterior estimate which will be valid for all X_i .

- Estimate the highest density region of the posterior estimate by solving:

$$\widehat{c}_{q_\psi(\theta \in \cdot | X^{(i)})}(1 - \alpha) := \sup\{c: \frac{1}{n_c} \sum_{i=1}^{n_c} \mathbf{1}_{[c, \infty)}(q_\psi(\theta | X_i)) \geq 1 - \alpha\}$$

(yielding an approximate highest density region $\widehat{\Theta}_{q_\psi(\theta \in \cdot | X_i)}(1 - \alpha)$ given by $\{\theta: q_\psi(\theta | X_i) \geq \widehat{c}_{q_\psi(\theta \in \cdot | X_i)}(1 - \alpha)\}$, for each i , which is an approximate version of the optimization problem in (V.17), where the integral is approximated by a sum over a grid of points covering the support of the posterior estimate.

- Estimate the expected coverage probability by replacing the expectation over $p((x, \theta) \in \cdot)$ present in Equation (V.18) by an empirical expectation of over the samples $\{X_i, \theta_i\}_{i=1}^{n_c}$:

$$\frac{1}{n_c} \sum_{i=1}^{n_c} \mathbb{I}_{\{\theta_i \in \widehat{\Theta}_{q_\psi(\theta \in \cdot | X_i)}(1 - \alpha)\}}(\theta)$$

Using this recipe, an approximation of the true coverage curve for an arbitrary posterior estimate can be obtained. Note that the true posterior has an expected coverage probability of $1 - \alpha$ for all $\alpha \in [0, 1]$. In contrast, an overconfident posterior will have an expected coverage probability of less than $1 - \alpha$: this will be the case if, say the mode of the posterior estimate matches the mode of the true posterior, but the posterior estimate is too concentrated around its mode. As argued by [104], overconfident posterior can conceal user from credible regions of the true parameter, and to ensure reliable inference procedures, posteriors estimates from SBI method should exhibit limited (if not no) overconfidence. In the next section, we present the set of benchmarks used to evaluate the coverage properties of the posterior estimates of UNLE, and present the results obtained.

Coverage analysis of UNLE We compute the approximate coverage curves of AUNLE and SUNLE for a set of 4 benchmark problems which are inspired by the ones used in [104]: These benchmark problems are variants of the benchmark problems two moons, Gaussian linear uniform, Lotka-Volterra and SLCP. Since

the approximation of these approximate coverage curves requires discretization over a grid covering the prior support, and that then number of grid points scales exponentially with the dimension of the parameter space, the original gaussian linear uniform, lotka-volterra, and SLCP (of prior dimension respectively 10, 4 and 5) are modified to have a prior dimension of 2 by conditioning the original prior over the last $n - 2$ dimension using a fixed value $\theta_{2:n,0}$ for the last $n - 2$ dimensions. This conditioning operation is equivalent to replacing the prior $p(\theta)$ by the prior

$$p(\theta_{2:} | p_{\theta_{2:}}) \delta_{\theta_{2:,0}}(\theta_{2:})$$

We use $n = 10$ different pairs to compute the expectation. SUNLE was trained in 5 rounds. Both AUNLE and SUNLE were given a simulation budget of 1000 samples. The coverage probabilities were estimated for 10 different confidence levels, evenly-spaced between 0 and 10. For comparison, we also estimated the coverage curves of (S)NRE (S)NLE and SNPE. As the results show, UNLE exhibit comparable levels of over/under confidence as the other methods.

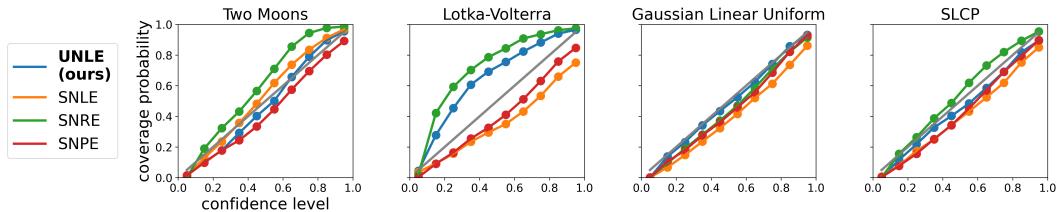


Figure V.B.1: Estimated coverage curves of UNLE, NLE, NPE, and NRE on 4 different benchmark problems. The posteriors were trained using a simulator budget of 1000 samples.

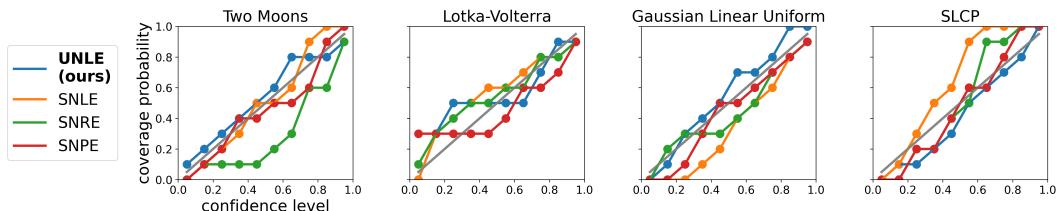


Figure V.B.2: Estimated coverage curves of SUNLE, SNLE, SNPE, and SNRE on 4 different benchmark problems. The posteriors were trained for 5 rounds with a simulator budget of 1000

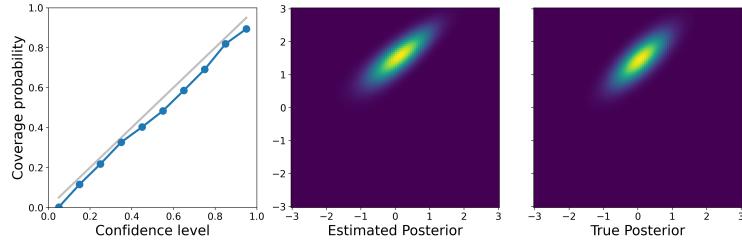


Figure V.B.3: Estimated coverage curves and posterior of UNLE on SLCP model. The posterior is plotted for an observation x_o sampled using a parameter θ from sampled from the prior.

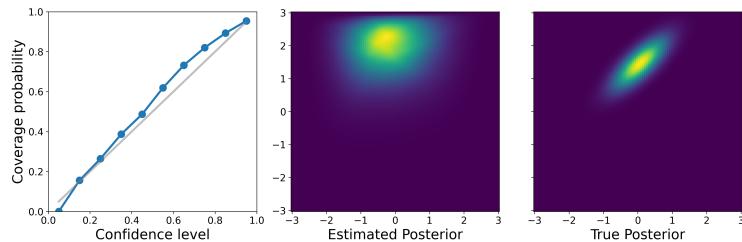


Figure V.B.4: Estimated coverage curves and posterior of NRE on SLCP model.

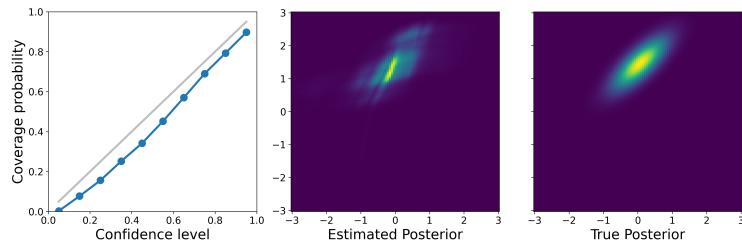


Figure V.B.5: Estimated coverage curves and posterior of NPE on SLCP model

V.C Proof of Lemma V.1.1

The proof relies on the following result of [215], which we recall below

Theorem (Corollary 2 of [215]). *Let $v = \lambda \mathcal{N}(m_1, \sigma^2 I) + (1 - \lambda) \mathcal{N}(m_2, \sigma^2 I)$ with $m_1, m_2 \in \mathbb{R}^d$, $\sigma > 0$ and $\lambda \in (0, 1)$. Assume there exists $g: \mathbb{R}^d \mapsto \mathbb{R}^p$ Lipschitz such that $g_\# \mu = v$. Then*

$$\text{Lip}(g) < \sigma \exp \left[\|m_1 - m_2\|^2 / (8\sigma^2) - \Phi^{-1}(\lambda)^2 / 2 \right]$$

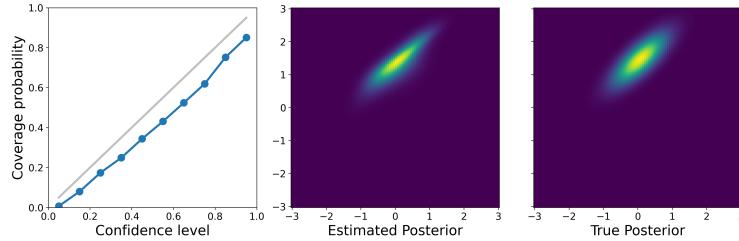


Figure V.B.6: Estimated coverage curves and posterior of NLE on SLCP model

With that in mind, we move on to the proof of Lemma V.1.1. We consider a probability space of the form $(\mathcal{X} \times \Theta, \mathcal{X} \otimes \mathcal{T}, \mathbb{P})$, where $\mathcal{X} \otimes \mathcal{T}$ is the product sigma algebra of \mathcal{X} and \mathcal{T} , and \mathbb{P} is a probability measure on $\mathcal{X} \times \Theta$ which by assumption is given by $\mathbb{P}(dx, d\theta) = p(dx, \theta) \otimes \pi(d\theta)$, where $p(dx, \theta)$ is regular version of the conditional probability given \mathcal{T} . Assume that there exists a some invertible mapping g which verifies the condition of Lemma V.1.1. Since $(A, \theta) \mapsto (g(\cdot | \theta) \mu)(A)$ is a regular version [see 43, Definition 2.4] of the conditional probability of X given θ , we have [see 43, Proposition 2.5] that for any measurable function $f : \Theta \times \mathcal{X} \mapsto \mathbb{R}$, that

$$\begin{aligned} \int_B \left(\int f(\theta, x) p(dx, \theta) \right) \pi(d\theta) &= \int_B \left(\int f(\theta, x) g(\cdot | \theta) \# \mu(dx) \right) \pi(d\theta) \\ \int_B \left(\int f(\theta, x) p(x, \theta) dx \right) \pi(d\theta) &= \int_B \left(\int f(\theta, x) g(\cdot | \theta) \# \mu(x) dx \right) \pi(d\theta) \end{aligned}$$

since $p(dx, \theta) = \lambda \mathcal{N}(m_1, \sigma^2 I) + (1 - \lambda) \mathcal{N}(m_2, \sigma^2 I)$ admits a Radon-Nikodym derivative on B , and so does $g(\cdot | \theta) \# \mu$ since g is invertible. Setting $f(\theta, x)$ to be the measurable (since both $\theta \mapsto g(\cdot | \theta) \# \mu(dA)$ and $\theta \mapsto p(A, \theta)$ are measurable since they are Markov kernels, and thus their Radon-Nikodym derivatives are measurable, see 210, Section 6.10) function $f(\theta, x) = (g(\cdot, \theta) \# \mu)(x) - p(x, \theta)$, we have that

$$\int \left(\int (g(\cdot, \theta) \# \mu(x) - p(x, \theta))^2 dx \right) \pi(d\theta) = 0$$

meaning that [see 242, Exercise 1.4.36(viii)] $\int (g(\cdot, \theta) \# \mu(x) - p(x, \theta))^2 dx = 0$, π -a.e. Let us note B_g the g -dependent, (π -) 0-measure set where this equality does not hold. Let $\theta \in B \setminus B_g$, which necessarily is not empty since $\pi(B \setminus B_g) \geq$

$\pi(B) - \pi(B_g) > 0$. On such θ , it holds that $g(\cdot, \theta)_\# \mu(x) = p(x, \theta)$, $\text{d}x$ -a.e, meaning that

$$\begin{aligned} \int f(x) g(\cdot, \theta)_\# \mu(x) \text{d}x &= \int f(x) p(x, \theta) \text{d}x \\ \iff \int f(x) g(\cdot, \theta)_\# \mu(\text{d}x) &= \int f(x) p(\text{d}x, \theta) \end{aligned}$$

(where the second equality holds by definition of $p(x, \theta)$ and $g(\cdot, \theta)_\# \mu(x)$) for any $\text{d}x$ -measurable function $f : \mathcal{X} \mapsto \mathbb{R}$. Consequently, $p(\cdot, \theta)_\# \mu(\text{d}x) = p(\text{d}x, \theta)$. We can now apply Theorem V.C, which yields that

$$\text{Lip}_x(g(\cdot | \theta)) < \sigma \exp \left[\|m_1 - m_2\|^2 / (8\sigma^2) - \Phi^{-1}(\lambda)^2 / 2 \right]$$

from which the result of Lemma V.1.1 follows.

V.D Proof of Theorem V.2.1

This section is devoted to the proof of Theorem V.2.1. Our results are an adaptation of [247], which established convergence guarantees for the (unconditional) contrastive divergence algorithm, to the conditional setting. However, importantly, to be as general as possible, we do not require the assumptions of [247] to hold uniformly over the conditioned variables, as the resulting assumptions may become overly restrictive.

Section V.D.1 details the problem setup, the algorithm, the assumptions, and the theorem statement. Section V.D.2 presents the proof of Theorem V.2.1.

V.D.1 Full statements of the setup, algorithm, assumptions, and of the theorem

V.D.1.1 Problem Setup

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{X}, \mathcal{X})$ and (Θ, \mathcal{T}) be two measurable spaces. Let $\{X_t, \theta_t\}_{t=1}^n$ be i.i.d random variables such that

$$\theta_t \sim \pi, \quad X_t | \theta_t \sim p(\bullet | \theta_t) \quad \forall t \in [n].$$

Here, π is a probability measure on Θ , and $p(\bullet | \diamond)$ is some Markov kernel, the “true” conditional distribution of X given θ . In this section, we seek to learn a model q_ψ of p using a conditional extension of exponential families as our model class—which we call Conditional Exponential Families, or CEF—defined, for all $\theta \in \Theta$, by

$$\begin{aligned} \frac{dq_\psi}{d\mu}(x|\theta) &:= c(x, \theta) e^{\langle \psi, \phi(x, \theta) \rangle - \log Z(\theta, \psi)}, \\ \log Z(\theta, \psi) &:= \log \left(\int c(x, \theta) e^{\langle \psi, \phi(x, \theta) \rangle} \mu(dx) \right), \\ \mathcal{D}_\theta &:= \{ \psi \in \mathbb{R}^p ; \log Z(\theta, \psi) < +\infty \}. \end{aligned} \quad (\text{V.19})$$

Here, μ is a measure on \mathcal{X} , ϕ is a $(\mathcal{X} \times \mathcal{T}, \mathcal{B}(\mathbb{R}^d))$ -measurable function, and c is a non-negative $(\mathcal{X} \times \mathcal{T}, \mathcal{B}(\mathbb{R}))$ -measurable function. At each θ , the set $\{q_\psi(\cdot|\theta), \psi \in \mathcal{D}_\theta\}$ defines an exponential family (with natural parameter ψ), whose (1) sufficient statistics $\phi(\bullet, \theta)$, (2) carrier density $c(\bullet, \theta)$ as well as—as a by-product—(3) log-normalizing function $\log Z(\theta, \psi)$ and (4) domain \mathcal{D}_θ now depend on the conditioned parameter θ . Such conditional probabilities are specific instances of conditional Energy-Based Models, with energy function $E_\psi(x, \theta) := -\langle \psi, \phi(x, \theta) \rangle - \log c(x, \theta)$. In the following, we will note

$$\mathcal{D} := \{ \psi ; \psi \in \mathcal{D}_\theta, \pi\text{-a.e} \}.$$

We learn some $\hat{\psi}_n$ such that $q_{\hat{\psi}_n} \approx p$ by (approximately) minimizing the loss function

$$\bar{\mathcal{L}}(\psi) := -\mathbb{E} \left[\log \frac{dq_\psi}{d\mu}(X_1 | \theta_1) \right] = -\mathbb{E} \left[\mathbb{E} \left[\log \frac{dq_\psi}{d\mu}(X_1 | \theta_1) \mid \theta_1 \right] \right] \quad (\text{V.20})$$

over some set $\Psi \subset \mathcal{D}$ using the data $\{\theta_t, X_t\}_{t=1}^n$. The r.h.s of (V.20), which follows from the law of total expectation, shows that $\bar{\mathcal{L}}(\psi)$ is the average of each $p_\psi(\bullet | \theta)$ ’s maximum likelihood loss over π .

Link with maximum likelihood estimation Strictly speaking, maximum likelihood estimation applies to unconditional probability models by minimizing their cross-entropy to the data. In contrast, we propose here to fit a conditional probability

model by minimizing its average cross-entropy to the data's conditional distribution, deviating from the framework of maximum likelihood estimation. However, it turns out that minimizing $\bar{\mathcal{L}}$ is equivalent to maximize the (unconditional) log-likelihood of the joint model $q_{\psi,\pi}$ on $(\mathcal{X} \times \Theta, \mathcal{X} \times \mathcal{T})$, defined by

$$\frac{dq_{\psi,\pi}}{d(\pi \times \mu)}(x, \theta) := \frac{dq_\psi}{d\mu}(x | \theta). \quad (\text{V.21})$$

By construction, (i) the marginal of this model in θ equals the proposal π , e.g. we have $q_{\psi,\pi}(\theta \in \cdot) = \pi(\theta)$, and (ii) the conditional distribution of X given θ is $q_\psi(\bullet | \theta)$. Moreover, it follows from the definition of $\bar{\mathcal{L}}$ and Equation V.21 that

$$\bar{\mathcal{L}}(\psi) = -\mathbb{E} \left[\log \frac{dq_{\psi,\pi}}{d(\mu \times \pi)}(X_1, \theta_1) \right],$$

that is $\bar{\mathcal{L}}$ also equals the cross-entropy between the joint model $q_{\psi,\pi}$ and the joint data distribution p_π . As the marginal distribution in θ of this joint model is the true marginal distribution of θ , intuitively, the model does not have to "learn" this marginal distribution.

The algorithm Let ψ_0 be a $(\Psi, \mathcal{B}(\Psi))$ -valued random variable independent of $\{\theta_t, X_t\}_{t=1}^n$. For each $\psi \in \Psi$ and $\theta \in \Theta$, to a Markov kernel (see, e.g. 16, Section 1.2.2) $k_{\psi,\theta}(x, dx')$ with invariant probability $q_\psi(x \in \cdot | \theta)$. Given a number of MCMC steps m , we note $k_{\psi,\theta}^m(x, dx')$ the Markov kernel obtained by chaining $k_{\psi,\theta}(x, dx')$ m times starting from x , e.g.

$$k_{\psi,\theta}^m(x, dx') := \int k_{\psi,\theta}(x, dx_1) \dots k_{\psi,\theta}(x_{m-2}, dx_{m-1}) \dots k_{\psi,\theta}(x_{m-1}, dx') \quad (\text{V.22})$$

Our algorithm consists in the following sequence of random variables, given by

$$\psi_t = \psi_{t-1} - \eta_t h_t, \quad h_t := -\phi(X_t, \theta_t) + \phi(K_{\psi_t, \theta_t, t}^m(X_t), \theta_t), \quad \forall t \in [n] \quad (\text{V.23})$$

Where $K_{\psi_t, \theta_t, t}^m(X_t)$ is the last-iterate MCMC sample at time t , e.g. the random variable admitting $k_{\psi_{t-1}, \theta_t}^m(X_t, dx')$ as a version of its conditional distribution given $\mathcal{F}_{t-1}, \theta_t, X_t$. The rationale for this algorithm is that, for sufficiently large m , h_t should

become a low-bias approximation of $\nabla_{\psi} \bar{\mathcal{L}}(\psi_t)$, which, as we will see in Lemma V.D.8, is given by

$$\nabla_{\psi} \bar{\mathcal{L}}(\psi) = -\mathbb{E}[\phi(X_t, \theta_t)] + \mathbb{E}\left[\phi\left(X_t^{\psi, \theta_t}, \theta_t\right)\right]$$

where, for all $\psi, \theta \in \Psi \times \Theta$, $X_t^{\psi, \theta}$ is a sample from the CEBM with parameter ψ and conditioned on θ , e.g.

$$X_t^{\psi, \theta} \mid \theta \sim p_{\psi}(\bullet \mid \theta), \quad \forall t \in [n].$$

V.D.1.2 Notations

Restricted Spectral Gaps Let $p_{\psi, \theta} := q_{\psi}(x \in \cdot \mid \theta)$. For a function $f \in L^2(p_{\psi, \theta})$, we define

$$\alpha(f, \psi, \theta) = \frac{\left(\int \left(\int (f - \int f(z)p_{\psi}(dz \mid \theta))(y)k_{\psi, \theta}(x, dy)\right)^2 p_{\psi}(dx \mid \theta)\right)^{1/2}}{\left(\int (f - \int f(z)p_{\psi}(dz \mid \theta))(y)(x)^2 p_{\psi, \theta}(dx)\right)^{1/2}} \quad (\text{V.24})$$

which is a measure of how quick a Markov chain with kernel $k_{\psi, \theta}$ mixes, relative to the function f [152]. Define $\phi_{\theta} := x \mapsto \phi(x, \theta)$, and $\phi_{\theta, i}$ its i -th component. Let $\mathcal{F}_{\theta} := \{\phi_{\theta, i}\}_{i=1}^p \cup \{\phi_{\theta, i}\phi_{\theta, j}\}_{i,j=1}^p$, and note:

$$\alpha(\psi, \theta) := \sup \{\alpha(f, \psi, \theta) \mid f \in \mathcal{F}_{\theta}\}, \quad \vartheta(\psi, \theta) := -\log(1 - \alpha(\psi, \theta)). \quad (\text{V.25})$$

$\alpha(\theta, \psi)$ can be described using the language of Markov Operators [16]. Given some kernel $k_{\psi, \theta}^m$, we define the operator $P_{\psi, \theta}^m$ as an operator on $L^2(p_{\psi, \theta})$ defined by:

$$(P_{\psi, \theta}^m f)(x) := \int f(y)k_{\psi, \theta}^m(x, dy) = \mathbb{E}\left[f\left(K_{\psi, \theta_1}^m(X_1)\right) \mid X_1 = x\right] \quad \forall f \in L^2(p_{\psi, \theta}), x \in \mathcal{X}$$

Note that indeed, $P_{\psi, \theta}^m f \in L_2(p_{\psi}(\cdot \mid \theta))$, as :

$$\begin{aligned} \int \|P_{\psi, \theta} f\|^2 d p_{\psi}(\cdot \mid \theta) &\stackrel{(a)}{\leq} \int P_{\psi, \theta} \|f\|^2 d p_{\psi}(\cdot \mid \theta) = \int \|f\|^2 k_{\psi, \theta}(y, dx) d p_{\psi}(\cdot \mid \theta)(x) \\ &\stackrel{(b)}{=} \int \|f\|^2 d p_{\psi}(\cdot \mid \theta), \end{aligned}$$

where we noted $P_{\psi,\theta}^1 = P_{\psi,\theta}$, (a) follows from Jensen's inequality, and (b) holds by definition of $p_\psi(\cdot | \theta)$ being the invariant distribution of $k_{\psi,\theta}$. With such notations, we have

$$\alpha(f, \psi, \theta) = \frac{\left(\int (P_{\psi,\theta} f - \mathbb{E}_{p_{\psi,\theta}} f)^2 dP_{\psi,\theta}(x) \right)^{1/2}}{\left(\int (f - \mathbb{E}_{p_{\psi,\theta}} f)(x)^2 p_{\psi,\theta}(dx) \right)^{1/2}}, \quad (\text{V.26})$$

where we used the fact that $P_{\psi,\theta}(\mathbf{1}) = \mathbf{1}$, where $\mathbf{1}$ is the function equal to 1 everywhere. This identity will often be used in the sequel.

V.D.1.3 Assumptions

Our analysis requires four assumptions to be satisfied: [A7](#), [A8](#), [A9](#), and one variant of [A10](#) (either [A10](#), [A10'](#), [A10''](#), or [A10'''](#)).

Assumption A7. (i) ϕ is $(\mathcal{X} \times \mathcal{T}, \mathcal{B}(\mathbb{R}^{d_\psi}))$ -measurable, c is $(\mathcal{X} \times \mathcal{T}, \mathcal{B}(\mathbb{R}))$ -measurable and non-negative, (ii) $\mathbb{E}[\phi(X_1, \theta_1)] \in \mathbb{R}^d$ and $\mathbb{E}[\log c(X_1, \theta_1)] \in \mathbb{R}$, (iii) $\text{int}(\mathcal{D}) \neq \emptyset$, (iv) $\{p_\psi(\bullet | \theta) ; \psi \in \mathcal{D}_\theta\}$ is minimal π -almost-everywhere, and (v) Ψ is a convex and compact subset of \mathcal{D} with non-empty interior, and $\psi^* \in \text{int}(\Psi)$.

Assumption A8. There exists $\tau < +\infty$ s.t. $\sup_{\psi \in \Psi} (\mathbb{E}[\|\phi(X^{\psi, \theta_1}, \theta_1)\|^8 | \theta_1])^{1/8} \leq \tau$.

Assumption A9. There exists a function $C_\chi(\theta)$ such that $\chi^2(p_\psi(\cdot | \theta), p_{\pi, \psi^*}(\cdot | \theta)) \leq C_\chi^2(\theta) \|\psi - \psi^*\|^2$. Moreover, it holds that $\mathbb{E}[C_\chi^2(\theta_1)] =: \bar{C}_\chi^2 < +\infty$.

Assumption A10. It holds that $\vartheta_\kappa := \sup\{\mathbb{E}[\vartheta(\alpha_1, \psi)^\kappa]^{1/\kappa}, \psi \in \Psi\} < +\infty$, for some $\kappa \geq 1$.

Assumption A10'. For all $\psi \in \Psi$, $\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)]$ is λ -sub-exponential e.g. $\mathbb{E}[e^{s(\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)])}] \leq e^{\lambda^2 s^2 / 2}$ for all $s \leq 1/\lambda$.

Assumption A10''. For all $\psi \in \Psi$, $\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)]$ is σ -sub-Gaussian e.g. $\mathbb{E}[e^{s(\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)])}] \leq e^{\sigma^2 s^2 / 2}$ for all $s \in \mathbb{R}$.

Assumption A10'''. For all $\psi \in \Psi$, $\vartheta(\theta_1, \psi) \leq \vartheta_{\max} < 1$ almost surely.

Our assumptions adapt the ones of [121, 247], who studied the (unconditional) contrastive divergence algorithm, to the conditional setting.

Assumption A7 is basic assumption to ensure that this problem is well-defined is that $\bar{\mathcal{L}}$ indeed admits a unique minimizer. To guarantee it, we simply need to ensure that this loss is well-defined in the sense that the domain of $\bar{\mathcal{L}}$ is non-empty (we will then show that a unique minimizer exists by showing that $\bar{\mathcal{L}}$ is convex and appropriately restricting Ψ). We do so through A7: with it, the terms $\mathbb{E}[\langle \psi, \phi(X_1, \theta_1) \rangle]$ and $\mathbb{E}[\log Z(\theta_1, \psi)]$ are both well-defined, and it follows that $\bar{\mathcal{L}}$, given by:

$$\bar{\mathcal{L}}(\psi) = -\mathbb{E}[\langle \psi, \phi(X_1, \theta_1) \rangle] + \log c(x, \theta) - \log Z(\theta_1, \psi)$$

admits a non-empty domain. Furthermore, (ii) and (iii) also ensure that this domain has a non-empty interior, which is a prerequisite to establish the differentiability of $\bar{\mathcal{L}}$. Moreover, it becomes possible to break down $\bar{\mathcal{L}}$ in two separate ψ -dependent terms (and one constant in ψ):

$$\bar{\mathcal{L}}(\psi) = -\mathbb{E}[\langle \psi, \phi(X_1, \theta_1) \rangle] + \mathbb{E}[\log Z(\theta_1, \psi)] + \underbrace{\mathbb{E}[\log c(X_1, \theta_1)]}_{\text{constant in } \psi}, \quad \forall \psi \in \text{dom}(\bar{\mathcal{L}})$$

which is helpful as our algorithm uses separate stochastic estimates for each of the two ψ -dependent terms gradients.

Assumption A8 is specific to the conditional setting: in the unconditional setting, the log-partition function $\mathcal{Z}(\psi)$ is known to be smooth on the interior of the natural parameter space, without any additional assumption. In our setting however, it is not guaranteed that the *average* log-partition function $\mathbb{E} \log Z(\theta_1, \psi)$ is smooth, a property which is important to make gradient-based algorithm like ours well-defined. As we will see, A8, along with the other assumptions, will ensure that this function is indeed smooth. We refer to Section V.D.2.7 for a discussion on the necessity of this assumption, which is automatically verified in the unconditional setting, but does not necessarily hold in the conditional setting.

Assumptions A9 and A10 are conditional analogues of the assumptions made in [121, 247]: A9 allows to smoothly map differences in distribution spaces via differences in parameter space, while A10 ensures that the MCMC kernels used in the algorithm mix. Importantly however, we do not simply impose the same

assumptions as in [121, 247] uniformly over θ , as this would be overly restrictive. Instead, we only require these assumptions to hold on average over θ . The variants A10 were made to control (at various strengths) the mixing of the Markov Chains used in the algorithm. Recall that when $\alpha(\psi, \theta) = 1$, does not mix at all. As soon as $\alpha(\psi, \theta) < 1$, the Markov chain mixes, but the closer $\alpha(\psi, \theta)$ is to 1, the slower the mixing. When considering a random θ_1 , the concentration of $\alpha(\psi, \theta_1)$ around 1 will thus have a strong impact. We propose to quantify this concentration around 1 through the tails of the random variable $\vartheta(\psi, \theta_1) = -\log(1 - \alpha(\psi, \theta_1))$. As we will see, the lighter the tail will be, the smaller the number of MCMC steps m required to ensure convergence of the algorithm.

We note that A10'' is stronger than A10', which is itself stronger than A10. Replacing \mathcal{F}_θ by $\mathcal{L}_2(p_{\psi, \theta})$, Assumption A10 would imply that $p_{\psi, \theta}$ admits a $(1 - \alpha(\psi, \theta))$ - $\mathcal{L}_2(p_{\psi, \theta})$ spectral gap, for all ψ and $p_{\psi, \theta}$ -almost all θ . In comparison, ensuring A10 using \mathcal{F}_θ , as \mathcal{F}_θ is much smaller than $\mathcal{L}_2(p_{\psi, \theta})$.

V.D.1.4 Main Result

We are now ready to state our convergence result.

Theorem 2.1. *Assume Assumptions A7 to A9, and any variant of A10. Let $\{\psi_t\}_{t \geq 0}$ be the sequence generated by (V.23) using a step-size schedule given by $\eta_t := Ct^{-\beta}$, for $\beta \in (0, 1)$, and using a number of MCMC steps m satisfying $m := m(n) > \frac{1}{4} \times F^{-1}(n^{-2(1-\beta)})$, where F is given in Lemma V.D.4. Then, for all $n \geq 1$, it holds that Define $\bar{\psi}_n := \frac{1}{n} \sum_{i=1}^n \psi_i$. Then,*

$$\mathbb{E} [\|\hat{\psi}_n - \psi^*\|^2] \leq 2 \frac{\text{tr}(\mathcal{I}(\psi^*)^{-1})}{n} + o(n^{-1}),$$

where $\mathcal{I}(\psi^*) := \mathbb{E}[\text{Cov}[\phi(X_1, \theta_1) | \theta_1]]$ is the Fisher information matrix of the model at ψ^* (see Equation V.42). Consequently, we have that $\limsup_{n \rightarrow \infty} n \times \mathbb{E} [\|\hat{\psi}_n - \psi^*\|^2] \leq 2 \times \text{tr}(\mathcal{I}(\psi^*)^{-1})$.

In addition to handling challenges associated with working with conditional distributions, this result constitutes an improvement over 247, which contained a factor of 4

instead of 2 in front of the leading term.

V.D.2 Proof of Theorem 2.1

The proof follows a similar strategy as the one from [247]: we first derive convergence guarantees of the non-averaged iterates ψ_n . We then leverage this result to prove the convergence of the averaged iterates $\hat{\psi}_n$. Note that this result does not require m to depend on n . The rest of this section is structured as follows:

- Section V.D.2.2 contains the main proof of Theorem 2.1, proving the claimed bound on the mean-squared error of the averaged iterates $\hat{\psi}_n$ to ψ^* .
- Section V.D.2.3 establishes the convergence of the non-averaged iterates ψ_n to ψ^* .
- Section V.D.2.4 quantifies the (additional) bias and variance of the conditional CD gradient $h_n(\psi_{n-1})$, relative to the true gradient of the Maximum Likelihood Loss.
- Section V.D.2.5 provides some properties of the Maximum Likelihood Loss and of its (inaccessible) unbiased stochastic gradient.
- Section V.D.2.6 contains some auxiliary lemmas.
- Section V.D.2.7 discusses the necessity of a moment assumption specific to the conditional setting.
- Section V.D.2.8 contains some background on exponential families, cumulants and convexity.

V.D.2.1 Proof Notations

Here, we lay out some notations that will be used throughout the proof.

Conditional Moments Given a random variable $U \in \mathcal{U}$, n random variables V_1, \dots, V_n and a measurable function $f : \mathcal{U} \rightarrow \mathbb{R}^d$, we denote by $\mu_{k,\Sigma}(f(Y_t) \mid \theta_t)$ the k^{th} moment of $f(Y_t)$ given θ_t , e.g.

$$\mu_{k,\Sigma}(f(U) \mid V_1, \dots, V_n) := \mathbb{E} \left[(f(Y_t) - \mathbb{E}[f(Y_t) \mid V_1, \dots, V_n])^k \mid V_1, \dots, V_n \right]$$

For any $(\mathcal{X} \times \mathcal{T}, \mathcal{B}(\mathbb{R}))$ -measurable function f , and any $(\psi, \theta) \in \Psi \times \Theta$, we denote by $\bar{f}_\psi(x|\theta)$ the version of f , centered around the expectation of $f(X_t^{\psi, \theta})$ i.e.

$$\bar{f}_\psi(x|\theta) := \phi(x) - \mathbb{E} [\phi(X_t^{\psi, \theta})].$$

Random variables and Filtrations In the following, we will note $(\mathcal{F}_t)_{t \geq 1}$ the sigma-algebra generated by the sequence of random variables $(\theta_t, X_t, \psi_t, K_{\psi_{t-1}, \theta_t}^m(X_t))_{t \geq 1}$ (with $\mathcal{F}_0 = \sigma(\psi_0)$). Moreover, for a fixed triplet (ψ, x, θ) , we denote by $K_{\psi, \theta, t}^m(x)$, $t \in [n]$ i.i.d random variables independent of \mathcal{F}_n distributed according to $k_{\psi, \theta}^m(x, \bullet)$, e.g.

$$K_{\psi, \theta, 1}^m(x), \dots, K_{\psi, \theta, n}^m(x) \stackrel{\text{i.i.d.}}{\sim} k_{\psi, \theta}^m(x, \bullet).$$

Similarly, we denote, for all (ψ, x, θ) , by $h_t(\psi, x, \theta)$, $t \in [n]$ the i.i.d random variable independent of \mathcal{F}_n given by

$$h_t(\psi, x, \theta) := -\phi(x, \theta) + \phi(K_{\psi, \theta, t}^m(x), \theta), \quad t \in [n],$$

and we note

$$h_t := h_t(\psi_{t-1}, X_t, \theta_t).$$

Finally, in our proofs, we will often compare the conditional CD gradient to the following unbiased stochastic gradient of the Maximum Likelihood Loss:

$$g_t := g(X_t, \theta_t, \psi_{t-1}) := \phi(X_t, \theta_t) - \mathbb{E} [\phi(X^{\psi_{t-1}, \theta_t}, \theta_t) \mid \mathcal{F}_t, \theta_t]$$

V.D.2.2 Proof of Theorem 2.1

Proof of Theorem 2.1.

Notations Let us note

$$\bar{h}_t := \bar{h}(\psi_{t-1}) := \mathbb{E}[h_t \mid \mathcal{F}_{t-1}] = -\mathbb{E}[\phi(X_t, \theta_t)] + \mathbb{E} [\phi(K_{\psi_{t-1}, \theta_t, t}^m(X_t), \theta_t) \mid \mathcal{F}_{t-1}], \quad t \in [n],$$

which marginalizes h_n over the data and Markov chain kernel at time t , given ψ_{t-1} . Moreover, we denote by $f'(\psi_t), f''(\psi_t)$ the Gradient and Hessian the Maximum Likelihood Loss at ψ_t , given by

$$\begin{aligned} f'(\psi_{t-1}) &:= -\mathbb{E}[\phi(X_t, \theta_t)] + \mathbb{E}[\phi(X^{\psi_{t-1}, \theta_t}, \theta_t)] \\ f''(\psi_{t-1}) &:= -\mathbb{E}[\text{Cov}[\phi(X^{\psi_{t-1}, \theta_t}, \theta_t) | \theta_t]], \quad t \in [n] \end{aligned} \quad (\text{V.27})$$

It holds that:

$$\begin{aligned} f''(\psi^*)(\psi_{n-1} - \psi^*) &= f'(\psi_{n-1}) - f'(\psi^*) + (f''(\psi^*)(\psi_{n-1} - \psi^*) - f'(\psi_{n-1}) + f'(\psi^*)) \\ &= h_n + (f''(\psi^*)(\psi_{n-1} - \psi^*) - f'(\psi_{n-1}) + f'(\psi^*)) \\ &\quad + (f'(\psi_{n-1}) - \bar{h}_n) + (\bar{h}_n - h_n) \end{aligned}$$

Summing over n , we have:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \psi_i - \psi^* &= \underbrace{\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} h_i(\psi_{i-1})}_{(i)} \\ &\quad + \underbrace{\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (f''(\psi^*)(\psi_{i-1} - \psi^*) - f'(\psi_{i-1}) + f'(\psi^*))}_{(ii)} \\ &\quad + \underbrace{\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (f'(\psi_{i-1}) - \bar{h}(\psi_{i-1}))}_{(iii)} + \underbrace{\frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}_i - h_i)}_{(iv)}. \end{aligned}$$

Thus, by Minkowski's inequality, we have

$$\mathbb{E}[\|\bar{\psi}_n - \psi^*\|^2]^{1/2} \leq \mathbb{E}[(i)^2]^{1/2} + \mathbb{E}[(ii)^2]^{1/2} + \mathbb{E}[(iii)^2]^{1/2} + \mathbb{E}[(iv)^2]^{1/2}$$

All terms can be handled as in [169], aside from the third and fourth term, which contains both the bias and the variance of CD.

We now proceed to bound each term:

Bounding (i) (i) can be bounded using [247, Lemma D.6]: The only condition needed to reuse their steps is that ψ_i satisfies an upper bound of the same form as

theirs, and a finite bound on $\mathbb{E} \|h_n\|^2$.

Bounding (ii) To bound (ii) notice that $\bar{\mathcal{L}}$ is thrice differentiable, with a continuous third derivative. Consequently, there exists some M such that

$$\|(f''(\psi^*)(\psi_{i-1} - \psi^*) - (f'(\psi_{i-1}) - f'(\psi^*)))\| \leq M \|\psi_{i-1} - \psi^*\|^2$$

Implying

$$\begin{aligned} \mathbb{E} [(ii)^2]^{1/2} &\leq \frac{1}{n} \sum_{i=1}^n \left(\mathbb{E} [\|(f''(\psi^*)(\psi_{i-1} - \psi^*) - (f'(\psi_{i-1}) - f'(\psi^*)))\|^2] \right)^{1/2} \\ &\leq \frac{1}{n} \sum_{i=1}^n M \times \left(\mathbb{E} [\|\psi_{i-1} - \psi^*\|^4] \right)^{1/2} = \mathcal{O}(n^{-\beta}) \end{aligned}$$

Bounding (iii) To bound (iii), notice that

$$\begin{aligned} f'(\psi) - \bar{h}(\psi) &= \mathbb{E} [P_{\psi, \theta_t}^m \phi(X_t, \theta_t)] - \mathbb{E} [\phi(X_t^{\psi, \theta_t}, \theta_t)] \\ \implies \|f'(\psi_{i-1}) - \bar{h}(\psi_{i-1})\| &\leq p \times \tau C_\chi \|\psi_i - \psi^*\| \times F(4m)^{1/4}, \quad \forall i \in [n], \end{aligned}$$

where we used Lemma V.D.17. Consequently, by Minkowski's inequality, we have

$$\begin{aligned} \mathbb{E} [(iii)^2]^{1/2} &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|f'(\psi_{i-1}) - \bar{h}(\psi_{i-1})\|^2]^{1/2} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\mathbb{E} [\|f'(\psi_{i-1}) - \bar{h}(\psi_{i-1})\|^2 | \mathcal{F}_i] \right]^{1/2} \\ &\leq p \tau C_\chi F(4m(n))^{1/4} \times \frac{1}{n} \sum_{i=1}^n \mathbb{E} [\|\psi_{i-1} - \psi^*\|^2]^{1/2} \\ &= p \tau C_\chi F(4m(n))^{1/4} \times \frac{1}{n} \sum_{i=1}^n \delta_i^{1/2} \end{aligned}$$

Since, by Theorem V.D.1, for all $i \in [n]$, we have

$$\delta_i^{1/2} \leq 2 \exp(2\tilde{L}C^2\varphi_{1-2\beta}(i)) \exp\left(-\frac{\tilde{\mu}_m C}{2} i^{1-\beta}\right) \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right)^{1/2} + \left(\frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m}\right)^{1/2} \times \frac{1}{i^{\beta/2}},$$

It holds that

$$\begin{aligned} \frac{F(4m(n))^{1/4}}{n} \sum_{i=1}^n \delta_i^{1/2} &\leq \frac{F(4m(n))^{1/4}}{n} \left(\sum_{i=1}^n 2 \exp(2\tilde{L}C^2 \varphi_{1-2\beta}(i)) \exp\left(-\frac{\tilde{\mu}_m C}{2} i^{1-\beta}\right) \times \right. \\ &\quad \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2} \right)^{1/2} + \left(\frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m} \right)^{1/2} \varphi_{1-\frac{\beta}{2}}(n) \Big) \\ &= \mathcal{O}\left(F(4m(n))^{1/4} n^{-\beta}\right) = \mathcal{O}(n^\gamma) \end{aligned}$$

where we noted $\varphi_\beta(t) = \frac{t^\beta - 1}{\beta}$ if $\beta \neq 0$, and $\log t$ if $\beta = 0$, and

$$\gamma := -\frac{\beta}{2} + \frac{\log F(4m(n))}{4 \log n} < -\frac{1}{2}$$

as, by assumption, we have:

$$m(n) > \frac{1}{4} \times F^{-1}(n^{-2(1-\beta)})$$

Consequently, we have that $\mathbb{E}[(iii)^2]^{1/2} = \mathcal{O}(n^\gamma) = o(n^{-1/2})$.

Bounding (iv) (iv) represents the variance of the conditional CD gradient. Importantly, it is—by construction—a sequence of martingale differences. Consequently, we have:

$$\begin{aligned} &\left(\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}_i - h_i) \right\|^2 \right] \right)^{1/2} \\ &= \left(\frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}_i - h_i) \right\|^2 \right] \right)^{1/2} \\ &= \frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[\text{Cov}[h_i | \mathcal{F}_{i-1}]] f''(\psi^*)^{-1}) \right)^{1/2} \end{aligned}$$

where the second line holds by definition of $\bar{h}(\psi_{i-1})$. We can now use Lemma V.D.7 to continue using

$$\frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[\text{Cov}[h_i | \mathcal{F}_{i-1}]] f''(\psi^*)^{-1}) \right)^{1/2}$$

$$\begin{aligned}
&\leq \frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[\text{Cov}[\phi(X_i, \theta_i) | \theta_i] f''(\psi^*)^{-1}) \right)^{1/2} \\
&+ \frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[\text{Cov}[g(X_i, \theta_i, \psi_{i-1}) | \mathcal{F}_{i-1}] f''(\psi^*)^{-1}) \right)^{1/2} \\
&+ \frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[e_2(\psi_{i-1}, m)] f''(\psi^*)^{-1}) \right)^{1/2} \quad (\text{V.28})
\end{aligned}$$

As $\mathbb{E}[\text{Cov}[\phi(X_t, \theta_t)] | \theta_t] = f''(\psi^*)$, the first term equals $\sqrt{\frac{\text{Tr} f''(\psi^*)^{-1}}{n}}$. The second term, which comes from the (unbiased) estimation of the gradient of the log-normalizer $\log Z(\theta_1, \psi)$ also equals $\sqrt{\frac{\text{Tr} f''(\psi^*)^{-1}}{n}}$, up to negligible terms. Indeed, by Lemma V.D.13, we have

$$\begin{aligned}
\left\| \text{Cov}[g(X_t, \theta_t, \psi_{t-1}) | \mathcal{F}_{t-1}] - \text{Cov}\left[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \mathcal{F}_{t-1}\right] \right\| &\leq 2p^2\tau^2 C_\chi \times \|\psi_{t-1} - \psi^*\|, \\
\forall t \in [n], \quad (\text{V.29})
\end{aligned}$$

from which we have:

$$\begin{aligned}
&\frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[\text{Cov}[g(X_i, \theta_i, \psi_{i-1}) | \mathcal{F}_{i-1}] f''(\psi^*)^{-1}) \right)^{1/2} \\
&\stackrel{(a)}{\leq} \sqrt{\frac{\text{Tr} f''(\psi^*)^{-1}}{n}} + \frac{\sqrt{2p^2\tau^2 C_\chi}}{n\mu} \left(\sum_{i=1}^n f''(\psi^*)^{-1} \sqrt{\mathbb{E}[\|\psi_{i-1} - \psi^*\|^2]} f''(\psi^*)^{-1} \right)^{1/2} \\
&\stackrel{(b)}{\leq} \sqrt{\frac{\text{Tr} f''(\psi^*)^{-1}}{n}} + \mathcal{O}(n^{-\frac{1}{2}-\frac{\beta}{2}}) \quad (\text{V.30})
\end{aligned}$$

where, in (a), we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, the cyclicity of the trace, the fact that $\|AB\|_{\text{F}} \leq \|A\|_{\text{op}} \|B\|_{\text{F}}$, the fact that $\|f(\psi^*)^{-2}\| \leq \mu^{-2}$, and finally Jensen's inequality, and in (b), we used the bound of Theorem V.D.1. It remains to bound the last term of Equation V.28. From the definition of $e_2(\psi_{t-1})$ and $e_3(\psi_{t-1})$, we have that

$$\frac{1}{n} \left(\sum_{i=1}^n \text{Tr} (f''(\psi^*)^{-1} \mathbb{E}[e_2(\psi_{i-1}, m)] f''(\psi^*)^{-1}) \right)^{1/2}$$

$$\begin{aligned} &\leq \frac{AF(4m)^{1/4}}{\sqrt{n}} + \frac{BF(4m)^{1/16}}{n} \left(\sum_{i=1}^n \mathbb{E} [\|\psi_{i-1} - \psi^*\|^2]^{1/4} \right)^{1/2} \\ &\quad + \frac{CF(4m)^{1/4}}{n} \left(\sum_{i=1}^n \mathbb{E} [\|\psi_{i-1} - \psi^*\|^2] \right)^{1/2} \end{aligned} \quad (\text{V.31})$$

for some constants A, B, C independent of n . Using the bound on $m(n)$, we have that the first term is $\mathcal{O}(n^{-1+\frac{\beta}{2}})$, the second term is of order $\mathcal{O}(n^{-\frac{5-\beta}{8}})$, and the last term is of order $\mathcal{O}(n^{-1})$. Inserting Equation V.30 with Equation V.31 into Equation V.28, we obtain

$$\left(\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n f''(\psi^*)^{-1} (\bar{h}_i - h_i) \right\|^2 \right] \right)^{1/2} \leq 2 \sqrt{\frac{\text{Tr} f''(\psi^*)^{-1}}{n}} + o(n^{-\frac{1}{2}})$$

with, the $o(n^{-\frac{1}{2}})$ being of order $\mathcal{O}\left(n^{-\max\left(-1+\frac{\beta}{2}, \frac{-5-\beta}{8}, -\frac{1}{2}-\frac{\beta}{4}, -1+\frac{\beta}{2}\right)}\right)$. Combining this bound with the bounds on (i), (ii) and (iii) concludes the proof. \square

V.D.2.3 Convergence of the non-averaged iterates

Theorem V.D.1. Fix $n \geq 1$. Let $(\psi_t)_{0 \leq t \leq n}$ be the iterates produced by the conditional CD algorithm, and define $\delta_t := \mathbb{E} \|\psi_t - \psi^*\|^2$. Moreover, assume that m verifies $\mu - \sigma C_\chi G(m)$, where μ and σ are defined in Lemma V.D.9, and $G(m)$ is defined in Lemma V.D.2. Then, under Assumptions A7, A8, A9 and any variant of A10, for $\eta_t = Ct^{-\beta}$ with $C > 0$, we have:

$$\delta_n \leq \begin{cases} 2 \exp(4\tilde{L}C^2 \varphi_{1-2\beta}(n)) \exp\left(-\frac{\tilde{\mu}_m C}{4} n^{1-\beta}\right) \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + \frac{4C\tilde{\sigma}_m^2}{\tilde{\mu}_m n^\beta}, & \text{if } 0 \leq \beta < 1 \\ \frac{\exp(2\tilde{L}^2 C^2)}{n^{\tilde{\mu}_m C}} \left(\delta_0 + \frac{\tilde{\sigma}_m^2}{\tilde{L}^2}\right) + 2\tilde{\sigma}_m^2 C^2 \frac{\varphi_{\tilde{\mu}_m C/2-1}(n)}{n^{\tilde{\mu}_m C/2}}, & \text{if } \beta = 1, \end{cases}$$

where $\tilde{\sigma}_m = \sigma^2 + \sigma_*^2 + G(m) \times \tau^2 (C_\chi(p+1) + 1)$, $\tilde{L} = \tau^2 C_\chi (2 + G(m)(p+1))$, and $\tilde{\mu}_m = \mu - G(m) \times C_\chi \sigma$. Consequently, if $\eta_n = \frac{C}{n}$ with an initial learning rate $C > 2\tilde{\mu}_m^{-1}$, we have $\sqrt{\delta_n} \leq 2\tilde{\sigma}_m C \sqrt{\frac{\tilde{\mu}_m C}{\tilde{\mu}_m C - 2}} \frac{1}{\sqrt{n}} + o\left(\frac{1}{n}\right)$.

The proof of this theorem is based on a recursion obtained in the following Lemma. As we will see, we will then unroll this recursion to obtain an inequality similar to the one of [247, Theorem 3.2].

Proof of Theorem V.D.1. By Lemma V.D.2, we have, for all $t \geq 1$,

$$\begin{aligned}\delta_t &\leq (1 - 2\eta_t(\mu - C_\chi \sigma G(m)) + \eta_t^2 \tau^2 C_\chi (2 + G(m)(p+1))) \delta_{t-1} \\ &\quad + \eta_t^2 (2\sigma^2 + \sigma_*^2 + C_\chi \tau^2 (G(m)(p+1) + 1)) . \quad (\text{V.32})\end{aligned}$$

Note that the recursion obtained above is of the same form as the one studied in [169, Equation 6, Theorem 1] given by:

$$\delta_t \leq (1 - 2\mu\gamma_t + 2L^2\gamma_t^2) \delta_{t-1} + 2\sigma^2\gamma_t^2$$

by identifying:

$$\begin{aligned}\sigma^2 &\leftarrow \sigma^2 + \sigma_*^2 + C_\chi \tau^2 (G(m)(p+1) + 1) =: \tilde{\sigma}_m^2 \\ L^2 &\leftarrow \tau^2 C_\chi (2 + G(m)(p+1)) =: \tilde{L}_m^2 \\ \mu &\leftarrow \mu - G(m) \times C_\chi \sigma =: \tilde{\mu}_m \\ \gamma_t &\leftarrow \eta_t\end{aligned}$$

The proof follows by noting that, from Lemma V.D.2, δ_t follows the same recursion as in [247], and can be unrolled in the same way, leading to the same result. Note that $\tilde{L}_m > \tilde{\mu}_m$ for all $m \geq 0$, since $\tilde{L}_m > \mu$ and $G(m) \geq 0$. \square

Lemma V.D.2. *Let $(\psi_t)_{0 \leq t \leq n}$ be the iterates from Algorithm 5. Denote $\delta_t = \mathbb{E}[\|\psi_t - \psi^*\|^2]$. Then, under Assumptions A7 to A9, and any variant of A10, for all $t \geq 1$, it holds that:*

$$\begin{aligned}\delta_t &\leq (1 - 2\eta_t(\mu - C_\chi \sigma G(m)) + \eta_t^2 \tau^2 C_\chi (2 + G(m)(p+1))) \delta_{t-1} \\ &\quad + \eta_t^2 (2\sigma^2 + \sigma_*^2 + C_\chi \tau^2 (G(m)(p+1) + 1)) . \quad (\text{V.33})\end{aligned}$$

where μ and σ are defined in Lemma V.D.9, $G(m) := \max \left\{ F(4m)^{1/4}, F(4m)^{1/2}F(4m)^{1/8} \right\}$, and $F(m)$ is defined in Lemma V.D.4.

Proof of Lemma V.D.2. By definition of ψ_t , we have:

$$\begin{aligned}\|\psi_t - \psi^*\|^2 &= \|\text{Proj}_\Psi(\psi_{t-1} - \eta_t h_t) - \psi^*\|^2 \\ &\stackrel{(a)}{\leq} \|\psi_{t-1} - \eta_t h_t - \psi^*\|^2 \\ &= \|\psi_{t-1} - \psi^*\|^2 - 2\eta_t \langle h_t, \psi_{t-1} - \psi^* \rangle + \eta_t^2 \|h_t\|^2 \\ &= \|\psi_{t-1} - \psi^*\|^2 - 2\eta_t \langle g_t, \psi_{t-1} - \psi^* \rangle + \eta_t^2 \|h_t\|^2 - 2\eta_t \langle h_t - g_t, \psi_{t-1} - \psi^* \rangle\end{aligned}$$

Implying

$$\begin{aligned}\mathbb{E} [\|\psi_t - \psi^*\|^2 | \mathcal{F}_{t-1}] &\leq \|\psi_{t-1} - \psi^*\|^2 - 2\eta_t \langle \mathbb{E}[g_t | \mathcal{F}_{t-1}], \psi_{t-1} - \psi^* \rangle + \eta_t^2 \mathbb{E} [\|h_t\|^2 | \mathcal{F}_{t-1}] \\ &\quad - 2\eta_t \langle \mathbb{E}[h_t - g_t | \mathcal{F}_{t-1}], \psi_{t-1} - \psi^* \rangle,\end{aligned}$$

Where (a) employed [247, Lemma C.8]. The first term is the previous iterate distance; the second term is the gradient descent term, as $\mathbb{E}[g(X_t, \theta_t, \psi)] = \nabla_\psi \bar{\mathcal{L}}(\psi)$ by definition. As $\bar{\mathcal{L}}$ is μ -strongly convex, we have:

$$\begin{aligned}\langle g(X_t, \theta_t, \psi), \psi - \psi^* \rangle &\geq \mu \times \|\psi - \psi^*\|^2 \\ \implies \langle \mathbb{E}[g_t | \mathcal{F}_{t-1}], \psi_{t-1} - \psi^* \rangle &\geq \mu \times \|\psi_{t-1} - \psi^*\|^2.\end{aligned}$$

The third term is the CD gradient noise term, which is controlled by Lemma V.D.6. The last term is a gradient bias term, which we can control using the approximation results developed above. In particular, we have:

$$\begin{aligned}\langle \mathbb{E}[h_t - g_t | \mathcal{F}_{t-1}], \psi_{t-1} - \psi^* \rangle &= \left\langle \mathbb{E} \left[P_{\psi_{t-1}, \theta_t}^m \phi(X_t, \theta_t) \middle| \mathcal{F}_{t-1} \right] - \mathbb{E} \left[\phi(X^{\psi_{t-1}, \theta_t}, \theta_t) \middle| \mathcal{F}_{t-1} \right], \psi_{t-1} - \psi^* \right\rangle \\ &\leq C_\chi \sigma F (4m)^{1/4} \|\psi_{t-1} - \psi^*\|^2,\end{aligned}$$

where in the last line, we used V.D.17 on each of the ϕ_i , as well as the Cauchy-Schwarz inequality. Combining everything, we obtain

$$\begin{aligned} \mathbb{E} \left[\|\psi_t - \psi^*\|^2 \mid \mathcal{F}_{t-1} \right] &\leq \left(1 - 2\eta_t(\mu - C_\chi \sigma F(4m)^{1/4}) \right) \|\psi_{t-1} - \psi^*\|^2 \\ &+ \eta_t^2 \left(\sigma^2 + \sigma_*^2 + \tau^2 C_\chi (2 + F(4m)^{\frac{1}{4}} p) \|\psi_{t-1} - \psi^*\| \right. \\ &\left. + \sigma \times \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi_{t-1} - \psi^*\| \right)^{1/2} \right). \end{aligned}$$

To simplify the expression further, we use the bound:

$$\begin{aligned} \sigma \times \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi_{t-1} - \psi^*\| \right)^{1/2} \\ \leq \sigma^2 + \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi_{t-1} - \psi^*\| \right) \quad (\text{V.34}) \end{aligned}$$

and note $G(m) := \max \left\{ F(4m)^{1/4}, F(4m)^{1/2} F(4m)^{1/8} \right\}$, thanks to which we now have:

$$\begin{aligned} \mathbb{E} \left[\|\psi_t - \psi^*\|^2 \mid \mathcal{F}_{t-1} \right] &\leq \left(1 - 2\eta_t(\mu - C_\chi \sigma F(4m)^{1/4}) \right) \|\psi_{t-1} - \psi^*\|^2 \\ &+ \eta_t^2 [2\sigma^2 + \sigma_*^2 + C_\chi \tau^2 (2 + G(m)(p+1)) \|\psi_{t-1} - \psi^*\| + G(m)\tau^2]. \end{aligned}$$

Finally, to remove the ‘‘cross-term’’, scaling with $\eta_t^2 \|\psi_{t-1} - \psi^*\|$, we use the bound $\eta_t^2 \|\psi_{t-1} - \psi^*\| \leq \eta_t^2 \|\psi_{t-1} - \psi_t\|^2 + \eta_t^2$. Rearranging, we now have:

$$\begin{aligned} \mathbb{E} \left[\|\psi_t - \psi^*\|^2 \mid \mathcal{F}_{t-1} \right] &\leq \left(1 - 2\eta_t(\mu - C_\chi \sigma G(m)) + \eta_t^2 \tau^2 C_\chi (2 + G(m)(p+1)) \right) \|\psi_{t-1} - \psi^*\|^2 \\ &+ \eta_t^2 (2\sigma^2 + \sigma_*^2 + C_\chi \tau^2 (G(m)(p+1) + 1)). \end{aligned}$$

and averaging over \mathcal{F}_{t-1} , we obtain the desired result , e.g

$$\begin{aligned} \delta_t &\leq \left(1 - 2\eta_t(\mu - C_\chi \sigma G(m)) + \eta_t^2 \tau^2 C_\chi (2 + G(m)(p+1)) \right) \delta_{t-1} \\ &+ \eta_t^2 (2\sigma^2 + \sigma_*^2 + C_\chi \tau^2 (G(m)(p+1) + 1)). \quad (\text{V.35}) \end{aligned}$$

□

Lemma V.D.3. *Under Assumptions A7, A8, A9, and any variant of A10, the online*

Conditional CD iterates obtained using $\eta_t = Ct^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ verify

$$\frac{1}{n} \sum_{i=1}^n \left(\mathbb{E} [\|\psi_i - \psi^*\|^4] \right)^{1/2} \leq P_1(n)$$

where $P_1(n)$ is a term in n of order $O(n^{-\beta})$.

Proof. We have:

$$\begin{aligned} \mathbb{E} [\|\psi_n - \psi^*\|^4 | \mathcal{F}_{n-1}] &\leq \|\psi_{n-1} - \psi^*\|^4 + 6\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 \mathbb{E} [\|h_n\|^2 | \mathcal{F}_{n-1}] \\ &\quad + \eta_n^4 \mathbb{E} [\|h_n\|^4 | \mathcal{F}_{n-1}] \\ &- 4\eta_n \|\psi_{n-1} - \psi^*\|^2 \langle \psi_{n-1} - \psi^*, \mathbb{E}[h_n] \rangle + 4\eta_n^3 \|\psi_{n-1} - \psi^*\| \mathbb{E} [\|h_n(\psi_{n-1})\|^3 | \mathcal{F}_{n-1}]. \end{aligned} \tag{V.36}$$

We first need to control the higher-order moments of $\|h_n\|$, as only the second one was controlled above. We use the following cruder bound:

$$\begin{aligned} \mathbb{E} [\|h_n\|^k] &= \mathbb{E} \left[\left\| \phi(X_n, \theta_n) + \phi(K_{\psi_{n-1}, \theta_n, n}^m(X_n), \theta_n) \right\|^k \right] \\ &\leq 2^{k-1} \left(\mathbb{E} [\|\phi(X_n, \theta_n)\|^k] + \mathbb{E} \left[\left\| \phi(K_{\psi_{n-1}, \theta_n, n}^m(X_n), \theta_n) \right\|^k \right] \right) \\ &\leq 2^k \tau^k (1 + C_\chi \times \|\psi_{n-1} - \psi^*\|), \end{aligned}$$

where, in the last line, we used Lemma V.D.17. Plugging this into the previous equation, we obtain

$$\begin{aligned} \mathbb{E} [\|\psi_n - \psi^*\|^4 | \mathcal{F}_{n-1}] &\leq \|\psi_{n-1} - \psi^*\|^4 \\ &\quad + 6\eta_n^2 \|\psi_{n-1} - \psi^*\|^2 (4\tau^2 + 4C_\chi \tau^2 \|\psi_{n-1} - \psi^*\|) \\ &\quad + \eta_n^4 (16\tau^4 + 16C_\chi \tau^4 \|\psi_{n-1} - \psi^*\|) \\ &\quad - 4\eta_n \|\psi_{n-1} - \psi^*\|^2 \langle \psi_{n-1} - \psi^*, \mathbb{E}[h_n | \mathcal{F}_{n-1}] \rangle \\ &\quad + 4\eta_n^3 \|\psi_{n-1} - \psi^*\| (8\tau^3 + 8C_\chi \tau^3 \|\psi_{n-1} - \psi^*\|) \end{aligned}$$

which has precisely the same form as the recursion of [247, Lemma D.8], and can thus be controlled in the same way; the result follows from applying their arguments.

In particular, we have:

$$\begin{aligned}
& \frac{1}{n} \sqrt{\sum_{i=1}^n \mathbb{E} \|\psi_{i-1} - \psi^*\|^4} \\
& \leq \frac{C\tilde{\tau}_1^2}{2n} \left(C^{1/2} \varphi_{1-3\beta/2}(n) + \tilde{\mu}_m^{-1/2} \varphi_{1-\beta}(n) \right) \\
& \quad + \frac{\sqrt{20}C^{1/2}\tilde{\tau}_1}{2n} A_1 \exp(24\tilde{L}_1^4 C^4) \left(\delta_0 + \frac{\tilde{\mu}_m \mathbb{E} \|\psi_0 - \psi^*\|^4}{20C\tilde{\tau}_1^2} + 2\tilde{\tau}_1^2 C^3 \tilde{\mu}_m + 8\tilde{\tau}_1^2 C^2 \right)^{1/2} \\
& = \mathcal{O}(n^{-\beta}),
\end{aligned}$$

where

$$A_1 = \sum_{k=1}^n e^{\frac{-\tilde{\mu}_m C k^{1-\beta}}{16} + 16\tilde{L}_1^4 C^4 \varphi_{1-2\beta}(k)}$$

where we defined $\tilde{\tau}_1 := 2(1+C_\chi)\tau$ and $\tilde{L}_1 := 2(1+C_\chi)\tau$. We have $A(1) < +\infty$ if $\beta < 1$, and $A(1) = O(n)$ otherwise. \square

V.D.2.4 Bias and Variance of the conditional CD gradient

Before proceeding, we show a result regarding the behavior of averaged exponents of the spectral gaps $\alpha(\theta_t, \psi_t)$.

Lemma V.D.4. *Under Assumption A7, the functions $\alpha_t^m(\theta_t, \psi_t)$ are $(\mathcal{F}, \mathcal{B}(\mathbb{R}_+))$ -measurable for all $t, m \in \mathbb{N}$. Moreover, we have, for all $m \geq 1$,*

$$\mathbb{E}[\alpha^m(\theta_1, \psi)] \leq F(m)$$

for $F(m)$ given by

$$F(m) := \begin{cases} \frac{1}{m+1} + \frac{\vartheta_\kappa^\kappa}{(\log(m+1) - \log(\log(m+1)))^\kappa} = O\left(\left(\frac{1}{\log m}\right)^\kappa\right) & \text{if A10 holds} \\ \left(\frac{1}{1+m\lambda}\right)^{\frac{1}{\lambda}} + \frac{\vartheta_1}{\sqrt{e}} \left(\frac{e^{\vartheta_1} \log(1+m\lambda)}{m\lambda}\right)^{1/\lambda} = O\left(\left(\frac{\log m}{m}\right)^{\frac{1}{\lambda}}\right) & \text{if A10' holds} \\ J(\vartheta_1, \sigma^2, m) = O\left(\frac{\text{polylog}(m)}{m}\right)^{\frac{\log m + o(\log m)}{\sigma^2}} & \text{if A10'' holds} \\ (\vartheta_{\max})^m & \text{if A10''' holds.} \end{cases}$$

Here, $\text{polylog}(m)$ denotes an arbitrary polynomial of $\log m$, and the explicit

form of $J(\vartheta, \sigma^2, m)$ is given in Equation V.38.

Proof. Let $\mathcal{F} := \{\phi_i\}_{i=1}^n \cup \{\phi_i \phi_j\}_{i,j=1}^n$ and consider the functions

$$\begin{aligned}\tilde{\alpha}_f : (\psi, \theta) &\longmapsto \frac{\left(\int \left(\int (f(x, \theta) - \int f(z, \theta) p_\psi(dz|\theta))(y) k_{\psi, \theta}(x, dy) \right)^2 p_\psi(dx|\theta) \right)^{1/2}}{\left(\int (f(x, \theta) - \int f(z, \theta) p_\psi(dz|\theta))(y) (x)^2 p_{\psi, \theta}(dx) \right)^{1/2}} \\ &= \frac{\beta(f, \theta, \psi)}{\gamma(f, \theta, \psi)}, \quad f \in \mathcal{F} \\ \tilde{\alpha} : (\psi, \theta) &\longmapsto \max_{f \in \mathcal{F}} \tilde{\alpha}_f(\psi, \theta)\end{aligned}$$

with the convention $\frac{0}{0} = 0$. Note that that α_t as defined in (V.37) can be written as

$$\alpha_t : \omega \longmapsto \tilde{\alpha} \circ (\theta_t(\omega), \psi_t(\omega)), \quad t \in \mathbb{N} \quad (\text{V.37})$$

We will show that that $\tilde{\alpha}$ is $\mathcal{T} \times \blacksquare$ -measurable, from which it will follow that α_t is \mathcal{T} -measurable. Note first that both $\beta(f, \theta, \psi)$ and $\gamma(f, \theta, \psi)$ are composition of continuous maps and measure-kernel-functions applications [43, Proposition 6.9] where the functions are measurable w.r.t their respective product spaces, implying that for each f , $\beta(f, \bullet, \diamond)$ and $\gamma(f, \bullet, \diamond)$ are $\blacksquare \times \mathcal{T}$ -measurable. Moreover, consider the sequence:

$$\alpha_n(f, \psi, \theta) = \frac{\beta(\theta, \psi)}{\gamma(\theta, \psi) + \frac{1}{n}}$$

α_n is $\mathcal{T} \times \blacksquare$ -measurable as the denominator is now always positive. Moreover, $\lim_{n \rightarrow \infty} \alpha_n(\theta, \psi) = \alpha(\theta, \psi)$: indeed, if $\beta(\theta, \psi) > 0$, this is immediate. Otherwise, then f must be constant $p_{\psi, \theta}$ almost everywhere, implying we also have $\gamma(\theta, \psi)$, implying $\alpha_n(\theta, \psi) = \alpha(\theta, \psi) = 0$. Consequently, $\tilde{\alpha}_f(\psi, \theta)$ is measurable, as the pointwise-limit of measurable functions [43, Theorem 2.15], and so is $\alpha(\psi, \theta)$ as the maximum of measurable functions is measurable. Finally, as α_t is thus the composition of a $(\mathcal{T} \times \blacksquare, \mathcal{B}(\mathbb{R}_+))$ with the map $(\theta_t(\omega), \psi_t(\omega))$ which is $(\mathcal{F}, \mathcal{T} \times \blacksquare)$ -measurable, and is thus $(\mathcal{F}, \mathcal{B}(\mathbb{R}_+))$ -measurable. We thus have that $(\alpha_t)_{t \in \mathbb{N}}$ define valid random variables, and so do α_t^m .

Let $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$ the extended half-line [132, p. 5] endowed with its standard

order topology, and denote $\mathcal{B}(\overline{\mathbb{R}}_+)$ its associated Borel σ -algebra. Let ϑ_t the function defined

$$\vartheta(\psi, \theta_1) := -\log(1 - \alpha(\psi, \theta_1))$$

Note that ϑ_t is $(\mathcal{F}, \mathcal{B}(\overline{\mathbb{R}}_+))$ -measurable, as $\vartheta_t^{-1}\{+\infty\} = \alpha_t^{-1}\{0\}$, which has measure 0 (and is thus measurable). Consider now the quantity

$$\mathbb{E}[\alpha^m(\theta_1, \psi)]$$

and let $\delta \in (0, 1)$. We have:

$$\begin{aligned} & \mathbb{E}[\alpha^m(\theta_1, \psi)] \\ &= \mathbb{E}[\alpha^m(\theta_1, \psi) | \alpha(\theta_1, \psi) < 1 - \delta] \mathbb{P}[\alpha(\theta_1, \psi) < 1 - \delta] \\ &\quad + \mathbb{E}[\alpha^m(\theta_1, \psi) | \alpha(\theta_1, \psi) > 1 - \delta] \mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] \\ &\leq (1 - \delta)^m + \mathbb{E}[\alpha^m(\theta_1, \psi) | \alpha(\theta_1, \psi) > 1 - \delta] \mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] \\ &\leq (1 - \delta)^m + \mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] \end{aligned}$$

Case A10 We have

$$\begin{aligned} \mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] &= \mathbb{P}[(-\log(1 - \alpha(\theta_1, \psi)))^\kappa > (-\log \delta)^\kappa] \leq \frac{\mathbb{E}[\vartheta(\theta_1, \psi)^\kappa]}{(-\log \delta)^\kappa} \\ &= \frac{\vartheta_\kappa^\kappa}{(-\log \delta)^\kappa} \end{aligned}$$

where we used Markov's inequality as well as A10 in the last step. Alltogether, we thus have:

$$\mathbb{E}[\alpha^m(\theta_1, \psi)] \leq (1 - \delta)^m - \frac{\vartheta_\kappa^\kappa}{(-\log \delta)^\kappa}, \quad \forall \delta \in (0, 1).$$

And setting $\delta = \frac{\log(1+m)}{m} \in (0, 1)$ for all $m \geq 1$, we have:

$$\begin{aligned}\mathbb{E}[\alpha^m(\theta_1, \psi)] &\leq e^{m\log(1-\frac{\log(1+m)}{m})} + \frac{\vartheta_\kappa^\kappa}{(\log(1+m) - (\log\log(1+m)))^\kappa} \\ &\stackrel{(b)}{\leq} \frac{1}{m+1} + \frac{\vartheta_\kappa^\kappa}{(\log(m+1) - \log(\log(m+1)))^\kappa}\end{aligned}$$

where in (b), we used that $\log(1+x) \leq x$ for all $x > -1$.

Case A10' If instead, we assume A10', then by definition of sub-exponentiality, we have that, for all $s \leq 1/\lambda$,

$$\mathbb{E}\left[e^{s(\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)])}\right] = \mathbb{E}\left[\frac{1}{(1 - \alpha(\theta_1, \psi))^s}\right] \leq e^{\lambda^2 s^2/2}$$

implying that, using Markov's inequality on $e^{s\vartheta(\theta_1, \psi)}$, that:

$$\mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] = \mathbb{P}[(1 - \alpha(\theta_1, \psi))^{-s} > \delta^{-s}] \leq \delta^s e^{s\vartheta_1} e^{-\lambda^2 s^2/2}.$$

Setting $\delta = \frac{\log(1+m\lambda)}{m\lambda} \in (0, 1)$ and $s = \frac{1}{\lambda}$, we obtain

$$\mathbb{E}[\alpha^m(\theta_1, \psi)] \leq \left(\frac{1}{1+m\lambda}\right)^{\frac{1}{\lambda}} + \frac{\vartheta_1}{\sqrt{e}} \left(\frac{e^{\vartheta_1} \log(1+m\lambda)}{m\lambda}\right)^{1/\lambda}$$

Case A10'' Finally, under A10'', it holds, by properties of sub-Gaussian variables, that:

$$\mathbb{P}[\alpha(\theta_1, \psi) > 1 - \delta] = \mathbb{P}[\vartheta(\theta_1, \psi) - \mathbb{E}[\vartheta(\theta_1, \psi)] > -\log \delta - \mathbb{E}[\vartheta(\theta_1, \psi)]] \leq e^{-\frac{(\log \delta + \vartheta_1)^2}{2\sigma^2}}$$

Using Lemma V.D.16, we can set:

$$\delta := \frac{\left(\log\left(\frac{\sigma^2 m}{2} + 1\right)\right)^2 - 2\log\left(\frac{\sigma^2 m}{2} + 1\right)}{\sigma^2 m} \in (0, 1)$$

Implying we have

$$\begin{aligned}
\mathbb{E}[\alpha^m(\theta_1, \psi)] &\leq \left(\frac{\sigma^2 m}{2} + 1 \right)^{\frac{2 - \log\left(\frac{\sigma^2 m}{2} + 1\right)}{\sigma^2}} \\
&+ \left(e^{\vartheta_1} \frac{\left(\log\left(\frac{\sigma^2 m}{2} + 1\right)\right)^2 - 2\log\left(\frac{\sigma^2 m}{2} + 1\right)}{\sigma^2 m} \right)^{\frac{\log\left(\frac{\sigma^2 m}{2} + 1\right) - \vartheta_1 - \log\left(\left(\log\left(\frac{\sigma^2 m}{2} + 1\right)\right)^2 - 2\log\left(\frac{\sigma^2 m}{2} + 1\right)\right)}{\sigma^2}} \\
&:= J(\vartheta_1, \sigma^2, m)
\end{aligned} \tag{V.38}$$

The result for A10''' is immediate. \square

Lemma V.D.4 shows a clear separation of rates, depending of the tail behavior of a variable quantifying the concentration of spectral gaps around 1: Provided that a finite bound exists on the moment of $\vartheta(\alpha_1, \psi)$ across all ψ , then $\mathbb{E}[\alpha^m(\theta_1, \psi)]$ will decrease at a polylogarithmic scale. If instead, $\vartheta(\theta_1, \alpha)$ is sub-exponential, the rate will be polynomial in m (up to polylog), with an exponent depending on the sub-exponentiality parameter. Finally, if $\vartheta(\theta_1, \alpha)$ is sub-Gaussian, the rate ($\mathcal{O}(m^{-\frac{\log m}{\sigma^2}})$) is super-polynomial in m , with the exponent again depending on sub-Gaussianity parameter. To assess the tightness of these bounds, consider the case where $\alpha(\theta_1, m) \sim \mathcal{U}(0, 1)$. In that case, $\vartheta(\alpha_1, \theta) \sim \text{Exp}(1)$, meaning that A10' holds, with $\lambda = 1$. Moreover, explicit integral calculations show that $\mathbb{E}[\alpha(\theta_1, \psi)^m] \sim \frac{1}{m+1}$. Comparing with our bound, we see that the rate is tight up to a logarithmic factor, which is likely due to our Chernoff-type technique, which itself is known to be tight up to logarithmic factors [262, p.22].

Lemma V.D.5. Denote $h_i(\psi, X_t, \theta_t)$ be the i -th component of $h_t(\psi, X_t, \theta_t)$, and similarly for $g_i(X_t, \theta_t, \psi)$, for all $i \in [p]$. Then for all $i, j \in [p] \times [p]$, we have:

$$\begin{aligned}
\mathbb{E}[(h_i h_j)(X_t, \theta_t, \psi)] &= \mathbb{E}[(g_i g_j)(X_t, \theta_t, \psi)] + \mathbb{E}\left[\text{Cov}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t), \phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] \\
&+ e_{ij}(\psi, m),
\end{aligned} \tag{V.39}$$

where

$$|e_{ij}(\psi, m)| \leq (F(4m))^{\frac{1}{4}} C_\chi \tau^2 \|\psi - \psi^*\| + 2\tau \times \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi - \psi^*\| \right)^{1/2}.$$

Proof. We have:

$$\mathbb{E}[(h_i h_j)(X_t, \theta_t, \psi)] = (\mathbb{E}[(h_i h_j)(X_t, \theta_t, \psi)] - \mathbb{E}[(g_i g_j)(X_t, \theta_t, \psi)]) + \mathbb{E}[(g_i g_j)(X_t, \theta_t, \psi)]$$

Here, the second term is controlled by Lemma V.D.12. The first term is a residual term, which we now bound. We have:

$$\begin{aligned} & \mathbb{E}[(h_i h_j)(X_t, \theta_t, \psi)] - \mathbb{E}[(g_i g_j)(X_t, \theta_t, \psi)] = \\ & \mathbb{E}\left[(\phi_i \phi_j)(K_{\psi, \theta_t, t}^m(X_t), \theta_t)\right] - \mathbb{E}\left[\mathbb{E}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t\right] \times \mathbb{E}\left[\phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] \\ & - \mathbb{E}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \left(\phi_j(K_{\psi, \theta_t, t}^m(X_t), \theta_t) - \mathbb{E}\left[\phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right)\right] \\ & - \mathbb{E}\left[\phi_j(X^{\psi, \theta_t}, \theta_t) \times \left(\phi_i(K_{\psi, \theta_t, t}^m(X_t), \theta_t) - \mathbb{E}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right)\right] \quad (\text{V.40}) \end{aligned}$$

Term 1 For the first term, we first relate $\mathbb{E}[(\phi_i \phi_j)(K_{\psi, \theta_t, t}^m(X_t), \theta_t)]$ to $\mathbb{E}[(\phi_i \phi_j)(X^{\psi, \theta_t}, \theta_t)]$ using Lemma V.D.17:

$$\begin{aligned} & \left| \mathbb{E}\left[(\phi_i \phi_j)(K_{\psi, \theta_t, t}^m(X_t), \theta_t)\right] - \mathbb{E}\left[(\phi_i \phi_j)(X^{\psi, \theta_t}, \theta_t)\right] \right| \\ &= \left| \mathbb{E}\left[P_{\psi, \theta_t}^m(\phi_i \phi_j)(X^{\psi, \theta_t}, \theta_t)\right] - \mathbb{E}\left[(\phi_i \phi_j)(X^{\psi, \theta_t}, \theta_t)\right] \right| \\ &\stackrel{(a)}{\leq} (F(4m))^{\frac{1}{4}} C_\chi \|\psi - \psi^*\| \left(\mathbb{E}\left[\mu_2\left((\phi_i \phi_j)(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right)^2\right] \right)^{1/4} \\ &\leq (F(4m))^{\frac{1}{4}} C_\chi \|\psi - \psi^*\| \end{aligned}$$

where in (a), we used Lemma V.D.17 Consequently,

$$\begin{aligned} & \mathbb{E}\left[(\phi_i \phi_j)(K_{\psi, \theta_t, t}^m(X_t), \theta_t)\right] - \mathbb{E}\left[\mathbb{E}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] \times \mathbb{E}\left[\phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[(\phi_i \phi_j)(K_{\psi, \theta_t, t}^m(X_t), \theta_t) \mid \theta_t\right] - \mathbb{E}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t\right] \times \mathbb{E}\left[\phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] \\ &\quad + e_1(\psi, m) \\ &= \mathbb{E}\left[\text{Cov}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t), \phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] + e_1(\psi, m) \end{aligned}$$

where

$$|e_1(\psi, m)| \leq (F(4m))^{\frac{1}{4}} C_\chi \tau \times \|\psi - \psi^*\|$$

We now proceed to bound the error present in the last two terms.

Term 2 and 3 For the second term, we have:

$$\begin{aligned} & \left| \mathbb{E} \left[\phi_i(X_t^{\psi, \theta_t}, \theta_t) \left(\phi_j(K_{\psi, \theta_t, t}^m(X_t), \theta_t) - \mathbb{E} \left[\phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t \right] \right) \right] \right| \\ &= \left| \mathbb{E} \left[\phi_i(X_t^{\psi, \theta_t}, \theta_t), P_{\psi, \theta_t}^m \bar{\phi}_j(X_t, \theta_t) \right] \right| \\ &\leq \mathbb{E} \left[\phi_i(X_t^{\psi, \theta_t}, \theta_t)^2 \right]^{1/2} \mathbb{E} \left[P_{\psi, \theta_t}^m \bar{\phi}_i(X_t, \theta_t) \right]^{1/2} \\ &\leq \tau \times \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi - \psi^*\| \right)^{1/2} \end{aligned}$$

where the last line used Lemma V.D.18. The third term is bounded in the same way, and combining the three terms concludes the proof. \square

Lemma V.D.6. *We have:*

$$\mathbb{E} \left[\|h_t(\psi, X_t, \theta_t)\|^2 \right] = \mathbb{E} \left[\|g(\psi, X_t, \theta_t)\|^2 \right] + \sigma^2 + e(\psi, m)$$

where $e(\psi, m) \leq p \max_{i \in [p]} e_{ii}(\psi, m)$ and $e_{ii}(\psi, m)$ is defined as in Lemma V.D.5.

Proof.

$$\begin{aligned} & \mathbb{E} \left[\|h_t(\psi, X_t, \theta_t)\|^2 \right] \\ &= \sum_{i=1}^p \mathbb{E} \left[(h_t(\psi, X_t, \theta_t))_i^2 \mid \psi \right] \\ &\stackrel{(a)}{=} \sum_{i=1}^p \left(\mathbb{E} \left[(g(\psi, X_t, \theta_t))_i^2 \right] + \mathbb{E} \left[\text{Cov} \left[\phi_i(X_t^{\psi, \theta_t}, \theta_t), \phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t \right] \right] \right. \\ &\quad \left. + e_{ii}(\psi, m) \right) \\ &\stackrel{(b)}{=} \mathbb{E} \left[\|h_t(\psi, X_t, \theta_t)\|^2 \right] + \sigma^2 + p \times e_{ii}(\psi, m) \end{aligned}$$

where in (a), we used Lemma V.D.5, and in (b) we used the definition of σ . \square

Lemma V.D.7. *We have that*

$$\text{Cov}[h_t(\psi, X_t, \theta_t)] = \text{Cov}[g(\psi, X_t, \theta_t)] + \mathbb{E} \left[\text{Cov} \left[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t \right] \right] + e_2(\psi, m)$$

where

$$\begin{aligned} \|e_2\|_{\text{F}} &\leq 2p^2\tau^2 F(4m)^{1/4} + 2p^2\tau^2 C_{\chi}^{1/2} F(4m)^{1/16} \|\psi - \psi^*\|^{1/2} \\ &\quad + p^2 F(4m)^{1/4} C_{\chi} \tau^2 \|\psi - \psi^*\| \\ &\quad + C_{\chi}^2 \tau^2 (2F(4m)^{1/4} + F(4m)^{1/2}) \|\psi - \psi^*\|^2 \quad (\text{V.41}) \end{aligned}$$

Proof. We have:

$$\text{Cov}[h_t(\psi, X_t, \theta_t)] = \mathbb{E}\left[h_t h_t^\top(\psi, X_t, \theta_t)\right] + \mathbb{E}[h_t(\psi, X_t, \theta_t)] \mathbb{E}[h_t(\psi, X_t, \theta_t)]^\top$$

with the first term such that

$$\begin{aligned} &\left(\mathbb{E}\left[h_t h_t^\top(\psi, X_t, \theta_t)\right]\right)_{ij} \\ &= \mathbb{E}\left[(h_i h_j)(X_t, \theta_t, \psi)\right] \\ &\stackrel{(a)}{=} \left(\mathbb{E}\left[(g_i g_j)(X_t, \theta_t, \psi)\right] + \mathbb{E}\left[\text{Cov}\left[\phi_i(X_t^{\psi, \theta_t}, \theta_t), \phi_j(X^{\psi, \theta_t}, \theta_t) \mid \theta_t\right]\right] + e_{ij}(\psi, m)\right) \end{aligned}$$

where in (a), we used Lemma V.D.5. Next, we have

$$\begin{aligned} &\mathbb{E}[h_t(\psi, X_t, \theta_t)] \\ &= \mathbb{E}\left[\phi(X_t^{\psi, \theta_t}, \theta_t)\right] + \mathbb{E}\left[\phi(K_{\psi, \theta_t, t}^m(X_t), \theta_t)\right] \\ &= \mathbb{E}[g(\psi, X_t, \theta_t)] + \underbrace{\left(\mathbb{E}\left[P_{\psi, \theta_t}^m \phi(X_t^{\psi, \theta_t}, \theta_t)\right] - \mathbb{E}\left[\phi(K_{\psi, \theta_t, t}^m(X_t), \theta_t)\right]\right)}_{\Delta} \end{aligned}$$

Implying

$$\begin{aligned} \mathbb{E}[h_t(\psi, X_t, \theta_t)] \mathbb{E}[h_t(\psi, X_t, \theta_t)]^\top &= \mathbb{E}[g(\psi, X_t, \theta_t)] \mathbb{E}[g(\psi, X_t, \theta_t)]^\top \\ &\quad + 2\mathbb{E}[g(\psi, X_t, \theta_t)] \Delta^\top + \Delta \Delta^\top \end{aligned}$$

Putting the two equalities together, we obtain

$$\text{Cov}[h_t(\psi, X_t, \theta_t)] = \text{Cov}[g(\psi, X_t, \theta_t)] + \mathbb{E}\left[\text{Cov}\left[\phi(X_t^{\psi, \theta_t}, \theta_t)\right] \mid \theta_t\right] + e_2(\psi_{t-1}, m)$$

where

$$(e_2(\psi_{t-1}, m))_{ij} := e_{ij}(\psi_{t-1}, m) + 2\mathbb{E}[g_t | \mathcal{F}_{t-1}]_i \Delta_j + \Delta_i \Delta_j$$

implying

$$\begin{aligned} \|e_2\|_{\mathbb{F}} &\leq \|(e_{ij})_{i,j \in [p]}\|_{\mathbb{F}} + 2\|\mathbb{E}[g(\psi, X_t, \theta_t)]\| \|\Delta\| + \|\Delta\|_{\mathbb{F}}^2 \\ &\leq p^2 \times \max_{i,j \in [p]} |e_{ij}(\psi, m)| + 2C_\chi \tau \|\psi - \psi^*\| \|\Delta\| + \|\Delta\|^2 \\ &\leq p^2 \times \left((F(4m))^{\frac{1}{4}} C_\chi \tau^2 \|\psi - \psi^*\| \right. \\ &\quad \left. + 2\tau \times \left(F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi - \psi^*\| \right)^{1/2} \right) \\ &\quad + 2C_\chi \tau \times \|\psi - \psi^*\| \times \tau C_\chi \times \|\psi - \psi^*\| \times F(4m)^{1/4} \\ &\quad + (\tau C_\chi \times \|\psi - \psi^*\| \times F(4m)^{1/4})^2 \\ &\leq 2p^2 \tau^2 F(4m)^{1/4} + 2p^2 \tau^2 C_\chi^{1/2} F(4m)^{1/16} \|\psi - \psi^*\|^{1/2} \\ &\quad + p^2 F(4m)^{1/4} C_\chi \tau^2 \|\psi - \psi^*\| \\ &\quad + C_\chi^2 \tau^2 (2F(4m)^{1/4} + F(4m)^{1/2}) \|\psi - \psi^*\|^2 \end{aligned}$$

□

V.D.2.5 Properties of the loss function and its stochastic gradient

The main benefit of using (conditional) exponential families is that the cross entropy between $p_{\psi, \pi}$ and p_π^* is, not only differentiable, but also convex, thanks to which gradient-based optimization algorithms will be able to find the an approximate minimizer of $\bar{\mathcal{L}}$ efficiently. To show that, we first extend the smoothness and convexity properties of the log-partition function of exponential families to the conditional setting, establishing the same properties for the *average* log-partition function $\mathbb{E} \log Z(\theta_1, \psi)$.

Lemma V.D.8. *Under Assumption A7, A8, A9, the function $\bar{\mathcal{L}}(\psi)$ defined in Equation V.20 has a convex domain, and is four times differentiable on $\text{int}(\text{Dom}(\bar{\mathcal{L}}))$, with*

gradient and higher order derivatives given by

$$\begin{aligned}\nabla_{\psi} \bar{\mathcal{L}}(\psi) &= -\mathbb{E}[\phi(X_1, \theta_1)] + \mathbb{E}[\nabla_{\psi} \log Z(\theta_1, \psi)] \\ D^k \bar{\mathcal{L}}(\psi) &= \mathbb{E}[D^k \log Z(\theta_1, \psi)], \quad 2 \leq k \leq 4.\end{aligned}$$

Proof. We start from the form of $\bar{\mathcal{L}}$ given in Equation (V.20). Both terms are convex combinations of convex functions (the first integrand is linear, the second is the log of the partition function of an exponential family, and thus convex [263, Proposition 3.1]), and thus have convex domain. Thus, their sum has a convex domain. To derive the gradient of $\bar{\mathcal{L}}$, we show the differentiability of the two terms of the loss function separately. We have:

$$\mathbb{E}[\langle \psi, \phi(X_t, \theta_t) \rangle] = \mathbb{E}\left[\sum_{i=1}^d \psi_i \phi_i(X_t, \theta_t)\right] = \sum_{i=1}^d \psi_i \mathbb{E}[\phi_i(X_t, \theta_t)] = \langle \psi, \mathbb{E}[\phi(X_t, \theta_t)] \rangle$$

and $\mathbb{E}[\phi(X_t, \theta_t)] \in \mathbb{R}^p$ by Assumption A7. We thus have $\nabla_{\psi} \mathbb{E}[\langle \psi, \phi(X_t, \theta_t) \rangle] = \mathbb{E}[\phi(X_t, \theta_t)]$. Let us now show the first differentiability of $\mathbb{E}[\log Z(\theta_1, \psi)]$. The proof for the second differentiability is similar. Let $\psi_0 \in \Psi$, and let $(\psi_n)_{n \geq 0}$ a sequence converging to ψ_0 . Consider

$$d(\psi, \theta) := \frac{\log Z(\theta, \psi_n) - \log Z(\theta, \psi_0) - \langle \psi_n - \psi_0, \nabla_{\psi} \log Z(\theta, \psi_0) \rangle}{\|\psi_n - \psi_0\|}.$$

As the function $\log Z(\theta, \psi_n)$ is differentiable on its domain [32, Theorem 2.2] and thus on $\text{dom}(\bar{\mathcal{L}}) \subset \mathcal{D} \subset \mathcal{D}_{\theta}$, and we have that $\lim_{n \rightarrow \infty} d(\theta, \psi_n) = 0$. Moreover, we have:

$$\begin{aligned}\frac{\|\log Z(\theta, \psi_n) - \log Z(\theta, \psi_0)\|}{\|\psi_n - \psi_0\|} &\leq \max_{\psi \in \Psi} \mathbb{E}\left[\left\|\phi(X_t^{\psi, \theta}, \theta)\right\| \middle| \theta\right] \\ &\leq \mathbb{E}\left[\left\|\phi(X_t^{\psi_0, \theta}, \theta)\right\| \middle| \theta\right] + \mathbb{E}\left[\left\|\phi(X_t^{\psi_0, \theta}, \theta)\right\|^2 \middle| \theta\right]^{1/2} \times \chi_2(p_{\psi_0}(\cdot, \theta), p_{\psi}) \\ &\leq \mathbb{E}\left[\left\|\phi(X_t^{\psi_0, \theta}, \theta)\right\| \middle| \theta\right] + \mathbb{E}\left[\left\|\phi(X_t^{\psi_0, \theta}, \theta)\right\|^2 \middle| \theta\right]^{1/2} \times C_{\chi}(\theta) \times r_{\Psi} \\ &=: g(\theta)\end{aligned}$$

where we noted $\bar{\psi}(\theta) = \arg \max_{\psi \in \Psi} \mathbb{E}[\|\phi(X_t^{\psi_0, \theta} | \theta)\|]$ and $r_\Psi := \sup_{\psi, \psi' \in \Psi} \|\psi - \psi'\|$. Moreover, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[g(\theta_t)] &\leq \mathbb{E}\left\|\phi(X_t^{\psi_0, \theta_t}, \theta_t)\right\| + \mathbb{E}\left[\left\|\phi(X_t^{\psi_0, \theta_t}, \theta_t)\right\|^2\right]^{1/2} \times (\mathbb{E}[C_\chi(\theta)^2])^{1/2} \\ &< p \times \tau + p \times \tau \times C_\chi r_\Psi\end{aligned}$$

under assumption A8, A9. Similarly, $\frac{\langle \psi_n - \psi_0, \nabla \log Z(\theta, \psi_0) \rangle}{\|\psi_n - \psi_0\|}$ is also upper bounded by a π -integrable function. Thus, by the dominated convergence theorem, we have that $\mathbb{E} \log Z(\theta_t, \psi)$ is differentiable, and thus continuous. Combining both results finishes the proof. This argument can be repeated for the derivatives of order up to 4, as beyond that, the required moment on $\left\|\phi(X_t^{\psi_0, \theta}, \theta)\right\|$ may be infinite. \square

Combining Lemma V.D.8 with known formulas for the derivatives of $\log Z(\theta, \psi)$, we thus have that, under assumption A7, the gradient and hessian of $\bar{\mathcal{L}}$ are given by:

$$\nabla_\psi \bar{\mathcal{L}}(\psi) = -\mathbb{E}[\phi(X_t, \theta_t)] + \mathbb{E}\left[\phi\left(X_t^{\psi, \theta_t}, \theta_t\right)\right] \quad \nabla_\psi^2 \bar{\mathcal{L}}(\psi) = \mathbb{E}\left[\text{Cov}\left[\phi\left(X_t^{\psi, \theta_t}, \theta_t\right) \mid \theta_t\right]\right],$$

On the Fisher Information Matrix of p_ψ Previously, we saw that $\bar{\mathcal{L}}$ could be understood as a Maximum Likelihood objective for the joint model $p_{\psi, \pi}$. In this context, the Fisher Information Matrix $\mathcal{I}(\psi^*)$ of $p_{\psi, \pi}$ at ψ^* is given by:

$$\begin{aligned}\mathcal{I}(\psi^*) &:= \text{Cov}\left[\nabla_\psi \log \frac{dp_{\psi, \pi}}{d(\pi \times \mu)}(x, \theta)\right] \\ &= \text{Cov}\left[\nabla_\psi \log \frac{dp_\psi}{d\mu}(x, \theta)\right] \\ &= \text{Cov}[-\phi(X_t, \theta_t) + \mathbb{E}[\phi(X_t, \theta_t) | \theta_t]] \\ &= \mathbb{E}[\text{Cov}[-\phi(X_t, \theta_t) + \mathbb{E}[\phi(X_t, \theta_t) | \theta_t]] | \theta_t] \\ &\quad + \underbrace{\text{Cov}[\mathbb{E}[-\phi(X_t, \theta_t) + \mathbb{E}[\phi(X_t, \theta_t) | \theta_t]] | \theta_t]}_{=0} \\ &= \mathbb{E}[\text{Cov}[\phi(X_t, \theta_t) | \theta_t]]\end{aligned}\tag{V.42}$$

Since at $\psi = \psi^*$, $X_t^{\psi^*, \theta_t} \stackrel{d}{=} X_1$.

Lemma V.D.9. Under Assumption A7, A8, A9, we have:

$$\begin{aligned}\mu &:= \inf_{\psi \in \Psi} \lambda_{\min} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)) \in (0, +\infty), \quad L := \sup_{\psi \in \Psi} \lambda_{\max} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)) \in (0, +\infty), \\ \sigma^2 &:= \sup_{\psi \in \Psi} \text{Tr} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)) \in (0, +\infty) \quad (\text{V.43})\end{aligned}$$

and such infimum and supremum are attained.

Proof. We show that $\mu \in (0, +\infty)$, the proof for L, σ^2 being similar. Since $p_{\psi}(\cdot | \theta)$ is minimal, $\log Z(\theta, \psi)$ is strictly convex [263, Proposition 3.1], and we have

$$\left\langle x, \nabla_{\psi}^2 \log Z(\theta, \psi) x \right\rangle > 0, \quad \forall x \in \mathbb{R}^{d_{\psi}} \setminus \{0\}.$$

Consequently, we have [211, Section 11, Ex. 1]

$$0 < \mathbb{E} \left[\left\langle x, \nabla_{\psi}^2 \log Z(\theta_1, \psi) x \right\rangle \right] = \left\langle x, \mathbb{E} \left[\nabla_{\psi}^2 \log Z(\theta_1, \psi) \right] x \right\rangle = \left\langle x, \nabla_{\psi}^2 \bar{\mathcal{L}}(\psi) x \right\rangle,$$

meaning $\lambda_{\min} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)) > 0$, for all ψ . Since the coefficients of $\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)$ are continuous w.r.t ψ , we have [206, p. 39] that $\lambda_{\min} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi))$ is continuous w.r.t ψ .

By the Extreme Value Theorem, it follows that

$$(\mu :=) \inf_{\psi \in \Psi} \lambda_{\min} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)) = \min_{\psi \in \Psi} \lambda_{\min} (\nabla_{\psi}^2 \bar{\mathcal{L}}(\psi)),$$

and since this minimum is attained, we must have $\mu > 0$ from the equations above, as well as $\mu < +\infty$. \square

Lemma V.D.10. Under A7, A8, A9, we have

$$\mathbb{E} \left[(\partial_i \log Z(\theta_t, \psi) - \partial_i \log Z(\theta_t, \psi^*))^2 \right] \leq 2 \times \tau^2 C_{\chi} \times \|\psi - \psi^*\|.$$

Proof.

$$\begin{aligned}(\partial_i \log Z(\theta, \psi) - \partial_i \log Z(\theta, \psi^*))^2 &\leq |\partial_i \log Z(\theta, \psi^*)| \times |\partial_i \log Z(\theta, \psi) - \partial_i \log Z(\theta, \psi^*)| \\ &\quad + |\partial_i \log Z(\theta, \psi)| \times |\partial_i \log Z(\theta, \psi) - \partial_i \log Z(\theta, \psi^*)|\end{aligned}$$

Now,

$$\begin{aligned} & |\partial_i \log Z(\theta, \psi) - \partial_i \log Z(\theta, \psi^*)| \\ &= \left| \int \phi_i(x, \theta) p_\psi(dx | \theta) - \int \phi_i(x, \theta) p_{\psi^*}(dx | \theta) \right| \\ &\leq \mathbb{E} [\phi_i(X_t^{\psi, \theta}, \theta)^2 | \theta]^{1/2} \chi_2(p_\psi(\cdot | \theta), p_{\psi^*}(\cdot | \theta))^{1/2} \end{aligned}$$

Implying, using the Cauchy-Schwarz again, and using Lemma V.D.14, that

$$\begin{aligned} & \mathbb{E} [|\partial_i \log Z(\theta_t, \psi^*)| \times |\partial_i \log Z(\theta_t, \psi) - \partial_i \log Z(\theta_t, \psi^*)|] \\ &\leq \mathbb{E} \left[\left(\mathbb{E} [\phi_i(X_t, \theta_t) | \theta] \times \mathbb{E} [\phi_i(X_t^{\psi, \theta_t}, \theta_t)^2 | \theta]^{1/2} \right)^2 \right]^{1/2} \times \chi_2(p_{\psi, \pi}, p_{\psi^*, \pi})^{1/2} \\ &\stackrel{(a)}{\leq} \mathbb{E} \left[\frac{1}{4} \left(\mathbb{E} [\phi_i(X_t^{\psi, \theta_t}, \theta_t)^4 | \theta_t] + \mathbb{E} [\phi_i(X_t, \theta_t)^4 | \theta_t] \right) \right]^{1/2} \times \chi_2(p_{\psi, \pi}, p_{\psi^*, \pi})^{1/2} \\ &\leq \tau^2 C_\chi \times \|\psi - \psi^*\| \end{aligned}$$

where in (a), we used Jensen's inequality and $2ab \leq a^2 + b^2$. Repeating the same steps for the second term concludes the proof. \square

Lemma V.D.11. *Under A7, A8, A9, we have*

$$\mathbb{E} [\|\nabla_\psi \log Z(\theta_t, \psi) - \nabla_\psi \log Z(\theta_t, \psi^*)\|^2] \leq 2pC_\chi \tau^2 \times \|\psi - \psi^*\|$$

Proof. The proof follows from Lemma V.D.10, noting that

$$\|\nabla_\psi \log Z(\theta, \psi) - \nabla_\psi \log Z(\theta, \psi^*)\|^2 = \sum_{i=1}^p (\partial_i \log Z(\theta, \psi) - \partial_i \log Z(\theta, \psi^*))^2.$$

Note, morevoer, than since:

$$\left\| \mathbb{E} [\phi(X_t, \theta_t)] - \mathbb{E} [\phi(X_t^{\psi, \theta_t}, \theta_t)] \right\| \leq p\tau^2 C_\chi \|\psi - \psi^*\|,$$

it holds that, $\mu < 2p\tau^2 C_\chi$. \square

Lemma V.D.12. *Under A7,A8,A9, we have*

$$\mathbb{E} [\|g(\psi, X_t, \theta_t)\|^2] \leq \sigma^2 + 2\tau^2 C_\chi \|\psi - \psi^*\|$$

Proof. We have:

$$\begin{aligned}
& \mathbb{E} \left[\|g(\psi, X_t, \theta_t)\|^2 \mid \psi_{t-1} \right] \\
&= \mathbb{E} \left[\left\| (\phi(X_t, \theta_t) - \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t]) + \left(\mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] - \mathbb{E}[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t] \right) \right\|^2 \right] \\
&\leq \mathbb{E}[\text{TrCov}[\phi(X_t, \theta_t) \mid \theta_t]] + \mathbb{E} \left[\|\nabla_\psi \log Z(\theta_t, \psi_{t-1}) - \nabla_\psi Z(\theta_t, \psi^*)\|^2 \right] \\
&\leq \sigma^2 + 2\tau^2 C_\chi \|\psi - \psi^*\|^2.
\end{aligned}$$

□

Lemma V.D.13. *We have*

$$\text{Cov}[g(\psi, X_t, \theta_t)] = \mathbb{E}[\text{Cov}[\phi(X_t, \theta_t) \mid \theta_t]] + e_3(\psi_{t-1}, m)$$

where $\|e_3(\psi_{t-1}, m)\| \leq 2p^2 \tau C_\chi \|\psi_{t-1} - \psi^*\|$.

Proof.

$$\text{Cov}[g(\psi, X_t, \theta_t)]$$

$$\begin{aligned}
&= \text{Cov}[g(\psi, X_t, \theta_t) - \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] + \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t]] \\
&= \text{Cov} \left[\phi(X_t, \theta_t) - \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] + \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] - \mathbb{E}[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t] \right] \\
&= \text{Cov}[\phi(X_t, \theta_t) - \mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t]] + \text{Cov} \left[\mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] - \mathbb{E}[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t] \right]
\end{aligned}$$

and, using Lemma V.D.10, we have that

$$\begin{aligned}
& \|\text{Cov} \left[\mathbb{E}[\phi(X_t, \theta_t) \mid \theta_t] - \mathbb{E}[\phi(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t] \right]\|_F \\
&= \|\text{Cov}[\nabla \log Z(\theta_t, \psi) - \nabla \log Z(\theta_t, \psi^*)]\|_F \\
&\leq \left\| \mathbb{E} \left[(\nabla \log Z(\theta_t, \psi) - \nabla \log Z(\theta_t, \psi^*)) (\nabla \log Z(\theta_t, \psi) - \nabla \log Z(\theta_t, \psi^*))^\top \right] \right\|_F \\
&\leq p^2 \left\| \mathbb{E} \left[(\nabla \log Z(\theta_t, \psi) - \nabla \log Z(\theta_t, \psi^*)) (\nabla \log Z(\theta_t, \psi) - \nabla \log Z(\theta_t, \psi^*))^\top \right] \right\|_\infty \\
&\stackrel{(a)}{\leq} 2p^2 \tau C_\chi \|\psi - \psi^*\|
\end{aligned}$$

Where, in (a), we noted that the largest component must lie on the diagonal, and we used Lemma V.D.10. □

V.D.2.6 Auxiliary Lemmas

Lemma V.D.14. *We have:*

$$\mathbb{E} [\chi^2(p_\psi(\cdot|\theta_1), p_{\psi^*}(\cdot|\theta_1))] = \chi^2(p_{\psi,\pi}, p_{\psi^*,\pi})$$

Proof. Note first that for all $\psi \in \Psi$, $\frac{dp_{\psi,\pi}}{dp_{\psi^*,\pi}}$ exists, and additionally, for all $\theta \in \Theta$, $\frac{dp_\psi}{dp_{\psi^*}}(\cdot|\theta)$ exists. Moreover, we have:

$$\frac{dp_{\psi,\pi}}{dp_{\psi^*,\pi}} = \frac{dp_\psi}{dp_{\psi^*}}.$$

Consequently, we have:

$$\begin{aligned} \chi^2(p_{\psi,\pi}, p_{\psi^*,\pi}) &:= \int \left(\frac{dp_{\psi,\pi}}{dp_{\psi^*,\pi}}(x, \theta) - 1 \right)^2 dp_{\psi^*,\pi} \\ &= \int \left(\frac{dp_\psi}{dp_{\psi^*,\pi}}(x|\theta) - 1 \right)^2 \frac{dp_{\psi^*,\pi}}{d(\mu \times \pi)}(x, \theta) d(\mu \times \pi)(x, \theta) \\ &\stackrel{(a)}{=} \int \left(\frac{dp_\psi}{dp_{\psi^*}}(x|\theta) - 1 \right)^2 \frac{dp_{\psi^*}}{d\mu}(x|\theta) d(\mu \times \pi)(x, \theta) \\ &\stackrel{(b)}{=} \int \left(\int \left(\frac{dp_\psi}{dp_{\psi^*}}(x|\theta) - 1 \right)^2 \frac{dp_{\psi^*}}{d\mu}(x|\theta) d\mu(x) \right) d\pi(\theta) \\ &= \int \chi^2(p_\psi(\cdot|\theta), p_{\psi^*}(\cdot|\theta)) d\pi(\theta) = \mathbb{E} [\chi^2(p_\psi(\cdot|\theta_1), p_{\psi^*}(\cdot|\theta_1))]. \end{aligned}$$

In (a), we used (V.21), and in (b), we used Fubini's theorem [211, Theorem 8.8]. \square

Lemma V.D.15. *For all $\psi \in \Psi$, $k \in [4]$, we have:*

$$\mathbb{E} \left[\left\| P_{\psi, \theta_t}^m \phi(X_t, \theta_t) \right\|^k \right] < +\infty.$$

Proof.

$$\begin{aligned}
\mathbb{E} \left[\left\| P_{\psi, \theta_t}^m \phi(X_t^{\psi, \theta_t}, \theta_t) \right\|^k \right] &= \int \left\| P_{\psi, \theta_t}^m \phi(X_t^{\psi, \theta_t}, \theta_t) \right\|^k \frac{dp_\psi}{d\mu}(\cdot | \theta) d(\mu \times \pi)(x, \theta) \\
&\stackrel{(a)}{=} \int \left(\int \left\| P_{\psi, \theta}^m \phi(x, \theta) \right\|^k dp_\psi(x | \theta) \right) d\pi(\theta) \\
&\stackrel{(b)}{\leq} \int \left(\int \|\phi(x, \theta)\|^k dp_\psi(x | \theta) \right) d\pi(\theta) \\
&\stackrel{(c)}{=} \mathbb{E} \left[\left\| \phi(X_t^{\psi, \theta_t}, \theta_t) \right\|^k \right] \stackrel{(d)}{<} +\infty
\end{aligned}$$

Where (a) and (c) follow from Fubini's Theorem, and (b) uses Jensen inequality. Moreover, we have that $\chi_2(p_{\psi, \pi}, p_{\psi^*, \pi}) \leq C_\chi \|\psi - \psi^*\| < +\infty$. We can apply [247, Lemma C.5] to conclude that $\mathbb{E} [\|P_{\psi, \theta_t}^m \phi(X_t, \theta_t)\|^k] < +\infty$. \square

Lemma V.D.16. *We have that, for all $x \geq 0$,*

$$0 \leq \frac{(\log(\frac{x}{2} + 1))^2 - 2\log(\frac{x}{2} + 1)}{x} \leq 1$$

Proof. Let $f(x) = (\log x)^2$, defined for all $x > 0$. We have $f'(x) = 2\log x/x$, and $f''(x) = 2(1 - \log x)/x^2$. Consequently, $f''(x) > 0$ for all $x \geq e$, and we have, since $f(e) = 1$ and $f'(e) = \frac{2}{e}$,

$$\begin{aligned}
1 &\leq (\log x)^2 \leq \frac{2x}{e} - 1, & \forall x \geq e \\
\iff 1 &\leq (\log(x+e))^2 \leq \frac{2x}{e} + 1, & \forall x \geq 0 \\
\iff 0 &\leq \log\left(\frac{e}{2}x + e\right)^2 - 1 \leq x, & \forall x \geq 0 \\
\iff 0 &\leq \log\left(\frac{x}{2} + 1\right)^2 - 2\log\left(\frac{x}{2} + 1\right) \leq x, & \forall x \geq 0
\end{aligned}$$

\square

Lemma V.D.17. *Under A7, A8, A9, for any function $f \in \{\phi_i\}_{i=1}^p$, and for all $\psi \in \Psi$, we have, for $k \in [4]$,*

$$\left| \mathbb{E} \left[(P_{\psi, \theta_t}^m f^k)(X_t, \theta_t) \right] - \mathbb{E} \left[f(X_t^{\psi, \theta_t}, \theta_t)^k \right] \right| \leq \tau^k \times C_\chi \|\psi - \psi^*\|.$$

If, additionally, any variant of A10 holds, for all $f \in \{\phi_i\}_{i=1}^p \cup \{\phi_i \phi_j\}_{i,j=1}^p$, we have:

$$\left| \mathbb{E} \left[(\mathbf{P}_{\psi, \theta_t}^m f)(X_t, \theta_t) \right] - \mathbb{E} \left[f(X_t^{\psi, \theta_t}, \theta_t) \right] \right| \leq \begin{cases} \tau C_\chi \times \|\psi - \psi^*\| \times F(4m)^{1/4} & \text{if } f \in \{\phi_i\}_{i=1}^p \\ \tau^2 C_\chi \times \|\psi - \psi^*\| \times F(4m)^{1/4} & \text{if } f \in \{\phi_i \phi_j\}_{i,j=1}^p \end{cases} \quad (\text{V.44})$$

where $F(m)$ is given by Lemma V.D.4.

Proof. We have, for all $f \in \mathcal{L}_2(p_{\psi, \pi})$,

$$\begin{aligned} & \left| \mathbb{E} \left[(\mathbf{P}_{\psi, \theta_t}^m f)(X_t, \theta_t) \right] - \mathbb{E} \left[f(X_t^{\psi, \theta_t}, \theta_t) \right] \right| \\ & \stackrel{(a)}{=} \left| \mathbb{E} \left[\mathbb{E} \left[(\mathbf{P}_{\psi, \theta_t}^m f)(X_t, \theta_t) \mid X_t, \theta_t \right] \mathbb{E} \right] - \mathbb{E} \left[\mathbb{E} \left[f(X_t^{\psi, \theta_t}, \theta_t) \right] \mid \theta_t \right] \right| \\ & \stackrel{(b)}{=} \left| \mathbb{E} \left[\mathbb{E} \left[\mathbf{P}_{\psi, \theta_t}^m \left(f - \mathbb{E} \left[f(X_t^{\psi, \theta_t}, \theta_t) \mid \theta_t \right] \right) (X_t, \theta_t) \right] \mid X_t, \theta_t \right] \right| \\ & = \int \left(\int \mathbf{P}_{\psi, \theta}^m \bar{f}_\psi(x, \theta) \left(\frac{dp_{\psi^*, \pi}}{dp_{\psi, \pi}}(x, \theta) - 1 \right) p_{\psi, \pi}(dx, d\theta) \right) \\ & \stackrel{(c)}{\leq} \chi^2(p_{\psi, \pi}, p_{\psi^*, \pi})^{1/2} \left(\int \mathbf{P}_{\psi, \theta}^m \bar{f}_\psi(x, \theta)^2 p_{\psi, \pi}(dx, d\theta) \right)^{1/2} \\ & \leq C_\chi \|\psi - \psi^*\| \left(\int \left(\int \mathbf{P}_{\psi, \theta}^m \bar{f}_\psi(x, \theta)^2 p_\psi(dx \mid \theta) \right) \pi(d\theta) \right)^{1/2} \end{aligned}$$

where, in (a) we used the law of total expectation, in (b) we used the fact that $\mathbf{P}_{\psi, \theta}^m$ leaves constant functions (w.r.t x) unchanged; in (c), we used the Cauchy-Schwarz inequality and the definition of the χ^2 divergence. Now, since $\mathbf{P}_{\psi, \theta}^m$ is a contraction in $\mathcal{L}^2(p_\psi(\cdot \mid \theta)) / \text{Vec}(\mathbf{1})$, using A8 and setting $f := \phi_i^k$, we have:

$$\begin{aligned} \left| \mathbb{E} \left[(\mathbf{P}_{\psi, \theta_t}^m f^k)(X_t, \theta_t) \right] - \mathbb{E} \left[\phi_i^k(X_t^{\psi, \theta_t}, \theta_t) \right] \right| & \leq C_\chi \|\psi - \psi^*\| \mathbb{E} \left[\phi_i^{2k}(X_t^{\psi, \theta_t}, \theta_t) \right]^{1/2} \\ & \leq \tau^k \times C_\chi \|\psi - \psi^*\| \end{aligned}$$

For the second inequality, we set $f \in \{\phi_i\}_{i=1}^p \cup \{\phi_i\phi_j\}_{i,j=1}^p$, and, using A10, we have:

$$\begin{aligned} C_\chi \|\psi - \psi^*\| & \left(\int \left(\int P_{\psi,\theta}^m \bar{f}_\psi^2(x, \theta) p_\psi(dx | \theta) \right) \pi(d\theta) \right)^{1/2} \\ & \stackrel{(d)}{\leq} C_\chi \|\psi - \psi^*\| \times \left(\alpha(\theta)^{2m} \int \bar{f}_\psi^{-2}(x, \theta) p_{\psi,\pi}(dx, d\theta) \right)^{1/2} \\ & \stackrel{(e)}{\leq} C_\chi \|\psi - \psi^*\| \times \left(\int \alpha(\theta)^{4m} \pi(d\theta) \right)^{1/4} \left(\int \left(\int \bar{f}_\psi^{-2}(x, \theta) p_\psi(dx | \theta) \right)^2 d\pi(\theta) \right)^{1/4} \\ & \stackrel{(f)}{\leq} C_\chi \|\psi - \psi^*\| \times F(4m)^{1/4} \times \left(\mathbb{E} \left[\mu_2 \left(f(X_t^{\psi,\theta_t}, \theta_t) | \theta_t \right)^2 \right] \right)^{1/4} \end{aligned}$$

in (d), we used assumption A10, in (e), we used the Cauchy-Schwarz inequality on $\pi(\theta)$, and in (f) we used assumption A9 and Lemma V.D.14. Now if $f \in \{\phi_i\}_{i=1}^p$, we have:

$$\mathbb{E} \left[\mu_2 \left(f(X_t^{\psi,\theta_t}, \theta_t) | \theta_t \right)^2 \right] \leq \mathbb{E} \left[\mu^4 \left(f(X_t^{\psi,\theta_t}, \theta_t) | \theta_t \right) \right] \leq \tau^4$$

while, if $f \in \{\phi_i\phi_j\}_{i,j=1}^p$, we have:

$$\begin{aligned} \mathbb{E} \left[\mu_2 \left((\phi_i\phi_j)(X_t^{\psi,\theta_t}, \theta_t) | \theta_t \right)^2 \right] & \leq \mathbb{E} \left[(\phi_i\phi_j)(X_t^{\psi,\theta_t}, \theta_t)^4 | \theta_t \right] \\ & \leq \frac{1}{2} \left[\mathbb{E} \left[\phi_i(X_t^{\psi,\theta_t}, \theta_t)^8 \right] + \mathbb{E} \left[\phi_j(X_t^{\psi,\theta_t}, \theta_t)^8 \right] \right] \leq \tau^8. \end{aligned}$$

Here, we used the inequality $(a+b)^k \leq 2^{k-1}(a^k + b^k)$, as well as $ab < a^2 + b^2$, both for $a, b \in \mathbb{R}$, as well as A8. \square

We now establish a second approximation lemma in a stronger norm.

Lemma V.D.18. *For $f \in \{\phi_i\}_{i=1}^p$, Under A7, A8, A9, A10, we have that*

$$\mathbb{E} \left[\left\| P_{\psi,\theta_t}^m f(X_t, \theta_t) - \mathbb{E}[f(X_t, \theta_t) | \theta_t] \right\|^2 \right] \leq F(4m)^{1/2} \tau^2 + C_\chi F(4m)^{1/8} \tau^2 \|\psi - \psi^*\|$$

Proof. We have:

$$\begin{aligned}
& \mathbb{E} \left[\left\| P_{\psi, \theta_t}^m f(X_t, \theta_t) - \mathbb{E}[f(X_t, \theta_t) | \theta_t] \right\|^2 \right] \\
&= \int \left\| P_{\psi, \theta}^m \bar{f}_\psi(x, \theta) \right\|^2 d p_{\psi^\star, \pi}(x, \theta) \\
&= \int \left\| P_{\psi, \theta}^m \bar{f}_\psi(x, \theta) \right\|^2 d p_{\psi, \pi}(x, \theta) \\
&\quad + \int \left\| P_{\psi, \theta}^m \bar{f}_{\psi_{t-1}}(x, \theta) \right\|^2 \left(\frac{d p_{\psi^\star, \pi}}{d p_{\psi, \pi}} - 1 \right) d p_{\psi, \pi}(x, \theta) \\
&\leq F(4m)^{1/2} \left(\mathbb{E} \left[\mu_2(f(X_1^\psi, \theta_1) | \theta_1) \right]^2 \right)^{1/2} \\
&\quad + \chi^2(p_{\psi^\star, \pi}, p_{\psi, \pi})^{1/2} \left(\int \left\| P_{\psi, \theta_t}^m \bar{f}_\psi f(x, \theta) \right\|^4 d p_{\psi, \pi}(x, \theta) \right)^{1/2}
\end{aligned}$$

where the last line used the derivations (d), (e) of Lemma V.D.17 for the first term, and the Cauchy-Schwarz inequality for the second term. Now, we have:

$$\begin{aligned}
& \int \left\| P_{\psi, \theta}^m \bar{f}_\psi(x, \theta) \right\|^4 d p_{\psi, \pi}(x, \theta) \\
&\leq \left(\int \left\| P_{\psi, \theta}^m \bar{f}_\psi(x, \theta) \right\|^2 d p_{\psi, \pi}(x, \theta) \right)^{1/2} \left(\int \left\| P_{\psi, \theta}^m \bar{f}_\psi f(x, \theta) \right\|^6 d p_{\psi, \pi}(x, \theta) \right)^{1/2} \\
&\leq F(4m)^{1/4} \left(\mathbb{E} \left[\mu_2(f(X_t^{\psi, \theta_t}, \theta_1) | \theta_1) \right]^2 \right)^{1/4} \times \left(\mathbb{E} \left[\mu_6(f(X_t^{\psi, \theta_t}, \theta_1) | \theta_1) \right] \right)^{1/2} \\
&\leq F(4m)^{1/4} \tau^4
\end{aligned}$$

Where for the first term, we used derivations (d), (e) of Lemma V.D.17, and for the second term, we used the fact that $P_{\psi, \theta}^m$ is a contraction in $\mathcal{L}^6(p_\psi(\cdot | \theta)) / \text{Vec}(\mathbf{1})$, as well as A9 and A10. Putting everything together concludes the proof. \square

V.D.2.7 On the existence of moments in conditional exponential families

We comment on the fact that our analysis requires moment assumptions that are not required in the unconditional case. Recall that combining Lemma V.D.8 with (V.46),

we have that

$$\begin{aligned} \frac{\partial^{|\alpha|} \mathbb{E}[\log Z(\theta_t, \psi)]}{\partial \psi_1^{\alpha_1} \dots \partial \psi_d^{\alpha_d}} &= \mathbb{E} \left[\frac{\partial^{|\alpha|} \log \mathbb{E}[e^{\langle \xi, \phi(X_t^{\psi, \theta_t}) \rangle}]}{\partial \xi_1^{\alpha_1} \dots \partial \xi_d^{\alpha_d}} \Big|_{\xi=0} \right] \\ &:= \mathbb{E} \left[\kappa_\alpha(\phi(X_t^{\psi, \theta_t}, \theta_t)) \Big| \theta_t \right]. \end{aligned}$$

where we noted $\kappa_\alpha(\phi(X^\psi) \mid \theta_1)$ the “conditional cumulant” of order α of the sufficient statistics $\phi(X^\psi)$ given θ_1 . Thus, the average conditional cumulant $\kappa_\alpha(\phi(X^\psi) \mid \theta_1)$ is guaranteed to be finite. However, a finite average conditional cumulant does not imply a finite moment; Below, we provide an explicit construction where the former is finite while the latter is not. Let $\mathcal{X} = \mathbb{N}$, endowed with the discrete topology and its associated Borel σ -algebra. Let μ the counting measure on \mathbb{N} and $p = 1$. Let π be any distribution with a finite first moment but an infinite second moment (such as a Student distribution with $v = 2$). Consider a Poisson conditional exponential family given by,

$$\frac{dq_\psi}{d\mu}(x|\theta) := e^{-\theta\psi} \frac{(\theta\psi)^x}{x!} = \frac{e^{x\log\theta}}{x!} e^{x\log\psi - \theta\psi}.$$

Here, we have $c(x, \theta) = \frac{e^{x\log\theta}}{x!}$, $\phi(x, \theta) = x$, and $\log Z(\theta, \psi) = \theta\psi$, and, consequently, $\mathcal{D} = \mathbb{R}_+$. Moreover, by properties of Poisson distributions, we have:

$$\mathbb{E}[\kappa_k(\phi(X^\psi) \mid \theta_1)] = \psi \mathbb{E}[\theta_1], \quad \mathbb{E}\left[\mathbb{E}\left[\|\phi(X^\psi)\|^2 \mid \theta_1\right]\right] = \mathbb{E}[\psi\theta_1 + \psi^2\theta_1^2] = +\infty.$$

V.D.2.8 Background on exponential families, cumulants and convexity

Exponential Families: Definition Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathcal{Y}, \mathcal{Y})$ be a measurable space, μ a σ -finite measure on \mathcal{Y} , and $p \in \mathbb{N}$ a positive integer. Exponential families (EF) [32, 264] form a well-studied class of probability distributions on \mathcal{Y} , defined by

$$\frac{dp_\psi}{d\mu}(y) := c(y) e^{\langle \psi, \phi(y) \rangle - \log Z(\psi)}, \quad \log Z(\psi) := \log \left(\int c(y) e^{\langle \psi, \phi(y) \rangle} \mu(dy) \right),$$

$$\mathcal{D} := \{\psi \in \mathbb{R}^p; \log Z(\psi) < +\infty\} . \quad (\text{V.45})$$

Here, $\psi \in \Psi \subseteq \mathbb{R}^p$ is a finite-dimensional parameter called the *natural parameter*, $\phi : \mathcal{Y} \mapsto \mathbb{R}^p$ is a \mathcal{Y} -measurable function called the *sufficient statistics*, and $\log Z(\psi)$, the *log-normalizing* (or *log-partition*) function, is a quantity ensuring that p_ψ integrates to 1 over \mathcal{Y} . Finally, $c(y)$ a non-negative \mathcal{Y} -measurable function, called the carrier density. The expression for p_ψ defines a valid probability measure for all $\psi \in \mathcal{D}$, the set of all values of ψ for which the $\log Z(\psi)$ is finite. As we will see, our algorithm will not require such normalizer to be known. We refer to [264, 247] for an introduction to exponential families and their properties. Such densities are specific instances of Energy-Based Models, with energy function $E_\psi(y) := -\langle \psi, \phi(y) \rangle - \log c(y)$. Throughout the rest of Section V.D.2.8, we will denote by Y_ψ a random variable distributed according to p_ψ .

Properties of the log-partition function A well-known fact about exponential families is that the log-normalizing function is a smooth, convex function of ψ .

Theorem V.D.19 (32, Theorem 2.2, Corollary 2.3). *The function $\log Z(\cdot)$ infinitely differentiable on the interior of \mathcal{D} . Moreover, it is convex, and for all $\psi \in \text{int}(\mathcal{D})$, we have*

$$\nabla_\psi \log Z(\psi) = \mathbb{E}[\phi(Y^\psi)], \quad \nabla_\psi^2 \log Z(\psi) = \text{Cov}[\phi(Y^\psi)]$$

Higher-order derivatives The higher-order derivatives of $\log Z(\cdot)$ are related to the cumulants of the distribution p_ψ . Indeed, let $\psi \in \text{int}(\mathcal{D})$, and \mathcal{V} an open neighborhood of ψ included in \mathcal{D} . Then for all ξ s.t. $\psi + \xi \in \mathcal{V}_{\psi, \theta}$, we have

$$\begin{aligned} \frac{Z(\psi + \xi)}{Z(\psi)} &= \int e^{\langle \xi, \phi(y) \rangle} e^{\langle \psi, \phi(y) \rangle - \log Z(\psi)} dx = \mathbb{E}\left[e^{\langle \xi, \phi(Y^\psi) \rangle}\right], \\ \implies \log Z(\psi + \xi) - \log Z(\psi) &= \log \mathbb{E}\left[e^{\langle \xi, \phi(Y^\psi) \rangle}\right] =: K_{\phi(Y^\psi)}(\xi). \end{aligned}$$

The right-hand side of the last equation is precisely the *cumulant generating function* [164] of the random variable $\phi(Y^\psi)$ —formally, the *pushforward* of Y^ψ by the function ϕ — evaluated at ξ ; see Section V.D.2.9 for a more complete introduction to cumulants. As, by Theorem V.D.19, $\log Z$ is smooth, so is the cumulant

generating function: in particular, we have, for any multi-index $\alpha \in \mathbb{N}^p$ such that $|\alpha| := \sum_{i=1}^p \alpha_i \geq 1$,

$$\frac{\partial^{|\alpha|} K_{\phi(Y^\psi)}(\xi)}{\partial \xi_1^{\alpha_1} \dots \partial \xi_d^{\alpha_d}} \Big|_{\xi=0} = \frac{\partial^{|\alpha|} \log Z(\psi)}{\partial \psi_1^{\alpha_1} \dots \partial \psi_d^{\alpha_d}} \Big|_{\psi=\psi} := \kappa_\alpha(\phi(Y^\psi)), \quad (\text{V.46})$$

where κ_α is the α -th cumulant of $\phi(Y^\psi)$. Consequently, all cumulants of $\phi(Y^\psi)$ exist, and are given by the derivatives of $\log Z(\cdot)$. Plugging in the definition of the first and second order cumulants allows to recover the expression for the first and second derivatives of $\log Z$ given in Theorem V.D.19. This convexity will be instrumental to obtain our convergence guarantees for the algorithm introduced in this paper.

V.D.2.9 Background on (multivariate) cumulants

Let Z be a real-valued random variable. If there exists some $\mathcal{D}_{\kappa,Z} \subset \mathbb{R}$ open and containing 0 such that $\mathbb{E}[e^{tZ}] < +\infty$ for all $t \in \mathcal{D}_{\kappa,Z}$, then we can define the function K_Z , called the *cumulant-generating function* of Z , as

$$\begin{aligned} K_Z : \mathcal{D}_{\kappa,Z} &\longrightarrow \mathbb{R} \\ t &\longmapsto K_Z(t) := \log \mathbb{E}[e^{tZ}]. \end{aligned}$$

For such Z , the cumulant of order k of Z is defined as

$$\kappa_k(Z) := \left. \frac{\partial^k K_Z(t)}{\partial t^k} \right|_{t=0}$$

for all k such that the derivative exists. Given a d-dimensional random variable \tilde{Z} , if there is a set $\mathcal{D}_{\kappa,\tilde{Z}} \subset \mathbb{R}^d$ open and containing 0 such that $\mathbb{E}[e^{\langle \xi, \tilde{Z} \rangle}] < +\infty$ for all $\xi \in \mathcal{D}_{\kappa,\tilde{Z}}$, we can define the cumulant-generating function of \tilde{Z} as

$$\begin{aligned} K_{\tilde{Z}} : \mathcal{D}_{\kappa,\tilde{Z}} &\longrightarrow \mathbb{R} \\ \xi &\longmapsto K_{\tilde{Z}}(\xi) := \log \mathbb{E}[e^{\langle \xi, \tilde{Z} \rangle}], \end{aligned}$$

and for a multi index $\alpha := (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, the multivariate cumulant of order α is defined as

$$\kappa_\alpha(\tilde{Z}) := \frac{\partial^{|\alpha|} K_{\tilde{Z}}(\xi)}{\partial \xi_1^{\alpha_1} \dots \partial \xi_d^{\alpha_d}} \Big|_{\xi=0}$$

for all α such that the derivative exists. Such cumulants admit conditional extensions, obtained by conditionning the expectation present the cumulant generating function: given some other random variable Z , we define the conditional cumulant generating functions $K_{Z|Y}$ and $K_{\tilde{Z}|Y}$ as

$$K_{Z|Y}(t, y) := \log \mathbb{E}[e^{tZ} | Y = y] \quad \text{and} \quad K_{\tilde{Z}|Y}(\xi, y) := \log \mathbb{E}[e^{\langle \xi, \tilde{Z} \rangle} | Y = y].$$

and the conditional cumulants of order k and α are defined as

$$\kappa_k(Z|Y)(y) := \frac{\partial^k K_{Z|Y}(t, y)}{\partial t^k} \Big|_{t=0} \quad \text{and} \quad \kappa_\alpha(\tilde{Z}|Y)(y) := \frac{\partial^{|\alpha|} K_{\tilde{Z}|Y}(\xi, y)}{\partial \xi_1^{\alpha_1} \dots \partial \xi_d^{\alpha_d}} \Big|_{\xi=0}.$$

V.E Proof for Double Variational Inference

V.E.1 Proof of Proposition V.3.2

We repeat Proposition V.3.2 below.

Proposition V.E.1. *Let $\psi \in \Psi$. Assume that $E_\psi(\theta, x)$ is differentiable w.r.t θ , such that for all θ , there exists a neighborhood \mathcal{V}_θ containing θ such that $\nabla_\theta E_\psi(x, \theta) e^{-E_\psi(x, \theta)} \leq \ell(x)$ for all $x \in \mathcal{X}$ and for some integrable function ℓ . and let \mathcal{F} be the space of 1-differentiable real-valued functions on Θ . Let v be any distribution with full support on Θ , and let $f^* \in \mathcal{F}$. Then f^* is a solution of:*

$$\min_{f \in \mathcal{F}} \mathbb{E}_{q_\psi(x \in \cdot | \theta = \diamond) v(\theta \in \cdot)} \|\nabla f(\theta) + \nabla_\theta E_\psi(x, \theta)\|^2$$

if and only if $f^* = \log Z(\theta, \psi) + C$, for some constant C .

Proof. The proof stems from the following definition of the conditional expectation

$\mathbb{E}[Z|Y]$ for two random vectors Z and Y [99]:

$$\mathbb{E}[Z|Y] = \arg \min_{g \text{ measurable}} \mathbb{E} \left(\|Z - g(Y)\|^2 \right)$$

Applying this result to $Y = \theta$ (with distribution v) and $Z = -\nabla_\theta E_\psi(x, \theta)$, where $x|\theta$ is sampled according to $q_\psi(x \in \cdot | \theta = \diamond)$, the conditional expectation $\theta \mapsto -\mathbb{E}_{q_\psi(x \in \cdot | \theta = \diamond)} \nabla_\theta E_\psi(x, \theta)$ is thus given by

$$\arg \min_{g \text{ measurable}} \mathbb{E}_{(x, \theta) \sim q_\psi(x \in \cdot | \theta = \diamond) v(\theta \in \cdot)} \|\nabla_\theta E_\psi(x, \theta) + g(\theta)\|^2 \quad (\text{V.47})$$

As we show, the minimizers of Equation (V.47) and of Proposition V.3.2

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(x, \theta) \sim q_\psi(x \in \cdot | \theta = \diamond) v(\theta \in \cdot)} \|\nabla_\theta E_\psi(x, \theta) + \nabla f(\theta)\|^2.$$

are connected: Indeed, consider any primitive f_C^* of the conditional expectation function $g : \theta \mapsto -\mathbb{E}_{q_\psi(x \in \cdot | \theta = \diamond)} \nabla_\theta E_\psi(x, \theta)$. Under the conditions of the proposition, one can apply the dominated convergence theorem, such that the log-normalizer $\log Z(\theta, \psi)$ is differentiable with gradient given by

$$\nabla \log Z(\theta) = -\mathbb{E}_{q_\psi(x \in \cdot | \theta = \diamond)} \nabla_\theta E_\psi(x, \theta).$$

Consequently, f_C^* is given by $\log Z(\theta, \psi) + C$ for an additive constant C . By construction, f_C^* is differentiable and thus $f_C^* \in \mathcal{F}$. Moreover, for any $f \in \mathcal{F}$, we have, since ∇f is measurable,

$$\|\nabla_\theta E_\psi(x, \theta) + \nabla f(\theta)\|^2 \geq \min_{g \text{ measurable}} \|\nabla_\theta E_\psi(x, \theta) + \nabla f_C^*(\theta)\|^2 \geq \|\nabla_\theta E_\psi(x, \theta) + \nabla f^*(\theta)\|^2 \quad (\text{V.48})$$

Making all f_C^* the minimizers of Proposition V.3.2's problem. \square

V.E.2 Proof of Proposition V.3.3

Proof. Let us define $\mathcal{L}(\eta) = \mathbb{E}_{v(\theta \in \cdot)} q_{\psi}(x \in \cdot | \theta) \ell(x, \theta, \eta)$. We have

$$\begin{aligned} & \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \ell(x, \theta, \eta) \\ &= \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \|\nabla_{\theta} LZ_{\eta}(\theta) - \nabla_{\theta} E_{\psi}(x, \theta)\|^2 \\ &= \|\nabla_{\theta} LZ_{\eta}(\theta)\|^2 - 2 \langle \nabla_{\theta} LZ_{\eta}(\theta), \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \nabla_{\theta} E_{\psi}(x, \theta) \rangle + \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \|\nabla_{\theta} E_{\psi}(x, \theta)\|^2 \\ &= \|\nabla_{\theta} LZ_{\eta}(\theta)\|^2 - 2 \langle \nabla_{\theta} LZ_{\eta}(\theta), LZ_{\eta^*}(\theta) \rangle + \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \|\nabla_{\theta} E_{\psi}(x, \theta)\|^2 \end{aligned}$$

Applying this identity to both η_s and η^* , we obtain

$$\begin{aligned} & \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \ell(x, \theta, \eta_s) - \mathbb{E}_{q_{\psi}(x \in \cdot | \theta)} \ell(x, \theta, \eta^*) \\ &= \|\nabla_{\theta} LZ_{\eta_s}(\theta)\|^2 - \langle \nabla_{\theta} LZ_{\eta_s}(\theta), \eta^* \rangle + \|\nabla_{\theta} LZ_{\eta^*}(\theta)\|^2. \end{aligned}$$

Marginalizing out over $\theta \sim v(\theta \in \cdot)$, we recover the difference of the population losses, e.g.

$$\begin{aligned} \tilde{\mathcal{L}}(\eta_s) - \tilde{\mathcal{L}}(\eta^*) &= \mathbb{E}_{v(\theta \in \cdot)} \|\nabla_{\theta} \log q_{\psi^* \eta_s^*}(\theta, x) - \nabla_{\theta} \log q_{\psi^*, \eta^*}(\theta, x)\|^2 \\ &\leq 2\mathcal{R}_s \end{aligned}$$

Where the last inequality follows from using the standard symmetrization argument [see Theorem 26.3 of 222]. \square

V.E.3 Empirical improvements to DVI

We propose a few improvements to the DVI method outlined in Algorithm 7.

Choice of v . The DVI method allows for any choice v of proposal on θ . In practice, we set $v(\theta \in \cdot) = q_{\psi_r^*}(\theta | x_o)$, the previous round's posterior estimate. By doing so, we concentrate the log-Z network training data around parameters most relevant to the observation x_o , ensuring that our $LZ_{\eta}(\theta)$ is accurate on the regions of parameter space that are most relevant to the problem at hand.

Variance reduction. The training signal for $\nabla_{\theta} LZ_{\eta}(\theta)$ is given by data points $\left\{ \theta_i, -\mathbb{E}_{x \sim q_{\psi}(x \in \cdot | \theta_i)} [\nabla_{\theta} E_{\psi}(x, \theta_i)] \right\}_i$. The version of DVI in Algorithm 7 effectively approximates this conditional expectation with an empirical one-sample estimate:

$-\nabla_{\theta} E_{\psi}(X_i, \theta_i)$, for $X_i \sim q_{\psi}(x \in \cdot | \theta_i)$. We can reduce the variance of this estimate by sampling multiple points from the likelihood. The approximation then becomes

$$-\frac{1}{M} \sum_{m=1}^M E_{\psi}(X_i^m, \theta_i) \quad X_i^m \stackrel{\text{i.i.d.}}{\sim} q_{\psi}(x \in \cdot | \theta_i).$$

Hyperparameter tuning. For all experiments in the main paper and appendix, we use the same set of hyperparameters, with the exception of in the following problems:

- Gaussian Linear Uniform: `max_iter=10` (default: 500). We reduce the number of iterations of EBM training in order to avoid overfitting, because the true likelihood of this model is a very simple multivariate Gaussian [160]. We do this only when `num_samples==100 or 1000`.
- Lotka-Volterra: `learning_rate=0.001` (default: 0.01).
- Pyloric: `learning_rate=0.0001` (default: 0.01).

V.F Additional Experimental and Inferential Details

V.F.1 On the performance metric used for the Pyloric network

We motivate below the use of the Energy Distance to assess posterior estimators for the pyloric network model, and discuss the performance of SUNLE using alternative metrics used in prior work.

The Energy Distance (ED, 240, 80) is a metric between probability distributions that we used to produce the scalar values plotted in Figure 4. It was previously used in SBI in [188, Appendix 2], and proceeds by comparing the true likelihood $p_{\theta^*}(x) := p(x|\theta^*)$ and the posterior predictive distribution $p_p(x) := \int p(x|\theta)q_\psi(\theta|x_o)d\theta$, which can be seen as a blurred version of the true likelihood, using the formula:

$$ED(p_p, p_{\theta^*}) = 2\mathbb{E}_{x \sim p_p, y \sim p_{\theta^*}} |x - y|_2^\beta - \mathbb{E}_{p_p, p_p} |x - x'|_2^\beta - \mathbb{E}_{p_{\theta^*}, p_{\theta^*}} |y - y'|_2^\beta \quad \beta \in (0, 2)$$

Given $x_o \sim p_{\theta^*}$ and samples $(x_i)_{i=1}^n$ from p_p (which can be sampled from by sampling from the posterior $q_\psi(\theta|x_o)$ and feeding these samples to the simulator), it can be estimated (up to a constant in p_{θ^*}) using (see [2, Appendix D])

$$ES(x_o, (x_i)_{i=1}^n) = \frac{2}{n} \sum_{i=1}^n |x_i - x_o|_2^\beta - \frac{1}{n(n-1)} \sum_{i,j=1, i \neq j}^n |x_i - x_j|_2^\beta \quad \beta \in (0, 2),$$

a value called the "Energy Score" ES . Importantly [80], $ED(p_p, p_{\theta^*})$ is minimized if $p_{\theta^*} = p_p$, which notably happens if the posterior perfectly recovers θ^* .

Since the properties of ED's minimizers are only valid in expectation over $x_o \sim p(x|\theta^*)$, for a single x_o it is possible to minimize ES if the posterior concentrates around a $\tilde{\theta}$ such that $p(x|\tilde{\theta}) = \delta_{x_o}$ (provided that such a $\tilde{\theta}$ exists). However, the posterior of SUNLE (Figure 12) exhibits widespread pairwise marginals (similar to previous work on the pyloric network [78]), showing that this issue does not manifest itself in this case.

For completeness, we have added a comparison of the methods for the pyloric network using two other metrics: the median distance of [78] (which should be

used with caution, see [10]) and a kernel-based equivalent of the Energy Distance [188, Appendix D], which enjoys similar properties as the Energy Distance. SUNLE outperforms all alternatives for all 3 metrics.

V.F.2 Posterior pairplots on benchmark Problems

We report the ground truth estimated posterior pairplots on benchmark problems in Figure V.F.1. AUNLE and SUNLE exhibit satisfying mode coverage, and are able to capture complex posterior structures.

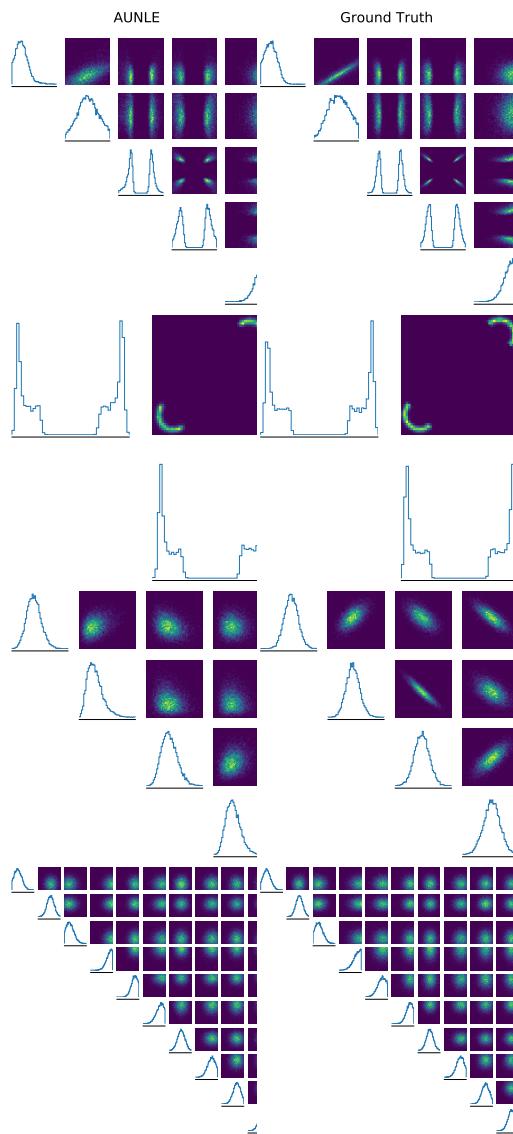


Figure V.F.1: Posterior marginal (empirical) pairplots for SUNLE’s posterior (first column), AUNLE’s posterior (second column) and the ground truth posterior for the four studied benchmark problems. Each row outlines a separate benchmark problem.

V.F.3 Manifestation of the short-run effect in UNLE

It was shown in [182] that EBMs trained by replacing the intractable expectation under the EBM with an expectation under a particle approximation obtained by running parallel runs of Langevin Dynamics initialized from random noise and updated for a fixed (and small) amount of steps can yield an EBM whose density is not proportional to the true density, but rather a generative model that can generate faithful images by running a few steps of Langevin Dynamics from random noise on it. Our design choices for both training and inference purposefully avoid this effect from manifesting in UNLE. During training, we estimate the intractable expectation using *persistent* MCMC or SMC chains, e.g. by initializing the MCMC (or SMC) algorithm of iteration k with the result of the MCMC (or SMC) algorithm at iteration $k - 1$, yielding a different training method than short-run EBMs. At inference, the posterior model is sampled from Markov Chains with a significant burn-in period, contrasting with the sampling model of short-run EBMs. Figure V.F.2 compares the density of UNLE’s posterior estimate for the Two Moons model (a 2D posterior which can be easily visualized) with the true posterior. As Figure V.F.2 shows, AUNLE and SUNLE’s posterior density match the ground truth very closely, demonstrating that UNLE’s EBM is not a short-run generative model, but a faithful *density estimator*.

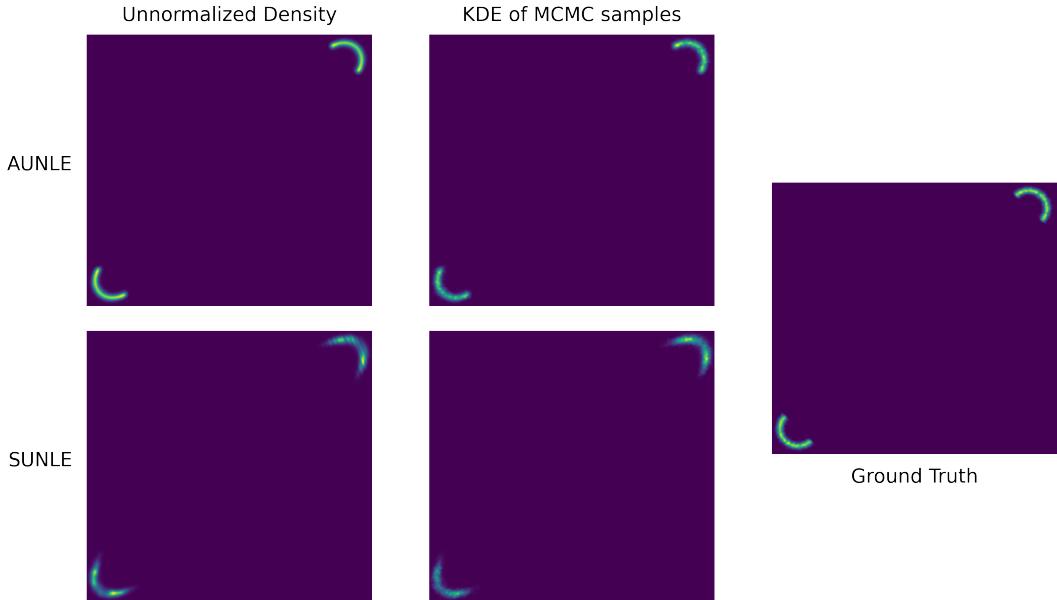


Figure V.F.2: Normalized posterior densities of AUNLE and SUNLE for the Two Moons model. Left: manually normalized posterior densities of AUNLE and SUNLE using a discretization of the posterior over a grid. Middle: kernel density estimation of the MCMC samples obtained from AUNLE’s and SUNLE’s posteriors. Right: Ground Truth posterior. AUNLE’s and SUNLE’s posterior densities closely match the true density, showing that these methods indeed learn a density estimator and a generative model [182].

V.F.4 Validating the (Z, θ) -uniformization of AUNLE’s posterior in practice

Proposition V.3.1 ensures that the normalizing constant $Z(\theta, \psi)$ present AUNLE’s posterior is independent of θ *provided that the problem is well-specified, and that $\psi = \psi^*$* , the optimum of AUNLE’s population objective. In practice, these conditions will not hold exactly, and the uniformization of AUNLE’s posterior thus only holds approximately. To assess the loss of precision associated with using a standard MCMC posterior in the context of approximate uniformization, we compare the quality of AUNLE’s posterior samples obtained using a standard MCMC sampler (which is valid only if uniformization holds), and using a doubly-intractable MCMC sampler, which handles non-uniformized posteriors. We mitigate the approximation error of doubly-intractable samplers by using a large number of steps (1000) when sampling from the likelihood using MCMC. As Figure V.F.3 shows, there is no gain in using a doubly-intractable sampler for inference in AUNLE, suggesting that the

uniformization property of AUNLE holds well in practice.

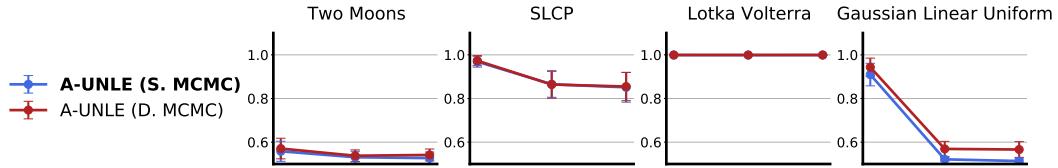


Figure V.F.3: Quality of AUNLE’s posterior samples (measured in classifier accuracy) obtained using a standard MCMC sampler (S. MCMC) and a doubly-intractable sampler (D. MCMC). The results show no gain in using a doubly-intractable sampler, justifying the use of standard samplers for AUNLE.

V.F.5 Computational Cost Analysis

Training unnormalized models using approximate likelihood is computationally intensive, as it requires running a sampler during training at each gradient step, yielding a computational cost of $O(T_1 T_2 N)$, where T_1 is the number of gradient steps, T_2 is the number of MCMC steps, and N is the number of parallel chains used to estimate the gradient.

To maximize the efficiency of training, we implement all samplers using jax [71], which provides a just-in-time compiler and an auto-vectorization primitive that generates efficient, custom parallel sampling routines. For AUNLE, we introduce a warm-started SMC approximation procedure to estimate gradients, yielding competitive performance with as little as 5 intermediate probabilities per gradient computation. For SUNLE, we warm-start the parameters of the EBM across training rounds, and warm-start the chains of the doubly-intractable sampler across inference rounds, which significantly reduces the need for burn-in steps and long training. Finally, all experiments are run on GPUs. Together, these techniques make AUNLE and SUNLE almost always the fastest methods for amortized and sequential inference, with total per-problem runtimes of 2 minutes for AUNLE and 15 minutes for SUNLE on benchmark models (which is significantly faster than NLE and SNLE on their canonical CPU setup, 160) and less than 3 hours for SUNLE on the pyloric network model (with half of this time spent simulating samples). The latter is *10 times faster than SNVI* (30 hours) on the same model. A breakdown of training, simulation and inference time is provided in Figure V.F.4. We note that (S)NLE was run on a CPU,

which is the advertised computational setting [160].

We note that the time spent performing inference is negligible for AUNLE, which uses standard MCMC for inference thanks to the tilting trick employed in its model. On the other hand, the runtime of SUNLE, which performs inference using a doubly-intractable sampler is dominated by its inference phase. This point demonstrates the computational benefits of the AUNLE’s tilting trick. Note that SUNLE performs inference in a multi-round procedure, and requires thus R training and inference phases (where R is the number of rounds), as opposed to 1 for AUNLE. We alleviate this effect by leveraging efficient warm-starting strategies for both training and inference, which to an extent amortizes these steps across rounds.

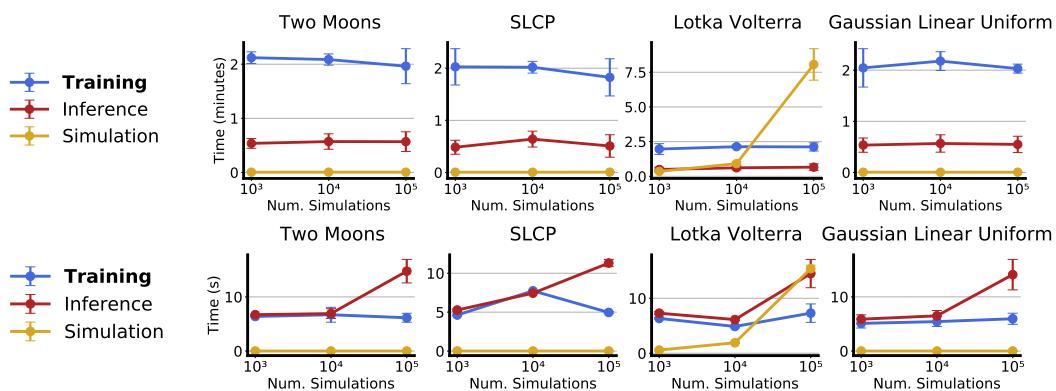


Figure V.F.4: Runtime of UNLE: Analysis and Comparisons. First row: time (in minutes) spent training, inferring, and simulating for AUNLE. Second row: time (in minutes) spent training, inferring, and simulating for SUNLE.

V.F.6 Experimental setup for SNLE and SMNLE

NLE, NRE, NPE The results reported for NLE, NRE, NPE are the one present in the SBI benchmark suite [160], which reports the performance of both NLE and SNLE on all benchmark problems studied in this paper.

V.F.7 Neuroscience Model: Details

Pairwise Marginals We provide the full pairwise marginals obtained after computing a kernel density estimation on the final posterior samples of SUNLE. We retrieve similar patterns as the one displayed in the pairwise marginals of SNVI samples. We refer to [78] for more details on the specificities of this model.

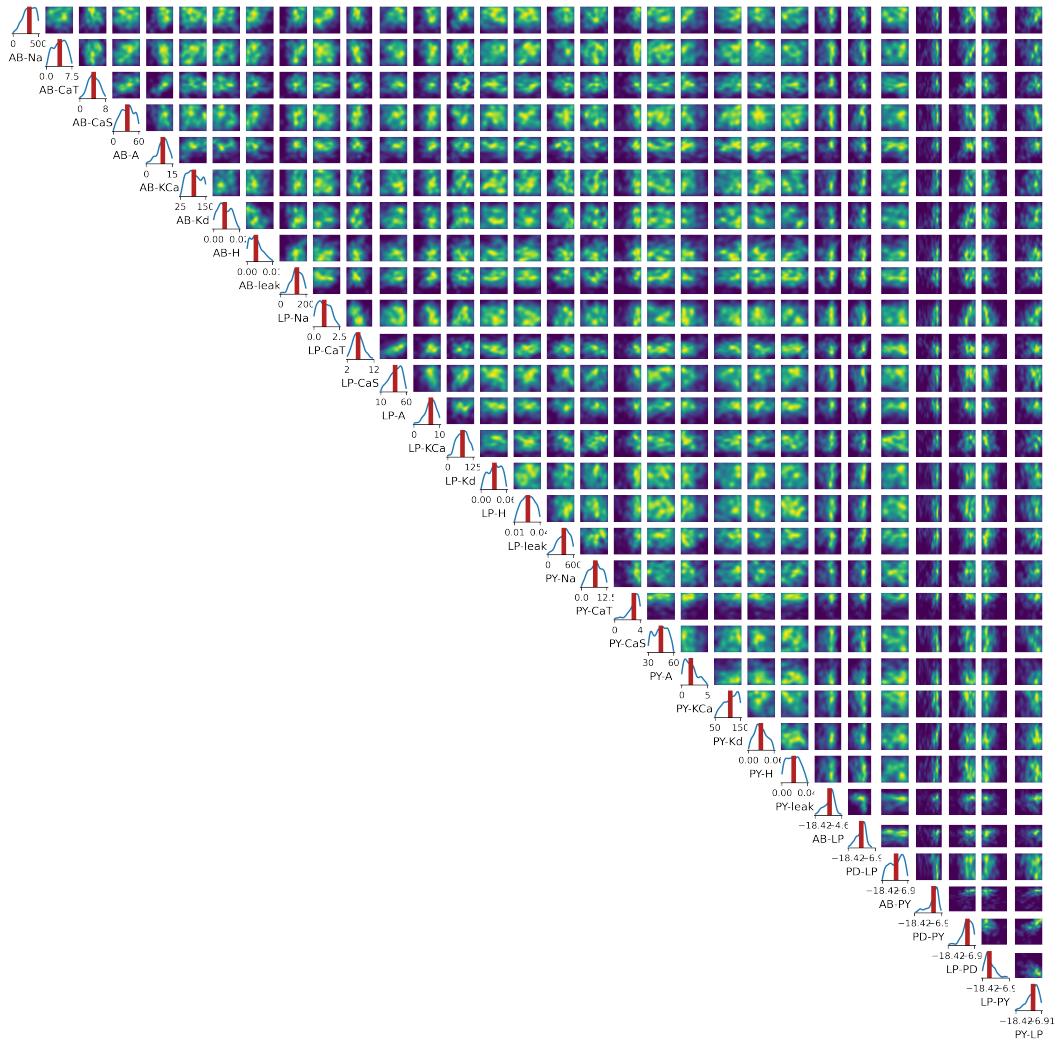


Figure V.F.5: Pairwise marginals of SUNLE's posterior estimate on the *C. borealis* simulator model.

Part II

Contributions to the evaluation of conditional density models

CHAPTER VI

Fast and Scalable Score-Based Calibration Tests

This Chapter is based on the following work:

Pierre Glaser, David Widmann, Fredrik Lindsten, and Arthur Gretton. Fast and scalable score-based kernel calibration tests. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2023. URL <https://proceedings.mlr.press/v216/glaser23a.html>

Abstract

We introduce the *Kernel Calibration Conditional Stein Discrepancy test* (KCCSD test), a non-parametric, kernel-based test for assessing the calibration of probabilistic models with well-defined scores. In contrast to previous methods, our test avoids the need for possibly expensive expectation approximations while providing control over its type-I error. We achieve these improvements by using a new family of kernels for score-based probabilities that can be estimated without probability density samples, and by using a conditional goodness-of-fit criterion for the KCCSD test’s U-statistic. We demonstrate the properties of our test on various synthetic settings.

VI.1 Introduction

Calibration is a statistical property of predictive probabilistic models that ensures that a model’s prediction matches the conditional distribution of the predicted variable given the prediction. A calibrated model expresses the uncertainty about its predictions reliably by being neither over- nor underconfident, and hence can be useful even if its accuracy is suboptimal. This property is essential in safety-critical applications such as autonomous driving. Unfortunately, empirical studies revealed that popular machine learning models such as deep neural networks tend to trade off calibration for accuracy [91]. This has lead to an increased interest in the study of calibrated models in recent years.

Calibration has been studied in the meteorological and statistical literature for many decades [e.g., 173, 52]. For a long time, research on calibration has been focused on different notions of calibration for probabilistic classifiers [e.g., 173, 52, 195, 275, 29, 176, 91, 140, 142, 141, 254, 267] and on calibration of quantiles and confidence intervals for real-valued regression problems [e.g., 106, 212, 241, 225, 67]. Regarding the calibration of classification models, different hypothesis tests have been proposed [e.g., 44, 30, 254, 267, 93, 150]. Given a predictive model and a validation dataset, these tests output whether a model is likely to be uncalibrated. The recent work of Widmann et al. [269] generalized the calibration-framework introduced for classification in [267] to (possibly multi-dimensional) continuous-valued predictive models. In particular, Widmann et al. [269] introduced a kernel-based hypothesis test for such general classes of models.

An important potential consumer of calibration tests is Bayesian inference, and in particular simulation-based inference (SBI), for which miscalibration is particularly undesirable. SBI [46] lies at the intersection of machine learning and domain sciences, and refers to the set of methods that train probabilistic models to estimate the posterior over scientific parameters of interest given some observed data. The models are trained using pairs composed of parameters drawn from a prior distribution, and their associated “synthetic” observed data, obtained by running a probabilistic program called the *simulator*, taking a parameter value as input, and that faithfully mimics the physical generative process of interest. The increasing number of use cases combined with advances in probabilistic modeling has elevated SBI to a critical role in solving complex scientific problems such as particle physics [76] and neuroscience [79, 249]. However, as discussed in [102], overconfidence in SBI models can conceal credible alternative scientific hypotheses, and result in incorrect discoveries [102], highlighting the need for principled and performant calibration tests suitable for such models.

While the theoretical framework of Widmann et al. [269] describes the calibration of any probabilistic model, applying its associated calibration test to Bayesian inference remains challenging: indeed, the test statistics require computing expectations against

the probabilistic models of interest, for reasons bearing both to the calibration setting, and to the limitations of currently available kernel-based tools for probabilistic models. Although such expectations can be computed exactly for classification models, expectations against generic probabilistic models are usually intractable and must be approximated. In cases where the models are *unnormalized*, these approximations are both computationally expensive—sometimes, prohibitively—and biased, thereby compromising theoretical guarantees of the calibration tests of Widmann et al. [269], including type-I error control.

Contributions In this paper, we introduce the kernel calibration-conditional Stein discrepancy (or KCCSD) test, a new nonparametric, score-based kernel calibration test which addresses the limitations of existing methods. The KCCSD test builds on the insight that the definition of calibration given by Vaicenavicius et al. [254] is a conditional goodness of fit property, as we remark in Section VI.3. This fact allows us to leverage the kernel conditional goodness of fit test proposed by Jitkrittum et al. [128] as the backbone of the KCCSD test. Unlike the test-statistics of Widmann et al. [269], the KCCSD test statistic does not contain *explicit* expectations against the probabilistic models; however, as in [269], it requires evaluating a kernel between probabilities densities, which in most cases of interest introduces an (intractable) expectation against the densities. To eliminate this limitation, we construct two new kernels between probability distributions that do not involve expectations against its input distributions, while remaining suitable for statistical testing. These kernels rely on a generalized version of the Fisher divergence and are of independent interest. We investigate a connection between these kernels and diffusion, akin to Stein methods, and discuss the relationships with other kernels on distributions. By using such kernels in the KCCSD test statistic, we obtain a fast and scalable calibration test that remains consistent and calibrated for unnormalized models, answering the need for such tests discussed above. We confirm in Section VI.6 the properties and benefits of the KCCSD test against alternatives on synthetic experiments.

VI.2 Background

Notation We consider probabilistic systems characterized by a joint distribution $\mathbb{P}(X, Y)$ of random variables (X, Y) taking values in $\mathcal{X} \times \mathcal{Y}$, and study *probabilistic models* $P_{\cdot| \cdot}: x \in \mathcal{X} \mapsto P_{|x}(\cdot) \in \mathcal{P}(\mathcal{Y})$ approximating the unknown conditional distribution of Y given $X = x$, $P_{|x}(\cdot) \simeq \mathbb{P}(Y \in \cdot | X = x)$. The target variable Y is typically a parameter of a probabilistic system of interest—like synaptic weights in biological neural networks—while the input variable X is observed data—like neuron voltage traces measured using electrophysiology.

VI.2.1 Calibration of Predictive Models

Calibration: General Definition A probabilistic model $P_{\cdot| \cdot}$ is called calibrated or reliable [31, 254, 269] if it satisfies

$$P_{|X} = \mathbb{P}(Y \in \cdot | P_{|X}) \quad \mathbb{P}(X)\text{-a.s.} \quad (\text{VI.1})$$

Note that this definition applies to general predictive probabilistic models, also beyond classification, and only assumes that the conditional distributions on the right-hand side exist.

Hypothesis Testing: Kernel Calibration Error There are multiple ways to test whether a given predictive probabilistic model is calibrated. In this section, we introduce the kernel-based tests of Widmann et al. [267] and their later generalization [269], since our KCCSD test is built on these approaches. These tests turn the equality between conditional distributions present in Equation (VI.1) into a more classical equality between two joint distributions. The transformation is achieved by noting that

$$\begin{aligned} P_{|X} = \mathbb{P}(Y \in \cdot | P_{|X}) &\quad \mathbb{P}(X)\text{-a.s.} \\ &\iff (P_{|X}, Y) \stackrel{d}{=} (P_{|X}, Z) \end{aligned}$$

where Z is an “auxiliary” variable such that $Z | P_{|X} \sim P_{|X}(\cdot)$. This identity was used by Widmann et al. [269] to construct an MMD-type calibration test based on the

(squared) kernel calibration error (SKCE) criterion

$$\sup_{h \in \mathcal{B}(0_{\mathcal{H}}, 1)} \mathbb{E}_{(x,y,z) \sim \mathbb{P}(X,Y,Z)} [h(P_{|x}, y) - h(P_{|x}, z)]. \quad (\text{VI.2})$$

Here, $\mathcal{B}(0_{\mathcal{H}}, 1)$ is the unit ball of a reproducing kernel Hilbert space (RKHS) \mathcal{H} of functions with positive definite kernel $k_{\mathcal{H}}: (\mathcal{P}_{|\mathcal{X}} \times \mathcal{Y})^2 \rightarrow \mathbb{R}$. As noted by Widmann et al. [269], the SKCE generalizes the (squared) kernel classification calibration error (SKCCE) defined for the special case of discrete output spaces $\mathcal{Y} = \{1, \dots, d\}$ [267], to continuous ones. Given n pairs of samples $\{(P_{|x^i}, y^i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$, Widmann et al. [269] consider the following SKCE estimator

$$\widehat{\text{SKCE}} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} G((P_{|x^i}, y^i), (P_{|x^j}, y^j)) \quad (\text{VI.3})$$

where

$$\begin{aligned} G((p, y), (p', y')) := & k((p, y), (p', y')) \\ & - \mathbb{E}_{z \sim p} k((p, z), (p', y')) \\ & - \mathbb{E}_{z' \sim p'} k((p, y), (p', z')) \\ & + \mathbb{E}_{z \sim p} \mathbb{E}_{z' \sim p'} k((p, z), (p', z')). \end{aligned} \quad (\text{VI.4})$$

For a target false rejection rate $\alpha \in (0, 1)$, the test of Widmann et al. [269] follows standard methodology in recent nonparametric testing [90, 89, 41] by rejecting the null hypothesis that the model is calibrated if $\widehat{\text{SKCE}} > \gamma_{1-\alpha}$, where $\gamma_{1-\alpha}$ denotes the $(1 - \alpha)$ -quantile of $\widehat{\text{SKCE}}$ under the null. While various methods are available to estimate this quantile, all tests experiments in this paper use a *bootstrap* approach [13]. As discussed by Widmann et al. [269], Equation (VI.3) contains two important sources of possible intractability:

First Problem The last three terms in the sum are expectations under predictions of the probabilistic model of interest. However, closed-form expressions for these expectations are only available in restricted cases, such as for classification and for Gaussian models coupled with Gaussian kernels. When these expectations are not

available, they must be approximated numerically. If the distributions $P|_X$ are given in the form of unnormalized models, this approximation requires running expensive approximation methods that often take the form of an MCMC algorithm and must be performed for every sample of $P|_X$ used to estimate the test statistic.

Second Problem The second source is the evaluation of the kernel function k . We restrict our attention to the conventional form of tensor-product type kernels $k((p, y), (p', y')) = k_P(p, p')k_Y(y, y')$ chosen in this setting. While typically many tractable choices for the kernel k_Y exist (taking as input discrete or Euclidean values), the choices for k_P , taking as input two probability distributions p and p' , are more limited and require expensive approximations methods when working with unnormalized models.

A popular approach to design kernels on distributions [236, 238] is to first embed the probability distributions in a Hilbert space \mathcal{H} using a map ϕ , and then compose it with a kernel $k_{\mathcal{H}}$ on \mathcal{H} :

$$k_P(p, p') = k_{\mathcal{H}}(\phi(p), \phi(p')).$$

Any valid kernel on \mathcal{H} , like the linear kernel $k_{\mathcal{H}}(z, z') = \langle z, z' \rangle_{\mathcal{H}}$, the Gaussian kernel $k_{\mathcal{H}}(z, z') = e^{-\|z-z'\|_{\mathcal{H}}^2}$, or the inverse multiquadric kernel $k_{\mathcal{H}}(z, z') = (1 + \|z - z'\|_{\mathcal{H}}^2)^{-1}$ can be used. In practice, the map ϕ can be set to be the *mean embedding* map to an RKHS \mathcal{H} , e.g., $\phi(\mu) = \int k_{\mathcal{H}}(z, \cdot) \mu(dz)$. Kernels $k_{\mathcal{H}}$ that are functions of $\|\phi(\mu) - \phi(\nu)\|_{\mathcal{H}}^2 := \text{MMD}^2(\mu, \nu)$, are often referred to as MMD-type kernels [165]. Other distances, like the Wasserstein distance in 1 dimension or the sliced Wasserstein distance [26] in multiple dimensions, also take this form for some choice of ϕ and \mathcal{H} , and can thus be used to construct kernels on distributions [165]. In general, however, computing $k_P(p, p')$ becomes intractable apart from special cases such as when p and p' are Gaussian distributions. While there exist finite-samples estimators for such kernels, a fast calibration estimation method based on Equation (VI.2) would require an estimator that does not require samples from p and p' .

VI.2.2 Kernel Conditional Goodness-of-Fit Test

We briefly introduce the background on goodness-of-fit methods relevant to our new test. *Conditional goodness-of-fit* (or CGOF) testing adapts the familiar goodness of fit tests to the conditional case. In particular, CGOF tests whether

$$H_0: Q|_Z = \mathbb{P}(Y \in \cdot | Z) \quad \mathbb{P}(Z)\text{-a.s.} \quad (\text{VI.5})$$

given a candidate $Q|_z$ for the conditional distribution $\mathbb{P}(Y \in \cdot | Z = z)$ and samples $\{(z^i, y^i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(Z, Y)$. This problem was studied by Jitkrittum et al. [128] for the case $\mathcal{Z} \times \mathcal{Y} \subset \mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$ and models $Q|_z$ with a differentiable, strictly positive density $f_{Q|_z}$. They proposed a kernel CGOF test for Equation (VI.5) based on the (squared) kernel conditional Stein discrepancy (KCSD)

$$D_{Q|_z}(\mathbb{P}) := \left\| \mathbb{E}_{(z,y) \sim \mathbb{P}(Z,Y)} \left[K_z \xi_{Q|_z}(y, \cdot) \right] \right\|_{\mathcal{F}_K}^2 \quad (\text{VI.6})$$

Here, \mathcal{F}_K is an $\mathcal{F}_l^{d_y}$ (e.g. $\overbrace{\mathcal{F}_l \times \cdots \times \mathcal{F}_l}^{d_y \text{ times}}$)-vector-valued RKHS with kernel $K: \mathcal{Z} \times \mathcal{Z} \rightarrow \mathcal{L}(\mathcal{F}_l^{d_y}, \mathcal{F}_l^{d_y})$, K_z is its associated linear operator on $\mathcal{F}_l^{d_y}$ with $K_z g := K(z, \cdot)g \in \mathcal{L}(\mathcal{Z}, \mathcal{F}_l^{d_y})$ for $g \in \mathcal{F}_l^{d_y}$, \mathcal{F}_l is an RKHS on \mathcal{Y} with kernel $l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and $\xi_{Q|_z}$ is the ‘‘kernelized score’’:

$$\xi_{Q|_z}(y, \cdot) = l(y, \cdot) \nabla_y \log f_{Q|_z}(y) + \nabla_y l(y, \cdot) \in \mathcal{F}_l^{d_y}.$$

We refer to Jitkrittum et al. [128, Section 2 and 3] for an intuition behind the KCSD formula, and its relationship to the more familiar Kernel Stein Discrepancy [41, 84]. Under certain assumptions, the null hypothesis in Equation (VI.5) is true if and only if $D_{Q|_z}(\mathbb{P}) = 0$. In particular, the latter will hold Jitkrittum et al. [128, Theorem 1] if \mathcal{Y} and \mathcal{Z} are compact and the kernels K and l are universal, meaning that \mathcal{F}_K (resp. \mathcal{F}_l) is dense with respect to $\mathcal{C}(\mathcal{Z}, \mathcal{F}_l^{d_y})$ (resp. $\mathcal{C}(\mathcal{Y}, \mathbb{R})$), the space of continuous functions from \mathcal{Z} to $\mathcal{F}_l^{d_y}$ (resp. \mathcal{Y} to \mathbb{R})¹. An instance of a universal \mathcal{F}^{d_y} -reproducing

¹These statements hold for noncompact \mathcal{Y}, \mathcal{Z} by replacing continuous functions by continuous functions vanishing at infinity [128, Theorem 1].

kernel is given by

$$K(z, z') = k(z, z') I_{\mathcal{F}_l^{d_y}} \quad (\text{VI.7})$$

where $I_{\mathcal{F}_l^{d_y}} \in \mathcal{L}(\mathcal{F}_l^{d_y}, \mathcal{F}_l^{d_y})$ is the identity operator and k is a real-valued universal kernel [33]. Jitkrittum et al. [128] showed that the CGOF statistic $D_{Q|_z}(\mathbb{P})$ admits an unbiased consistent estimator and used it to construct hypothesis tests of Equation (VI.5) with operator-valued kernels of the form in Equation (VI.7).

VI.3 Kernel Calibration-Conditional Stein Discrepancy

Calibration testing in the sense of Equation (VI.1) is an instance of *conditional goodness-of-fit* testing of Equation (VI.5) with input $Z = P|_X$, target Y , and models $Q|_z = z = P|_x$. Assuming that $\mathcal{Y} \subset \mathbb{R}^{d_y}$ and that distributions $P|_x$ have a differentiable, strictly positive density $f_{P|_x}$. In that case, the (squared) kernel conditional Stein discrepancy in Equation (VI.6) becomes

$$C_{P|_z}(\mathbb{P}) := \left\| \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \left[K_{P|_x} \xi_{P|_x}(y, \cdot) \right] \right\|_{\mathcal{F}_K}^2, \quad (\text{VI.8})$$

where now K is a kernel on $P|_X$. To emphasize the calibration setting, we call $C_{P|_z}$ the kernel calibration-conditional Stein discrepancy (KCCSD). Similar to the KCSD, given samples $\{P|_{x^i}, y^i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(P|_X, Y)$ and assuming a kernel K of the form in Equation (VI.7), statistic $C_{P|_z}(\mathbb{P})$ has an unbiased consistent estimator

$$\widehat{C}_{P|_z} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H((P|_{x^i}, y^i), (P|_{x^j}, y^j))$$

where

$$H((p, y), (p', y')) := k(p, p') h((p, y), (p', y')) \quad (\text{VI.9})$$

with

$$\begin{aligned} h((p, y), (p', y')) &:= l(y, y') s_p(y)^\top s_{p'}(y') \\ &+ \sum_{i=1}^{d_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(y, y') + s_p(y)^\top \nabla_{y'} l(y, y') \\ &+ s_{p'}(y')^\top \nabla_y l(y, y'), \end{aligned} \quad (\text{VI.10})$$

where $s_p(y) := \nabla_y \log f_p(y)$ (resp. $s_{p'}(y)$) is the *score* of p (resp. p'). In Section A in the supplement we discuss how the formula of $\widehat{C}_{P|}$ generalizes to operator-valued kernels that are not of the form in Equation (VI.7).

The above framing of the calibration problem conveniently avoids the first source of possible intractability present in the SKCE. For instance, for Gaussian models the test statistic can be evaluated exactly for arbitrary kernels l on \mathcal{Y} whereas a closed-form expression of the SKCE is known only in the special case where l is a Gaussian kernel.

Proposition VI.3.1 shows that the KCCSD can be viewed as a special case of the SKCE. More generally, as shown in Section B, the KCSD is a special form of the MMD.

Proposition VI.3.1 (Special case of Lemma B.1). *Under certain weak assumptions (see Lemma B.1), the KCCSD with respect to kernels $l: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and $k: P_{|\mathcal{X}} \times P_{|\mathcal{X}} \rightarrow \mathbb{R}$ is equivalent to the SKCE with kernel $H: (P_{|\mathcal{X}} \times \mathcal{Y}) \times (P_{|\mathcal{X}} \times \mathcal{Y}) \rightarrow \mathbb{R}$ defined in Equation (VI.9).*

The full testing procedure is outlined in Algorithm 8. The computations can be performed with kernels K of the form in Equation (VI.7) or more general operator-valued kernels, but crucially the method requires that K is tractable. Thus for general models of probability distributions, such as energy-based models and other unnormalized density models, it remains to address the second source of intractability, namely to construct a kernel K that can be evaluated efficiently.

Algorithm 8 CGOF Calibration Test (Tractable Kernel)

```

1: Input: Pairs  $\{(P_{|x^i}, y^i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(P|X, Y)$ 
2: Parameters: Number of data samples  $n$ , kernel  $l: \mathcal{Y}^2 \rightarrow \mathbb{R}$ , kernel  $k: (P|X)^2 \rightarrow \mathbb{R}$ , level  $\alpha$ 
3: Output: Decision on  $H_0$ : model is calibrated
4: // Estimate KCCSD using Equation VI.10 or (A.1)
5:  $\hat{C} \leftarrow \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H((P_{|x^i}, y^i), (P_{|x^j}, y^j))$ 
6: // Use e.g., bootstrap [13]
7:  $\hat{C}_\alpha \leftarrow$  approximate  $(1 - \alpha)$ -quantile of  $\hat{C}$ 
8: if  $\hat{C} < \hat{C}_\alpha$  then
9:   Fail to reject  $H_0$ 
10: else
11:   Reject  $H_0$ 
12: end if

```

VI.4 Tractable Kernels for General Unnormalized Densities

In this section, we introduce two kernels between (density-based) probability distributions that admit unbiased estimates that neither require samples from the said distributions nor require access their normalizing constant. Crucially, the properties of these new kernels allow to extend the scope of calibration tests to a more general setting, including Bayesian inference.

General Recipe As in prior work on kernels for distributions [165, 238], our proposed kernels take the form of exponentiated Hilbertian metrics

$$k(p, q) = e^{-\|\phi(p) - \phi(q)\|_H^2 / (2\sigma^2)}$$

for two probability densities p and q , defined on some set $\mathcal{X} \subset \mathbb{R}^d$, where H is some separable Hilbert space, $\phi: p \mapsto \phi(p) \in H$ is a feature map, and $\sigma > 0$ is a bandwidth parameter. Our contributions in this section consist in pairs of carefully designed ϕ and H that will allow approximating k easily.

VI.4.1 The Generalized Fisher Divergence (Kernel)

Our starting point is the *Fisher Divergence* [162, 231, 117], also known as the *Relative Fisher Information* [186], between two probability densities p and q , which is given by

$$\text{FD}(p, q) := \int_{\mathcal{X}} \|s_p(x) - s_q(x)\|^2 p(x) dx.$$

The Fisher Divergence is a convenient tool to compare unnormalized densities of the form

$$p(x) := \underbrace{\frac{f(x)}{Z_f}}_{\text{intractable}} \quad \text{where} \quad Z_f := \int_{\mathcal{X}} f(x) dx$$

as the score of p can be evaluated without knowing Z_f :

$$s_p(x) = \nabla_x (\log f(x)/Z_f) = \nabla_x \log f(x).$$

This property confers to the (squared) Fisher Divergence a tractable unbiased estimator given n i.i.d. samples $\{X^i\}_{i=1}^n$ from p , which takes the form:

$$\widehat{\text{FD}}(p, q) = \frac{1}{n} \sum_{i=1}^n \|s_p(X^i) - s_q(X^i)\|^2.$$

While the assumption ensuring access to samples from p is realistic in the unsupervised learning literature [117], or when dealing with special instances of unnormalized densities such as truncated densities $f(x) = p(x)\mathbf{1}_{x \in \mathcal{C}}$, it does not hold in the context of unnormalized models. We overcome this issue by constructing a generalized version of the Fisher Divergence:

Definition VI.4.1 (Generalized Fisher Divergence). Let p, q be two probability densities on \mathcal{X} , and v a probability measure on \mathcal{X} . The *Generalized Fisher Divergence* between p and q is defined as

$$\text{GFD}_v(p, q) := \int_{\mathcal{X}} \|s_p(x) - s_q(x)\|^2 v(dx)$$

if $\mathbb{E}_v \|s_p\|^2, \mathbb{E}_v \|s_q\|^2 < +\infty$, and $+\infty$ otherwise.

The Generalized Fisher Divergence differs from the Fisher Divergence in that the integration is performed with respect to some given base measure ν instead of p . If the support of ν covers the support of p and q , then we have that $\text{GFD}_\nu(p, q) = 0$ iff. $p = q$. Moreover, if ν can be sampled from in a tractable manner, then $\text{GFD}_\nu(p, q)$ admits a tractable estimator given samples $\{Z^i\}_{i=1}^n$ from ν of the form

$$\widehat{\text{GFD}_\nu(p, q)} = \frac{1}{n} \sum_{i=1}^n \|s_p(Z^i) - s_q(Z^i)\|^2.$$

In practice, the tractability assumption as well as the support assumption for any p , q are verified by setting ν to be a standard Gaussian distribution.

The Exponentiated-GFD Kernel Importantly, the (square root of the) Generalized Fisher Divergence is a Hilbertian metric on the space of probability densities. Indeed, for p, q such that $\mathbb{E}_\nu \|s_p\|^2, \mathbb{E}_\nu \|s_q\|^2 < +\infty$, we have that

$$\text{GFD}_\nu(p, q) = \|\phi(p) - \phi(q)\|_{\mathcal{L}_2(\nu)}^2$$

where $\phi: p \mapsto s_p(\cdot) \in \mathcal{L}_2(\nu)$ can be checked to be injective. The latter fact allows to construct a kernel K_ν on the space of probability densities based on the Generalized Fisher Divergence as follows:

Definition VI.4.2 (Exponentiated GFD Kernel). Let p, q be two probability densities on \mathcal{X} , and ν a probability measure on \mathcal{X} . The *exponentiated GFD kernel* between p and q is defined as

$$K_\nu(p, q) := e^{-\text{GFD}_\nu(p, q)/(2\sigma^2)}$$

Since the (square root of the) GFD is a Hilbertian metric, K_ν is positive definite [165], and can be estimated given samples of ν by replacing GFD_ν with its empirical counterpart. We summarize the computation method for K_ν in Algorithm 9.

Use in hypothesis testing In addition to being tractable to estimate, we show that when \mathcal{X} is compact (for instance, a bounded subset of \mathbb{R}^d), the exponentiated GFD kernels K_ν are *universal*. As a consequence, our KCCSD test, which is an instance of a KCSD test, will be able to distinguish the null-hypothesis from *any* alternative

Algorithm 9 Exponentiated GFD Kernel

```

1: Input: Probability densities  $p, q$  on  $\mathcal{X}$ 
2: Parameters: Base measure  $v$ , number of base samples  $m$ 
3: Output: Approx.  $\widehat{K}_v(p, q)$  of  $K_v(p, q)$  in Definition VI.4.2
4: for  $i = 1$  to  $m$  do
5:   Draw  $Z^i \sim v$ 
6: end for
7: return  $\exp\left(-\frac{1}{2m\sigma^2} \sum_{i=1}^m \|s_p(Z^i) - s_q(Z^i)\|^2\right)$ 

```

satisfying mild smoothness assumptions, as guaranteed by Jitkrittum et al. [128, Theorem 1].

Proposition VI.4.1. *Assume that \mathcal{X} is compact, v has full support on \mathcal{X} , and let $\mathcal{P}_{\mathcal{X}}$ be the set of twice-differentiable probability densities on \mathcal{X} equipped with the norm $\|p\|^2 = \|p\|_{\mathcal{L}_2(v)}^2 + \sum_{i=1}^d \|\partial_i p\|_{\mathcal{L}_2(v)}^2 + \sum_{i,j=1}^d \|\partial_i \partial_j p\|_{\mathcal{L}_2(v)}^2$. Then K_v is universal for any bounded subset of $\mathcal{P}_{\mathcal{X}}$.*

Proof. The proof is given in Section D.2. □

VI.4.2 The Kernelized Generalized Fisher Divergence (kernel)

While the recipe given above suffices to obtain a valid kernel on the space of probability densities, the approximation error arising from the discretization of the base measure v may scale unfavorably with the dimension of the underlying space \mathcal{X} . To address this issue, it is possible to apply a kernel-smoothing step to the GFD feature map $\phi(p)$ by composing it with an integral operator $T_{K,v}$ associated with a \mathcal{X} -vector-valued kernel K and its RKHS \mathcal{H}_K

$$T_{K,v}: f \in \mathcal{L}(\mathcal{X}, \mathbb{R}^d) \longmapsto \int_{\mathcal{X}} K_x f(x) v(dx) \in \mathcal{H}_K$$

and comparing the difference in feature map using the squared RKHS norm $\|\cdot\|_{\mathcal{H}_K}^2$. This choice of feature map yields another metric, which we call the “kernelized” GFD:

$$\text{KGFD}(p, q) := \|T_{K,v}s_p - T_{K,v}s_q\|_{\mathcal{H}_K}^2.$$

which, like the GFD, admits a tractable, unbiased estimator:

$$\frac{1}{m^2} \sum_{i,j=1}^m \langle K(Z^i, Z^j)(s_p - s_q)(Z^i), (s_p - s_q)(Z^j) \rangle_{\mathcal{X}}.$$

Since the KGFD is also a Hilbertian metric, we build upon it to construct our second proposal kernel:

Definition VI.4.3 (Exponentiated KGFD Kernel). Consider the setting of Definition VI.4.2, and let k be a bounded positive definite kernel. The *exponentiated KGFD kernel* is given by:

$$K_{K,v} := e^{-\text{KGFD}(p,q)/(2\sigma^2)}$$

For characteristic kernels K , the integral operator $T_{K,v}$ is a Hilbertian isometry between $\mathcal{L}_2(v, \mathbb{R}^d)$ and \mathcal{H}_K , making the exponentiated KGFD kernel positive definite. Additionally, $K_{K,v}$ enjoys a similar universality property as its GFD analogue, as discussed in the next proposition.

Proposition VI.4.2. *Assume that \mathcal{X} is compact, v has full-support on \mathcal{X} , and let $\mathcal{P}_{\mathcal{X}}$ be the set of twice-differentiable probability densities equipped with the norm $\|p\|^2 = \|p\|_{\mathcal{L}_2(v)}^2 + \sum_{i=1}^d \|\partial_i p\|_{\mathcal{L}_2(v)}^2 + \sum_{i,j=1}^d \|\partial_i \partial_j p\|_{\mathcal{L}_2(v)}^2$. Then $K_{K,v}$ is universal for any bounded subset of $\mathcal{P}_{\mathcal{X}}$.*

A diffusion interpretation of the KGFD In this section, we establish a relationship between the KGFD and diffusion processes [208], further anchoring the KGFD to the array of previously known divergences while opening the door for possible refinements and generalizations. Diffusion processes are well-known instances of stochastic processes $(X_t)_{t \geq 0}$ that evolve from some initial distribution μ_0 towards a target distribution p according to the differential update rule

$$dX_t = s_p(X_t) dt + \sqrt{2} dW_t, \quad X_0 \sim \mu_0,$$

where W_t is a standard Brownian motion. For any time $t \geq 0$, the probability density of X_t is the solution $\mu_{\mu_0, p}(\cdot, t)$ of the so-called Fokker-Planck equation

$$\frac{\partial \mu(x, t)}{\partial t} = \operatorname{div}(-\mu(x, t) s_p(x)) + \Delta_x \mu(x, t) \quad (\text{VI.11})$$

with initial condition $\mu(\cdot, 0) = \mu_0$. Proposition VI.4.3 establishes a link between these solutions and the KGFD:

Proposition VI.4.3 (Diffusion interpretation of the KGFD). *Let $\mu_{v, p}$ (resp. $\mu_{v, q}$) be the solution of Equation (VI.11) with initial condition v and target p (resp. q). Let k be a real-valued, twice-differentiable kernel. Then, we have that*

$$\lim_{t \rightarrow 0} \frac{1}{t} \operatorname{MMD}(\mu_{v, p}(\cdot, t), \mu_{v, q}(\cdot, t)) = \sqrt{\operatorname{KGFD}(p, q)}$$

where the MMD is w.r.t. the kernel k , and the KGFD is with respect to the matrix-valued kernel $\nabla_x \nabla_y k(x, y)$.

Proof. See Section D of the Appendix. □

Proposition VI.4.3 frames the exponentiated KGFD kernel as the $t \rightarrow 0$ limit of the kernel obtained by setting

$$\phi_t : p \longmapsto \nabla_x \log \mu_{v, p}(\cdot, t)$$

which is the score of the solution of the Fokker-Planck equation Equation (VI.11) with target p and initial measure v , and setting $H = \mathcal{H}$. Interestingly, the other limit case $t \rightarrow \infty$ recovers the exponentiated MMD kernel. Indeed, under mild conditions, the Fokker-Planck solution converges to the target and thus we have that $\lim_{t \rightarrow \infty} \phi_t(p) = p$: the feature map converges to the identity. Thus, the diffusion framework introduced above allows to recover both the KGFD and the MMD as special cases. However, while the limit $t \rightarrow 0$ and $t \rightarrow \infty$ both yield Hilbertian metrics, it is an open question whether for a given time $0 < t < \infty$, ϕ_t is also Hilbertian. A positive answer to this question would allow to construct positive definite kernels

that can possibly overcome the pitfalls of score-based tools [266, 276], while being computable in finite time.

VI.5 Fast and scalable calibration tests

The framing of the calibration testing problem of Section VI.3 alongside with the GFD-based kernels of Section VI.4 allows us to design a fast and scalable alternative to the pioneering tests of Widmann et al. [267]. The full testing procedure is outlined in Algorithm 10.

Algorithm 10 CGOF Calibration Test (GFD Kernel)

```

1: Input: Pairs  $\{(P_{|x^i}, y^i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(P_X, Y)$ 
2: Parameters: Base measure  $v$ , number of base samples  $m$ , number of data samples  $n$ , kernel  $l: \mathcal{Y}^2 \rightarrow \mathbb{R}$ , significance level  $\alpha$ 
3: Output: Decision on  $H_0$ : model is calibrated
4: for  $i = 1$  to  $m$  do
5:   Draw  $z^i \sim v$ 
6: end for
7: for  $i = 1$  to  $n - 1$  do
8:   for  $j = i + 1$  to  $n$  do
9:     //Use Algorithm 9 with base
10:    // samples  $\{z^k\}_{k=1}^m$ 
11:     $\kappa^{i,j} \leftarrow \widehat{K}_v(P_{|x^i}, P_{|x^j})$ 
12:   end for
13: end for
14: Run Algorithm 8 with kernel  $k(P_{|x^i}, P_{|x^j}) := \kappa^{i,j}$ 
```

Calibration tests as a reliability tests in Bayesian inference As one main motivation for studying calibration of generic probabilistic models is Bayesian inference, it is important to note that reliability metrics traditionally used in Bayesian inference such as conservativeness [102] differ from the notion of calibration in Equation (VI.1).

We first briefly recall the notion of posterior coverage:

Definition VI.5.1 (Conservativeness of a Bayesian model [102]). Let $P_{|x}(\cdot)$ be a conditional distribution model for $\mathbb{P}(Y \in \cdot | X = x)$, and assume that $P_{|x}$ has a density $f_{P_{|x}}$ for $\mathbb{P}(X)$ -almost every x . For level $1 - \alpha \in [0, 1]$, let $\Theta_{P_{|x}}(1 - \alpha)$ be the highest

density region of $P_{|x}$.² Then $P_{|x}$ is said to be *conservative* if

$$\mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{1}_{\Theta_{P_{|x}}(1-\alpha)}(y) \geq 1 - \alpha.$$

In the following proposition, we show that a probabilistic model that is calibrated according to Equation (VI.1) is also conservative in the sense of Hermans et al. [102], grounding the use of our tests in Bayesian inference.

Proposition VI.5.1 (Calibrated models are conservative). *If a model $P_{|x}$ is calibrated in the sense of Equation (VI.1), then it is conservative.*

The proof is given in Section C of the appendix.

VI.6 Experiments

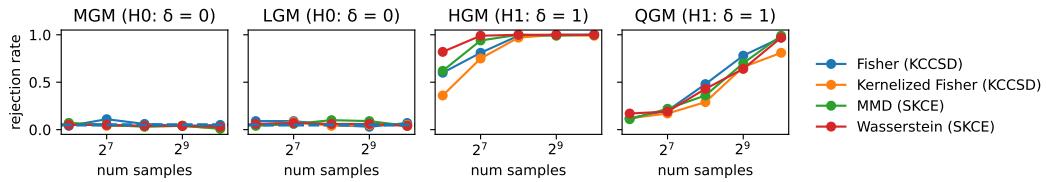


Figure VI.6.1: Rejection rates of the KCCSD and SKCE tests with a Gaussian kernel on the target space \mathcal{Y} (significance level $\alpha = 0.05$). All kernels and test statistics are evaluated exactly using closed-form expressions.

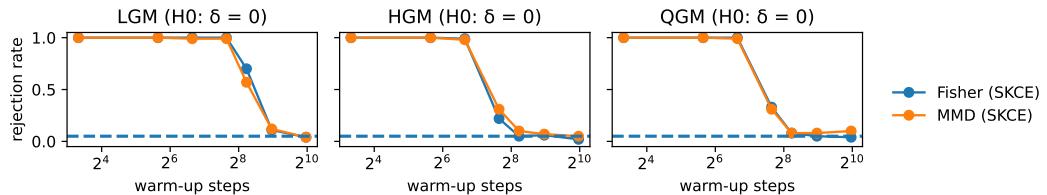


Figure VI.6.2: False rejection rates of the SKCE tests for the calibrated LGM, HMC, and QGM ($n = 200$ data points, significance level $\alpha = 0.05$). The expectations in the test statistic are estimated with 2 samples obtained with the Metropolis-adjusted Langevin algorithm (MALA) without step size tuning.

We validate the properties of our proposed calibration tests with synthetic data and

²The highest density region of a probabilistic model $P_{|x}$ with density $f_{P_{|x}}$ is defined [see, e.g., 116] by $\Theta_{P_{|x}}(1-\alpha) := \{y: f_{P_{|x}}(y) \geq c_{P_{|x}}(1-\alpha)\}$ where $c_{P_{|x}}(1-\alpha) := \sup\{c: \int \mathbb{1}_{[c,\infty)}(f_{P_{|x}}(y)) P_{|x}(dy) \geq 1 - \alpha\}$.

compare them with existing tests based on the SKCE.³ More concretely, we run KCCSD tests using either a exponentiated GFD kernel or kernelized exponentiated GFD kernel with a matrix-valued kernel of the form in Equation (VI.7) with real-valued Gaussian kernel k ; and compare them with SKCE tests using two already investigated kernels on distributions: the exponentiated MMD kernel with a Gaussian kernel on the ground space, and, for isotropic Gaussian distributions, the exponentiated Wasserstein kernel with closed-form expression

$$\begin{aligned} k_W(\mathcal{N}(\mu, \sigma^2 I_d), \mathcal{N}(\mu', \sigma'^2 I_d)) \\ = \exp(-(\|\mu - \mu'\|_2^2 + d(\sigma^2 - \sigma'^2))/(2\ell^2)). \end{aligned}$$

We set the base measure ν of the GFD and kernelized GFD kernels to be a standard Gaussian. On \mathcal{Y} , we study the Gaussian and the inverse multi-quadratic (IMQ) kernel. We repeated all experiments with 100 resampled datasets and used a wild bootstrap with 500 samples for approximating the quantiles of the test statistic with a prescribed significance level of $\alpha = 0.05$. The bandwidths of the kernels are selected with the median heuristic. A "second-order" median heuristic is used for the ground-space kernels of the KGFD and the exponentiated MMD kernel: For each pair of distributions, we compute the median distance between samples from an equally weighted mixture of these distributions (numerically for tractable cases such as Gaussian distributions and using samples otherwise), and then the bandwidth of the kernel is set to the median of these evaluations.

We repeatedly generate datasets $\{(P_{|x^i}, y^i)\}_i$ in a two-step procedure: First we sample distributions $P_{|x^i}$ and then we draw a corresponding target y^i for each $P_{|x^i}$. We compare different setups of targets Y and Gaussian distributions $P_{|X}$ with varying degree $\delta \geq 0$ of miscalibration (models are calibrated for $\delta = 0$ and miscalibrated otherwise):⁴

³The code to reproduce the experiments is available at <https://github.com/pierreglaser/kccsd>.

⁴MGM is adapted from a model used by Widmann et al. [269], and LGM, HGM, and QGM were used by Jitkrittum et al. [128].

Mean Gaussian Model (MGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}^5$, $\mathbb{P}(X) = \mathcal{N}(0, I_5)$, $\mathbb{P}(Y|X = x) = \mathcal{N}(x, I_5)$, and $P_{|x} = \mathcal{N}(x + \delta c, I_5)$ for $c \in \{\mathbf{1}_5, e_1\} \subset \mathbb{R}^5$ (miscalibration of all dimensions or only the first one).

Linear Gaussian Model (LGM) Here $\mathcal{X} = \mathbb{R}^5$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_5)$, and $P_{|x} = \mathcal{N}(\delta + \sum_{i=1}^5 ix_i, 1)$.

Heteroscedastic Gaussian Model (HGM) Here $\mathcal{X} = \mathbb{R}^3$, $\mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{N}(0, I_3)$, $\mathbb{P}(Y|X = x) = \mathcal{N}(m(x), 1)$, and $P_{|x} = \mathcal{N}(m(x), \sigma^2(x))$ with $m(x) = \sum_{i=1}^3 x_i$ and $\sigma^2(x) = 1 + 10\delta \exp(-\|x - c\|_2^2/(2 \cdot 0.8^2))$ for $c = 2/3 \mathbf{1}_3$.

Quadratic Gaussian Model (QGM) Here $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, $\mathbb{P}(X) = \mathcal{U}(-2, 2)$, $\mathbb{P}(Y|X = x) = \mathcal{N}(0.1x^2 + x + 1, 1)$, and $P_{|x} = \mathcal{N}(0.1(1 - \delta)x^2 + x + 1, 1)$.

Figure VI.6.1 demonstrates that the proposed KCCSD tests are calibrated: The false rejection rates (type I errors) of the calibrated MGM and LGM do not exceed the set significance level, apart from sampling noise. Figures F.1 and F.7 in the supplementary material confirm empirically that this is the case also when we approximate the Fisher and MMD kernels using samples.

Moreover, we see in Figure VI.6.1 that for the miscalibrated HGM the SKCE tests exhibit larger rejection rates, and hence test power, than the KCCSD tests in the small sample regime, regardless of the kernel choice. This specific setting with Gaussian distributions and a Gaussian kernel on the target space \mathcal{Y} is favourable to the SKCE test as both its test statistic, as well as the exponentiated MMD or Wasserstein kernel evaluations are available in closed-form. In such analytical scenarios we expect the score-based KCCSD tests to perform worse [266, 276]. However, the KCCSD tests present themselves as a practically useful alternative even in this example: For the miscalibrated HGM their rejection rates are close to 100% with ≥ 256 data points, and for the miscalibrated QGM they show very similar performance as the SKCE tests. Overall, as expected, we see in Figure VI.6.1 that for all studied tests rejection rates for the miscalibrated models increases with increasing number of samples.

One main advantage of the KCCSD over the SKCE is that it has first-class support for unnormalized models for which only the score function is available: In contrast to the SKCE its test statistic only involves scores but no expectations. In principle,

for unnormalized models these expectations in the test statistic of the SKCE can be approximated with, e.g., MCMC sampling. However, Figure VI.6.2 shows that there is a major caveat: If the MCMC method is not tuned sufficiently well (e.g., if the chain is too short or the proposal step size is not tuned properly), it might return biased samples which causes the SKCE tests to be miscalibrated. On the other hand, increasing the number of MCMC samples increases the computational advantage of the KCCSD even more.

Another difference between the KCCSD and SKCE is highlighted in Figures F.1 and F.2: The number of combinations of kernels for which the test statistic can be evaluated exactly is smaller for the SKCE (in these Gaussian examples, it requires Gaussian kernels on the target space).

One limitation of the (kernelized) exponentiated GFD Kernel is that it necessitates setting an additional hyperparameter: the base measure ν , which weights the score differences between its two input distributions p and q at all points of the ground space \mathcal{X} . While our experiments have set ν to be a Gaussian measure in order to obtain closed-form expressions for Gaussian p, q , other choices may be more adequate depending on the problem at hand. For instance, when p and q are posterior models for a given prior π , we hypothesize that setting ν to π constitutes a better default choice.

VI.7 Conclusion

In this paper, we introduced the Kernel Calibration Conditional Stein Discrepancy test, a fast and reliable alternative to prior calibration tests for general, density-based probabilistic models, thereby addressing an important need in the Bayesian inference community. In doing so, we introduced kernels for density-based inputs, which we believe are of independent interest and could be used in other domains such as distribution regression [238] or meta-learning [55]. Moreover, while the set of experiments conducted in this paper focused on “offline” calibration testing, its low computational cost opens the door to promising new use cases. One particularly interesting avenue would consist in using the KCCSD test criterion as a regularizer

directly within the training procedure of a probabilistic model, allowing not only to detect miscalibration but also to prevent it in the first place. We look forward to seeing extensions and applications of the tools introduced in this paper.

Acknowledgements

This research was financially supported by the Centre for Interdisciplinary Mathematics (CIM) at Uppsala University, Sweden, by the projects *NewLEADS - New Directions in Learning Dynamical Systems* (contract number: 621-2016-06079) and *Handling Uncertainty in Machine Learning Systems* (contract number: 2020-04122), funded by the Swedish Research Council, by the *Kjell och Märta Beijer Foundation*, by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and by the Excellence Center at Linköping-Lund in Information Technology (ELLIIT). Pierre Glaser and Arthur Gretton acknowledge support from the Gatsby Charitable Foundation.

Appendix

VI.A Conditional Goodness-of-Fit: General Operator-Valued Kernel

Assume that

- kernel $l \in \mathcal{C}^2(\mathcal{Y} \times \mathcal{Y}, \mathbb{R})$,
- densities $P_{|x} \in C^1(\mathcal{Y}, \mathbb{R})$ for $\mathbb{P}(X)$ -almost all x , and that
- $\mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \left\| K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right\|_{\mathcal{F}_K} < \infty$.

Due to the Bochner integrability of $(x, y) \mapsto K_{P_{|x}} \xi_{P_{|x}}(y, \cdot)$ expectation and inner product commute [see 12, Definition A.5.20], and hence we have

$$\begin{aligned} C_{P_{| \cdot}}(\mathbb{P}) &= \left\| \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \left[K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right] \right\|_{\mathcal{F}_K}^2 \\ &= \left\langle \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \left[K_{P_{|x}} \xi_{P_{|x}}(y, \cdot) \right], \mathbb{E}_{(x',y') \sim \mathbb{P}(X,Y)} \left[K_{P_{|x'}} \xi_{P_{|x'}}(y', \cdot) \right] \right\rangle_{\mathcal{F}_K} \\ &= \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{E}_{(x',y') \sim \mathbb{P}(X,Y)} \left\langle K_{P_{|x}} \xi_{P_{|x}}(y, \cdot), K_{P_{|x'}} \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_K} \\ &= \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{E}_{(x',y') \sim \mathbb{P}(X,Y)} \left\langle K_{P_{|x'}}^* K_{P_{|x}} \xi_{P_{|x}}(y, \cdot), \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}}, \end{aligned}$$

where $K_{P_{|x'}}^*$ is the adjoint of $K_{P_{|x'}}$. The reproducing property implies $K_{P_{|x'}}^* K_{P_{|x}} = K(P_{|x}, P_{|x'})$, and therefore we get

$$\begin{aligned} C_{P_{| \cdot}}(\mathbb{P}) &= \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{E}_{(x',y') \sim \mathbb{P}(X,Y)} \left\langle K(P_{|x}, P_{|x'}) \xi_{P_{|x}}(y, \cdot), \xi_{P_{|x'}}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}} \\ &= \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{E}_{(x',y') \sim \mathbb{P}(X,Y)} H((P_{|x}, y), (P_{|x'}, y')) \end{aligned}$$

where

$$\begin{aligned} H((p,y), (p',y')) &:= \left\langle K(p, p') \xi_p(y, \cdot), \xi_{p'}(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}} \\ &= \left\langle K(p, p') \xi_p(y, \cdot), l(y', \cdot) \nabla_{y'} \log f_{p'}(y') + \nabla_{y'} l(y', \cdot) \right\rangle_{\mathcal{F}_l^{d_y}}. \end{aligned}$$

For $i \in \{1, \dots, d_y\}$, let $\text{proj}_i: \mathcal{F}_l^{d_y} \rightarrow \mathcal{F}_l$ be the projection map to the i th subspace of the product space $\mathcal{F}_l^{d_y}$, and similarly let $\iota_i: \mathcal{F}_l \rightarrow \mathcal{F}_l^{d_y}$ be the embedding of \mathcal{F}_l in the

ith subspace of $\mathcal{F}_l^{d_y}$ via $x \mapsto (0, \dots, 0, x, 0, \dots, 0)$. Then we can write

$$\begin{aligned} H((p, y), (p', y')) &= \sum_{i=1}^{d_y} \left\langle \text{proj}_i K(p, p') \xi_p(y, \cdot), l(y', \cdot) \frac{\partial}{\partial y'_i} \log f_{p'}(y') + \frac{\partial}{\partial y'_i} l(y', \cdot) \right\rangle_{\mathcal{F}_l} \\ &= \sum_{i=1}^{d_y} \left[(\text{proj}_i K(p, p') \xi_p(y, \cdot))(y') \frac{\partial}{\partial y'_i} \log f_{p'}(y') + \frac{\partial}{\partial y'_i} (\text{proj}_i K(p, p') \xi_p(y, \cdot))(y') \right]. \end{aligned}$$

Since $K(p, p') \in \mathcal{L}(\mathcal{F}_l^{d_y})$ is a linear operator, we have

$$K(p, p') \xi_p(y, \cdot) = K(p, p')(l(y, \cdot) \nabla_y \log f_p(y)) + K(p, p') \nabla_y l(y, \cdot).$$

For $1 \leq i, j \leq d_y$, define $K_{i,j}(p, p'): \mathcal{F}_l \rightarrow \mathcal{F}_l$ as the continuous linear operator

$$K_{i,j}(p, p') := \text{proj}_i K(p, p') \iota_j.$$

Thus we have

$$\text{proj}_i K(p, p') \xi_p(y, \cdot) = \sum_{j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] K_{i,j}(p, p') l(y, \cdot) + \sum_{j=1}^{d_y} \frac{\partial}{\partial y_j} K_{i,j}(p, p') l(y, \cdot),$$

and therefore

$$(\text{proj}_i K(p, p') \xi_p(y, \cdot))(y') = \sum_{j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] (K_{i,j}(p, p') l(y, \cdot))(y') + \sum_{j=1}^{d_y} \frac{\partial}{\partial y_j} (K_{i,j}(p, p') l(y, \cdot))(y').$$

Due to the differentiability of kernel l we can interchange inner product and differentiation [12, Lemma 4.34], and thus we obtain

$$\begin{aligned}
 H((p, y), (p', y')) &= \sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] \left[\frac{\partial}{\partial y'_i} \log f_{p'}(y') \right] (K_{i,j}(p, p') l(y, \cdot))(y') \\
 &\quad + \sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y'_i} \log f_{p'}(y') \right] \frac{\partial}{\partial y_j} (K_{i,j}(p, p') l(y, \cdot))(y') \\
 &\quad + \sum_{i,j=1}^{d_y} \left[\frac{\partial}{\partial y_j} \log f_p(y) \right] \frac{\partial}{\partial y'_i} (K_{i,j}(p, p') l(y, \cdot))(y') \\
 &\quad + \sum_{i,j=1}^{d_y} \frac{\partial}{\partial y'_i} \frac{\partial}{\partial y_j} (K_{i,j}(p, p') l(y, \cdot))(y'),
 \end{aligned}$$

Define $A: (P_{|\mathcal{X}} \times \mathcal{Y})^2 \rightarrow \mathbb{R}^{d_y \times d_y}$ by

$$[A((p, y), (p', y'))]_{i,j} := (K_{i,j}(p, p') l(y, \cdot))(y') \quad (1 \leq i, j \leq d_y).$$

Thus we obtain

$$H((p, y), (p', y')) = (s_{p'}(y') + \nabla_{y'})^\top A((p, y), (p', y')) (s_p(y) + \nabla_y), \quad (\text{VI.12})$$

where for $x, x' \in \mathbb{R}^d, M(x, x') \in \mathbb{R}^{d \times d}$ we use the notation

$$\nabla_x^\top M(x, x') = \begin{bmatrix} \nabla_x^\top [M(x, x')]_{:,1} & \cdots & \nabla_x^\top [M(x, x')]_{:,d} \end{bmatrix} = \begin{bmatrix} \text{div}_x[M(x, x')]_{:,1} & \cdots & \text{div}_x[M(x, x')]_{:,d} \end{bmatrix},$$

and similarly

$$M(x, x') \nabla_{x'} = \left(\nabla_{x'}^\top M(x, x')^\top \right)^\top = \left[\text{div}_{x'}[M(x, x')]_{1,:} \quad \cdots \quad \text{div}_{x'}[M(x, x')]_{d,:} \right]^\top$$

and

$$\nabla_x^\top M(x, x') \nabla_{x'} = \nabla_x^\top (M(x, x') \nabla_{x'}^\top) = \sum_{i,j=1}^d \frac{\partial^2}{\partial x_i \partial x'_j} [M(x, x')]_{i,j}.$$

Thus, given samples $\{(P_{|x^i}, y^i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}(P_{|X}, Y)$, an unbiased estimator of statistic

$C_{P|}(\mathbb{P})$ is

$$\widehat{C_{P|}} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} H((P_{|x^i}, y^i), (P_{|x^j}, y^j)),$$

where H is given by Equation (VI.12).

If kernel K is of the form in ??, we recover the simpler formula in ???. In this case $A((p, y), (p', y')) = k(p, p')l(y, y')I_{d_y} \in \mathbb{R}^{d_y \times d_y}$, i.e., A is a scaled identity matrix.

VI.B KCCSD as a special case of SKCE

We prove the following general lemma that establishes the KCSD as a special case of the MMD. Then ?? follows immediately by considering random variables $Z = P_{|X}$ and Y , and models $Q_{|z} = z = P_{|x}$.

Lemma VI.B.1 (KCSD as a special case of the MMD). *Let $Q_{|z}$ be models of the conditional distributions $\mathbb{P}(Y \in \cdot | Z = z)$. Moreover, we assume that*

- $Q_{|z}$ has a density $f_{Q_{|z}} \in C^1(\mathcal{Y}, \mathbb{R})$ for $\mathbb{P}(Z)$ -almost all z ,
- kernel $l \in C^2(\mathcal{Y} \times \mathcal{Y}, \mathbb{R})$,
- $\mathbb{E}_{(z,y) \sim \mathbb{P}(Z,Y)} \left\| K_z \xi_{Q_{|z}}(y, \cdot) \right\|_{\mathcal{F}_K} < \infty$, and
- $\oint_{\partial \mathcal{Y}} l(y, y') f_{Q_{|z}}(y) n(y) dS(y') = 0$ and $\oint_{\partial \mathcal{Y}} \nabla_y l(y, y') f_{Q_{|z}}(y') n(y') dS(y') = 0$ for $\mathbb{P}(Z)$ -almost all z ,

where $n(y)$ is the unit vector normal to the boundary $\partial \mathcal{Y}$ of \mathcal{Y} at $y \in \mathcal{Y}$.⁵

Then

$$D_{Q|}(\mathbb{P}) = \text{MMD}_{k_{Q|}}^2(\mathbb{P}(Z, Y), \mathbb{P}_{Q|}(Z, Y))$$

where we define distribution $\mathbb{P}_{Q|}$ by

$$\mathbb{P}_{Q|}(Z \in A, Y \in B) := \int_A Q_{|z}(Y \in B) \mathbb{P}(Z \in dz)$$

⁵These assumptions are not restrictive in practice since they are satisfied if the conditions of [128, Theorem 1] hold which are required to ensure that $D_{Q|}(\mathbb{P}) = 0$ if and only if $Q_{|Z}(\cdot) = \mathbb{P}(Y \in \cdot | Z)$ $\mathbb{P}(Z)$ -almost surely.

and kernel $k_{Q|} : (\mathcal{Z} \times \mathcal{Y}) \times (\mathcal{Z} \times \mathcal{Y}) \rightarrow \mathbb{R}$ as

$$k_{Q|}((z, y), (z', y')) := (s_{Q|z'}(y') + \nabla_{y'})^\top A((z, y), (z', y'))(s_{Q|z}(y) + \nabla_y),$$

using the same notation as in Section VI.A and similarly defining $A((z, y), (z', y')) \in \mathbb{R}^{d_y \times d_y}$ by

$$[A((z, y), (z', y'))]_{i,j} := (K_{i,j}(z, z')l(y, \cdot))(y') \quad (1 \leq i, j \leq d_y).$$

If K is of the form $k(\cdot, \cdot)I_{\mathcal{F}_l^{d_y}}$, function A simplifies to

$$A((z, y), (z', y')) = k(z, z')l(y, y')I_{d_y}$$

and kernel $k_{Q|}$ is given by

$$\begin{aligned} & k_{Q|}((z, y), (z', y')) \\ &= k(z, z') \left[l(y, y')s_{Q|z}(y)^\top s_{Q|z'}(y') + s_{Q|z}(y)^\top \nabla_{y'}l(y, y') + s_{Q|z'}(y')^\top \nabla_y l(y, y') + \sum_{i=1}^{d_y} \frac{\partial^2}{\partial y_i \partial y'_i} l(y, y') \right]. \end{aligned}$$

Proof. From a similar calculation as in Section VI.A [cf. 128, Section A.2] we obtain that

$$k_{Q|}((z, y), (z', y')) = \left\langle K_z \xi_{Q|z}(y, \cdot), K_{z'} \xi_{Q|z'}(y', \cdot) \right\rangle_{\mathcal{F}_K}.$$

Thus $k_{Q|}$ is an inner product of the features of (z, y) and (z', y') given by the feature map $(z, y) \mapsto K_z \xi_{Q|z}(y, \cdot) \in \mathcal{F}_K$, and therefore $k_{Q|}$ is a positive-definite kernel. Moreover, from our assumption we obtain

$$\mathbb{E}_{(z, y) \sim \mathbb{P}(Z, Y)} |k_{Q|}((z, y), (z, y))|^{1/2} = \mathbb{E}_{(z, y) \sim \mathbb{P}(Z, Y)} \left\| K_z \xi_{Q|z}(y, \cdot) \right\|_{\mathcal{F}_K} < \infty.$$

Thus the mean embedding $\mu_{\mathbb{P}(Z, Y)} \in \mathcal{F}_K$ of $\mathbb{P}(Z, Y)$ exists [90, Lemma 3].

Due to the Bochner integrability of $(z, y) \mapsto K_z \xi_{Q|z}(y, \cdot)$ expectation and inner product

commute [see 12, Definition A.5.20], and hence we have

$$\begin{aligned}\mathbb{E}_{(z,y) \sim \mathbb{P}_{Q|}(Z,Y)} \mathbb{E}_{(z',y') \sim \mathbb{P}_{Q|}(Z,Y)} k_{Q|}((z,y), (z',y')) &= \left\| \mathbb{E}_{(z,y) \sim \mathbb{P}_{Q|}(Z,Y)} K_z \xi_{Q|z}(y, \cdot) \right\|_{\mathcal{F}_K}^2 \\ &= \left\| \mathbb{E}_{z \sim \mathbb{P}(Z)} \mathbb{E}_{y \sim Q|z} K_z \xi_{Q|z}(y, \cdot) \right\|_{\mathcal{F}_K}^2 \\ &= \left\| \mathbb{E}_{z \sim \mathbb{P}(Z)} K_z \mathbb{E}_{y \sim Q|z} \xi_{Q|z}(y, \cdot) \right\|_{\mathcal{F}_K}^2.\end{aligned}$$

Due to the last assumption [41, Lemma 5.1] we know that

$$\mathbb{E}_{y \sim Q|z} \xi_{Q|z}(y, \cdot) = 0,$$

which implies

$$\mathbb{E}_{(z,y) \sim \mathbb{P}_{Q|}(Z,Y)} \mathbb{E}_{(z',y') \sim \mathbb{P}_{Q|}(Z,Y)} k_{Q|}((z,y), (z',y')) = 0.$$

Thus the mean embedding $\mu_{\mathbb{P}_{Q|}(Z,Y)} \in \mathcal{F}_K$ of $\mathbb{P}_{Q|}(Z,Y)$ exists and satisfies $\|\mu_{\mathbb{P}_{Q|}(Z,Y)}\|_{\mathcal{F}_K}^2 = 0$, and hence $\mu_{\mathbb{P}_{Q|}(Z,Y)} = 0$. We obtain [90, Lemma 4] that

$$\begin{aligned}\text{MMD}_{k_{Q|}}^2(\mathbb{P}(Z,Y), \mathbb{P}_{Q|}(Z,Y)) &= \|\mu_{\mathbb{P}(Z,Y)} - \mu_{\mathbb{P}_{Q|}(Z,Y)}\|_{\mathcal{F}_K}^2 \\ &= \|\mu_{\mathbb{P}(Z,Y)}\|_{\mathcal{F}_K}^2 \\ &= \mathbb{E}_{(z,y) \sim \mathbb{P}(Z,Y)} \mathbb{E}_{(z',y') \sim \mathbb{P}(Z,Y)} k_{Q|}((z,y), (z',y')) \\ &= \mathbb{E}_{(z,y) \sim \mathbb{P}(Z,Y)} \mathbb{E}_{(z',y') \sim \mathbb{P}(Z,Y)} \left\langle K_z \xi_{Q|z}(y, \cdot), K_{z'} \xi_{Q|z'}(y', \cdot) \right\rangle_{\mathcal{F}_K} \\ &= D_{Q|}(\mathbb{P}),\end{aligned}$$

where the last equality follows from [128, Section A.2]. \square

VI.C Calibration implies expected coverage

We show that the sense of calibration employed by our tests implies posterior coverage in the sense of Hermans et al. [102]. Again let us note $P_{|x}(\cdot)$ for a model of the conditional distribution $\mathbb{P}(Y \in \cdot | X = x)$. Moreover, we assume that $P_{|x}$ has a density $f_{P_{|x}}$ for $\mathbb{P}(X)$ -almost every x .

For level $1 - \alpha \in [0, 1]$, let $\Theta_{P|x}(1 - \alpha)$ be the highest density region of a probabilistic model $P|x$ with density $f_{P|x}$. It is defined [see, e.g., 116] by

$$\Theta_{P|x}(1 - \alpha) := \left\{ y : f_{P|x}(y) \geq c_{P|x}(1 - \alpha) \right\}$$

where

$$c_{P|x}(1 - \alpha) := \sup \left\{ c : \int_{\{\tilde{y} : f_{P|x}(\tilde{y}) \geq c\}} P|x(dy) \geq 1 - \alpha \right\}.$$

Hence, by definition [see, e.g., 102]

$$\mathbb{E}_{y \sim P|x} \mathbb{1}\{y \in \Theta_{P|x}(1 - \alpha)\} = \int_{\Theta_{P|x}(1 - \alpha)} P|x(dy) \geq 1 - \alpha.$$

Assume that model $P|.$ is calibrated. By definition, it satisfies

$$\mathbb{P}(Y \in \cdot | P|X) = P|X \quad \mathbb{P}(X)\text{-almost surely.}$$

Hence, for all $\alpha \in [0, 1]$, we obtain

$$\begin{aligned} \mathbb{E}_{(x,y) \sim \mathbb{P}(X,Y)} \mathbb{1}\{y \in \Theta_{P|x}(1 - \alpha)\} &= \mathbb{E}_{(P|x,y) \sim \mathbb{P}(P|X,Y)} \mathbb{1}\{y \in \Theta_{P|x}(1 - \alpha)\} \\ &= \mathbb{E}_{P|x \sim \mathbb{P}(P|X)} \mathbb{E}_{y \sim P|x} \mathbb{1}\{y \in \Theta_{P|x}(1 - \alpha)\} \\ &\geq \mathbb{E}_{P|x \sim \mathbb{P}(P|X)} [1 - \alpha] \\ &= 1 - \alpha. \end{aligned}$$

Thus model $P|.$ has expected coverage for all $\alpha \in [0, 1]$.

VI.D Diffusion-Limit and Universality

VI.D.1 Fisher divergence as a diffusion limit

We recall that for a map f and a measure μ , the push-forward measure of μ by f , noted $f_\# \mu$, is the measure on the image space of f which verifies, for any measurable function g

$$\int g(x) f_\# \mu(dx) = \int g(f(x)) \mu(dx).$$

To prove the differential inequality linking the MMD and the KGFD, we rely on the following reformulation of the Fokker-Planck equation:

$$\begin{aligned}\frac{\partial \mu(x,t)}{\partial t} &= \operatorname{div}_x(-\mu(x,t)s_p(x)) + \Delta_x \mu(x,t) \\ &= \operatorname{div}_x(-\mu(x,t)s_p(x)) + \operatorname{div}_x \nabla_x \mu(x,t) \\ &= \operatorname{div}_x(-\mu(x,t)s_p(x)) + \operatorname{div}_x(\mu(x,t)\nabla_x \log \mu(x,t)) \\ &= \operatorname{div}_x(-\mu(x,t)(s_p(x) - \nabla_x \log \mu(x,t))).\end{aligned}$$

We remark that since the density $\mu(x,t)$ is twice differentiable in x and differentiable in t [129], this equation holds in the strong sense, and not only in the sense of distributions. Because of that, one has

$$\partial_t \mu(x,t) = \lim_{\Delta \rightarrow 0} \frac{\mu(x,t+\Delta) - \mu(x,t)}{\Delta}.$$

Let us consider an RKHS \mathcal{H} with kernel k , and let $h \in \mathcal{H}$. Let us define $m_t(x) := m(x,t) := \mu_{v,p}(x,t) - \mu_{v,q}(x,t)$ and we note $\text{MMD}(m_t)$ the function given by

$$\text{MMD}(m_t) = \left[\iint k(x,y)m_t(x)m_t(y) dx dy \right]^{1/2} = \text{MMD}(\mu_{v,p}(\cdot,t), \mu_{v,q}(\cdot,t)).$$

To show that $\lim_{t \rightarrow 0} \frac{d}{dt} \text{MMD}(m_t) = \text{KGFD}(p,q)$, we first analyze the differential properties of the easier to handle MMD^2 and complete the proof using a chain rule argument. The first variation (also called Gateaux Derivative) of $m \mapsto \text{MMD}^2(m)$ is a linear functional on the space of functions

$$\left\{ f - g \mid f, g: \mathcal{X} \times [0, \infty) \rightarrow \mathbb{R} \quad \text{with} \quad \forall t \geq 0: \int_{\mathcal{X}} f(x,t) dx = \int_{\mathcal{X}} g(x,t) dx = 1 \right\},$$

given by

$$\frac{\delta \text{MMD}^2}{\delta m}: f \mapsto \int 2k(x,y)m_t(x)f(y) dx dy.$$

Using the chain rule for Gateaux derivatives, we have that

$$\begin{aligned}\frac{d\text{MMD}^2(m)}{dt} &= \frac{d\text{MMD}^2}{dm}(m) \frac{dm}{dt} \\ &= \int 2k(x, y)m_t(x) \frac{dm}{dt}(y) dx dy.\end{aligned}$$

From the Fokker-Planck Equation, we have that

$$\begin{aligned}\frac{dm}{dt} &= \partial_t \mu_{v,p} - \partial_t \mu_{v,q} \\ &= \text{div}_x(\mu_{v,p} \nabla_x \log \frac{p}{\mu_{v,p}}) - \text{div}_x(\mu_{v,q} \nabla_x \log \frac{q}{\mu_{v,q}}) \\ &= \text{div}_x(v \nabla_x \log \frac{p}{v}) - \text{div}_x(v \nabla_x \log \frac{q}{v}) + o(1) \\ &= \text{div}_x(v \nabla_x \log \frac{p}{q}) + o(1)\end{aligned}$$

Plugging the last equation in the chain rule, we have:

$$\begin{aligned}\frac{d\text{MMD}^2(m)}{dt} &= \int 2m_t(x) \text{div}_y v(y) \nabla_y \log \frac{p}{q}(y) k(x, y) dx dy + o(1) \\ &= \int 2m_t(x) \left\langle \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle dx dy + o(1).\end{aligned}$$

Similarly, since $m_0 = \mu_{v,p}(\cdot, 0) - \mu_{v,q}(\cdot, 0) = v - v = 0$, we have $m_t(x) = t \partial_t m(x, 0) + o_x(t)$. The calculation follows as:

$$\begin{aligned}\frac{d\text{MMD}^2(m)}{dt} &= \int 2t \times \partial_t m(x, t) \left\langle \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle dx dy + o(t) \\ &= \int 2t \times \text{div}_x v(x) \nabla_x \log \frac{p}{q}(x) \left\langle \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle dx dy + o(t) \\ &= \int 2t \times \left\langle v(x) \nabla_x \log \frac{p}{q}(x), \nabla_x \left\langle \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle \right\rangle dx dy + o(t) \\ &= \int 2t \times \left\langle v(x) \nabla_x \log \frac{p}{q}(x), \nabla_x \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle dx dy + o(t).\end{aligned}$$

To get rid of the degenerate scaling as $t \rightarrow 0$, we now focus on (the derivative of) $\sqrt{\text{MMD}^2(m_t)}$ as $t \rightarrow 0$. Notice that since $\text{MMD}(m_0) = 0$, the derivative of $\sqrt{\text{MMD}^2(m_t)}$ does not exist a priori for $t = 0$: we consider instead

$\frac{d}{dt} \sqrt{\text{MMD}^2(m_t)} \Big|_{t=t}$, and extend it by continuity by setting $t \rightarrow 0$. We have:

$$\frac{d\sqrt{\text{MMD}^2(m_t)}}{dt} = \frac{1}{2\sqrt{\text{MMD}^2(m_t)}} \frac{d\text{MMD}^2(m_t)}{dt}.$$

As

$$\text{MMD}^2(m_t) = \int k(x, y) m_t(x) m_t(y) dx dy$$

we obtain through similar calculations that

$$\text{MMD}^2(m_t) = \int \int t^2 \left\langle v(x) \nabla_x \log \frac{p}{q}(x), \nabla_x \nabla_y k(x, y), v(y) \nabla_y \log \frac{p}{q}(y) \right\rangle dx dy + o(t)$$

from which the results follows. Note that the matrix-valued kernel $(K(x, y))_{ij} = (\nabla_x \nabla_y k(x, y))_{ij}$ is positive definite, a result akin to one of Zhou [279] but for the matrix-valued case. Indeed, for all $x, y \in \mathcal{X}, z, t \in \mathbb{R}^d$,

$$z K(x, y) t = \left\langle \sum_{i=1}^d z_i \partial_i k(x, \cdot), \sum_{i=1}^d t_i \partial_i k(y, \cdot) \right\rangle_{\mathcal{H}}$$

where $\partial_i k(x, \cdot) \in \mathcal{H}$ [279]. In the following, we write $\phi(x, y) = \sum_{i=1}^d y_i \partial_i k(x_i, \cdot)$. Now, for all sets of $\{x^i\}_{i=1}^n \in \mathcal{X}, \{y^j\}_{j=1}^n \in \mathbb{R}^d$, we have

$$\begin{aligned} \sum_{i,j=1}^n \langle K(x^i, x^j) y^j, y^i \rangle_{\mathbb{R}^d} &= \sum_{i,j=1}^n \langle \phi(x^i, y^i), \phi(x^j, y^j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^n \phi(x^i, y^i), \sum_{i=1}^n \phi(x^i, y^i) \right\rangle_{\mathcal{H}} \geq 0 \end{aligned}$$

from which it follows that K is indeed positive definite [167, Theorem 2.1].

VI.D.2 Universality of the Exponentiated-GFD and Exponentiated-KGFD kernel

To prove the universality of K_v and $K_{v,K}$ under the assumptions discussed in the related propositions, we rely on the following theorem [40, Theorem 2.2].

Theorem VI.D.1. *On a compact metric space $(\mathcal{Z}, d_{\mathcal{Z}})$ and for a continuous and*

injective map $\phi : \mathcal{Z} \mapsto H$, where H is a separable Hilbert space, the kernel $K(z, z') = e^{-\gamma \|\phi(z) - \phi(z')\|_H^2}$ is universal.

We first focus on the universality of K_v . We set as our goal to apply that theorem to our setting, in which $\mathcal{Z} := \mathcal{P}_X$ is a (sub)set of probability densities, which needs to be associated with a suitably chosen metric in order to make \mathcal{P}_X to be compact, and ϕ continuous. As bounded subsets of differentiable densities, whose elements can be framed as elements of the Sobolev space of first order $\mathcal{W}^{2,1}(v)$ [243]), are not compact a priori, we restrict ourselves to twice-differentiable densities with bounded Sobolev norm of second order, i.e., to $\mathcal{W}^{2,2}(v)$ with norm $\|p\|_{\mathcal{W}^{2,2}}^2 := \|p\|_{\mathcal{L}_2(v)}^2 + \sum_{i=1}^d \|\partial_i p\|_{\mathcal{L}_2(v)}^2 + \sum_{i,j=1}^d \|\partial_i \partial_j p\|_{\mathcal{L}_2(v)}^2$. From the Rellich-Kondrachov theorem [243], we know that when v has compact support, the canonical canonical injection $I : \mathcal{W}^{2,2}(v) \rightarrow \mathcal{W}^{2,1}(v)$ is a compact operator. As a consequence, for any bounded subset A of \mathcal{P}_X we thus have that $I(A)$ is compact for $\|f\|_{\mathcal{W}^{2,1}}^2 := \|f\|_{\mathcal{L}_2(v)}^2 + \sum_{i=1}^d \|\partial_i f\|_{\mathcal{L}_2(v)}^2$, which implies that any bounded subset A of \mathcal{P}_X is compact for $d(z, z') = \|z - z'\|_{\mathcal{W}^{2,1}}$. To apply the above theorem, it remains to prove the continuity and injectivity of $\phi : p \mapsto \nabla \log p$ under this metric (in that case the separable Hilbert space H is set to $\mathcal{L}_2(v)$). And indeed, for such a choice of d , ϕ and H , ϕ is continuous. To prove this fact, remark that differentiable densities with full support on X are bounded away from 0, making the use of a $\phi : p \mapsto \nabla \log p = \nabla p / p$ continuous. Moreover, ϕ is injective as $d_{\mathcal{W}^{2,1}}(p, q) := \|p - q\|_{\mathcal{W}^{2,1}} \neq 0$ implies $\|\nabla \log p - \nabla \log q\|_{\mathcal{L}_2(v)} \neq 0$. Thus, all conditions of [40, Theorem 2.2] are satisfied, and the result follows as a consequence.

We now move on to prove the universality of $K_{v,K}$. The proof follows the same reasoning as the proof of the universality of K_v , the only difference being the fact that the feature map $\tilde{\phi}$ of $K_{v,K}$ is given by $T_v \circ \phi$, where $\phi : p \mapsto \nabla \log p$ and $T_{K,v} : \mathcal{L}(X, \mathbb{R}^d) \rightarrow \mathcal{H}_K$ is given by

$$T_{K,v} : f \mapsto \int_X K_x f(x) v(dx).$$

However, if v is a probability measure and K is bounded, then $T_{K,v}$ is a bounded

operator, and thus continuous, making $\tilde{\phi}$ continuous. Moreover, if K is characteristic, $T_{K,v}$ is injective. Thus $\tilde{\phi}$ is injective and continuous, from which the result follows by [40].

VI.E Background on Stein and Fisher divergences

The Fisher Divergence Consider two continuously differentiable densities p and q on \mathbb{R}^d . Then the Fisher divergence [231, 129] between p and q is defined as:

$$\text{FD}(p||q) = \int_{\mathbb{R}^d} \|\nabla \log p(x) - \nabla \log q(x)\|_2^2 p(x) dx.$$

We refer to [231] for an overview of the properties of the Fisher divergence, including its relative strength w.r.t. other divergences, and other formulations. The Fisher divergence was used for learning statistical models of some training data in [117, 231], and more recently in [227].

Stein Discrepancies Of proximity to the Fisher divergence is the family of Stein discrepancies [10]. Stein discrepancies build upon the concept of Stein operators, which are operators $\mathcal{A}_{\mathbb{P}}$ such that

$$\mathbb{E}_{\mathbb{Q}} [\mathcal{A}_{\mathbb{P}} f] = 0 \iff \mathbb{Q} = \mathbb{P}$$

for any f within a set $\mathcal{G}(\mathcal{A}_{\mathbb{P}}) \subset \text{dom}(\mathcal{A}_{\mathbb{P}})$ called the *Stein class* of $\mathcal{A}_{\mathbb{P}}$. Following this definition, the $\mathcal{A}_{\mathbb{P}}$ -stein discrepancy is defined as

$$\text{SD}_{\mathcal{A}_{\mathbb{P}}}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{G}(\mathcal{A})} \|\mathbb{E}_{\mathbb{Q}} \mathcal{A} f\|$$

which satisfies by construction the axioms of a *dissimilarity* (or *divergence*) measure between \mathbb{P} and \mathbb{Q} .

Link Between the Fisher divergence and Diffusion Stein Discrepancies Perhaps the most famous Stein discrepancy is the one that sets $\mathcal{A}_{\mathbb{P}}$ to be the infinitesimal

generator of the isotropic diffusion process toward \mathbb{P} [85]:

$$\begin{cases} dX_t &= \nabla \log p(X_t) dt + \sqrt{2} dW_t \\ (\mathcal{A}_{d,\mathbb{P}} f)(\cdot) &= \langle \nabla \log p(\cdot), \nabla f \rangle + \langle \nabla, \nabla f \rangle \end{cases}$$

Recalling that $\mathbb{E}_{\mathbb{P}} [\mathcal{A}_{d,\mathbb{P}} f] = 0$ for all $f \in \mathcal{G}(\mathcal{A}_{d,\mathbb{P}})$, we obtain the following formulation for the diffusion Stein discrepancy

$$\begin{aligned} \text{SD}_{\mathcal{A}_{d,\mathbb{P}}}(\mathbb{P}, \mathbb{Q}) &:= \sup_f \|\mathbb{E}_{\mathbb{Q}} \mathcal{A}_{d,\mathbb{P}} f\| = \sup_f \left\| \mathbb{E}_{\mathbb{Q}} (\nabla \log p - \nabla \log q)^{\top} \nabla f \right\| \\ &= \sup_{g=\nabla f} \left\| \mathbb{E}_{\mathbb{Q}} (\nabla \log p - \nabla \log q)^{\top} g \right\|, \end{aligned}$$

highlighting the connection between the Fisher divergence and the diffusion Stein discrepancy.

Link Between the Fisher divergence and the Kernelized Stein Discrepancy Given a RKHS \mathcal{H} such that $B_{\mathcal{H}^{\otimes d}}(0_{\mathcal{H}^{\otimes d}}, 1)$ is a Stein class for $\mathcal{A}_{d,\mathbb{P}}$, the kernelized Stein discrepancy [84] is given by

$$\begin{aligned} \text{KSD}(\mathbb{P}, \mathbb{Q}) &:= \sup_{h=\nabla f \in \mathcal{H}^{\otimes d}: \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}} \langle \nabla \log p(x) - \nabla \log q(x), h(x) \rangle\| \\ &= \sup_{h=\nabla f \in \mathcal{H}^{\otimes d}: \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \langle h, \mathbb{E}_{\mathbb{Q}} (\nabla \log p(x) - \nabla \log q(x)) k(x, \cdot) \rangle_{\mathcal{H}^{\otimes d}}^{1/2} \\ &= \|\mathbb{E}_{\mathbb{Q}} [(\nabla \log p(x) - \nabla \log q(x)) k(x, \cdot)]\|_{\mathcal{H}^{\otimes d}} \\ &= \|I_{k,\mathbb{Q}}^*(\nabla \log p - \nabla \log q)\|_{\mathcal{H}^{\otimes d}} \end{aligned}$$

where $I_{k,\mathbb{Q}}^*$ is the adjoint of the canonical injection from $\mathcal{H}^{\otimes d}$ to $(L^2(\mathbb{Q}))^{\otimes d}$, also known as the *kernel integral operator*. This derivation shows that the KSD can be seen as a kernelized version of the Fisher divergence.

Link between MMD and KSD It is possible [84] to reframe the KSD as an MMD with a specific kernel. Indeed, given some base kernel $k(x, y)$, define the following “Stein” kernel

$$\tilde{k}(x, y) = \langle \nabla \log p(x) k(x, \cdot) + \nabla k(x, \cdot), \nabla \log p(y) k(y, \cdot) + \nabla k(y, \cdot) \rangle_{\mathcal{H}^{\otimes d}}$$

which is positive definite as an inner product of a feature map of x . Then $\mathcal{H}_{\tilde{k}} = \mathcal{A}_{d,\mathbb{P}}(\mathcal{H})$ and $\|f\|_{\mathcal{H}_{\tilde{k}}} = \|\mathcal{A}f\|_{\mathcal{H}_{\tilde{k}}^{\otimes d}}$. Moreover, we have that $\mathbb{E}_{\mathbb{P}}\tilde{h} = 0$ for all $\tilde{h} \in \mathcal{H}_{\tilde{k}}$. By the definition of the KSD, we have that

$$\begin{aligned}\text{KSD}(\mathbb{P}, \mathbb{Q}) &= \sup_{h \in \mathcal{H}^{\otimes d}: \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}} \mathcal{A}_{d,\mathbb{P}} h\| \\ &= \sup_{h \in \mathcal{H}^{\otimes d}: \|h\|_{\mathcal{H}^{\otimes d}} \leq 1} \|\mathbb{E}_{\mathbb{Q}} \mathcal{A}_{d,\mathbb{P}} h - \mathbb{E}_{\mathbb{P}} \mathcal{A}_{d,\mathbb{P}} h\|_{\mathcal{H}} \\ &= \sup_{h \in \mathcal{H}_{\tilde{k}}: \|h\|_{\mathcal{H}_{\tilde{k}}} \leq 1} \|\mathbb{E}_{\mathbb{Q}} h - \mathbb{E}_{\mathbb{P}} h\|_{\mathcal{H}_{\tilde{k}}} \\ &= \text{MMD}_{\tilde{k}}(\mathbb{P}, \mathbb{Q}).\end{aligned}$$

Differential Inequalities between the KL and the Fisher Divergence It is well known [34] that the KL divergence can be related to the Fisher divergence by considering the evolution of $\text{KL}(\mathbb{P}_t || \mathbb{Q})$ when \mathbb{P}_t evolves according to the Fokker-Planck equation

$$\partial_t p_t(x) = \text{div}(p_t(x)(\nabla \log q_t(x) - \nabla \log p_t(x))), \quad \mathbb{P}_0 = \mathbb{P}. \quad (\text{VI.13})$$

(Two relevant side notes: for any $t \geq 0$, \mathbb{P}_t is the law at time t of the Markov process $(X_t)_{t \geq 0}$ such that $X_0 \sim \mathbb{P}$ and undergoing an isotropic diffusion towards \mathbb{Q} . Moreover, Equation (VI.13) is also the Wasserstein gradient flow equation of $\text{KL}(\cdot || \mathbb{Q})$ starting from \mathbb{P}). Recalling that Equation (VI.13) is satisfied in the sense of distributions, and relying on Gateaux-Derivative formulas for Free Energy-type functionals [see 6, for more precise statements], we have:

$$\begin{aligned}\frac{d\text{KL}(\mathbb{P}_t || \mathbb{Q})}{dt} &= \frac{\partial \text{KL}}{\partial \mathbb{P}} \Big|_{\mathbb{P}_t} \frac{d\mathbb{P}_t}{dt} \\ &= \int \langle \nabla(\log p_t(x) - \log q_t(x)), (\nabla \log q_t - \nabla \log p_t) \rangle d\mathbb{P}_t(x) \\ &= -\text{FD}(\mathbb{P}_t, \mathbb{Q}).\end{aligned}$$

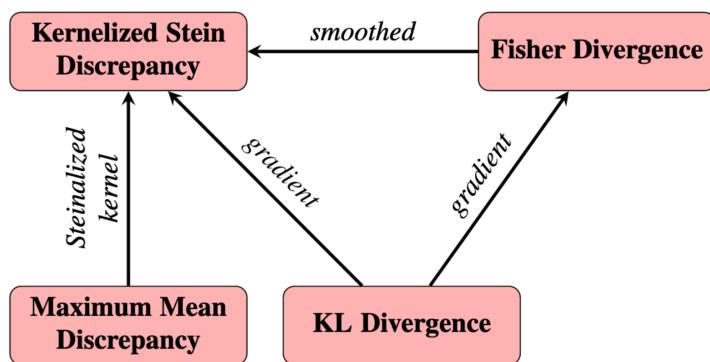


Figure VI.E.1: Relationships between the Fisher divergence, the KL divergence, the MMD, and the KSD [156].

VI.F Experimental Results

This section contains visualizations of all experiments discussed in VI.6, including figures contained in the main text. In all experiments we set the significance level to $\alpha = 0.05$. Every experiment is repeated for 100 randomly sampled datasets and with 500 bootstrap iterations for estimating the quantile of the test statistic.

We use Gaussian distributions and compare the KCCSD and the SKCE with different combinations of kernels. For the KCCSD, for Gaussian distributions all considered test statistics can be evaluated exactly. Alternatively, for the exponentiated (kernelized) Fisher kernel and the exponentiated MMD kernel one can resort to approximations using samples from the base measure. For the SKCE, however, the test statistic can be evaluated exactly on in special cases such as Gaussian kernels on the target space. All approximate evaluations are performed with 10 samples.

VI.F.1 Mean Gaussian Model

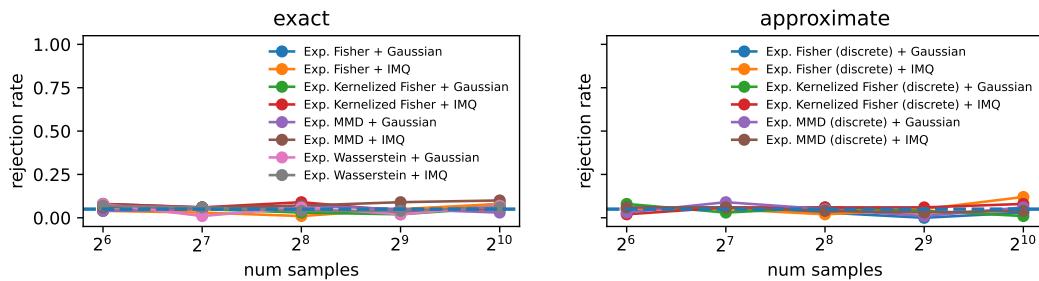


Figure VI.F.1: False rejection rate of the KCCSD for MGM ($\delta = 0$).

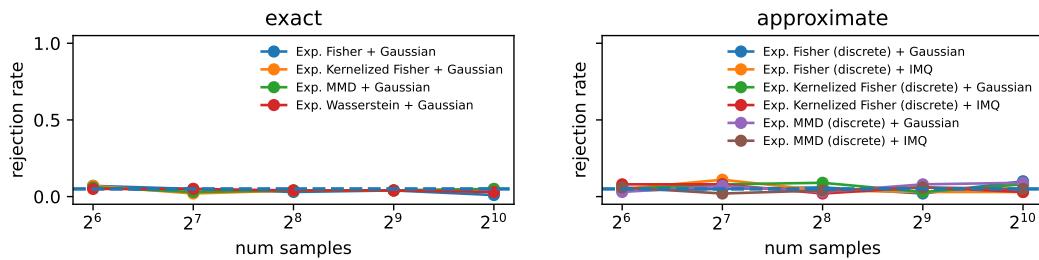


Figure VI.F.2: False rejection rate of the SKCE for MGM ($\delta = 0$).

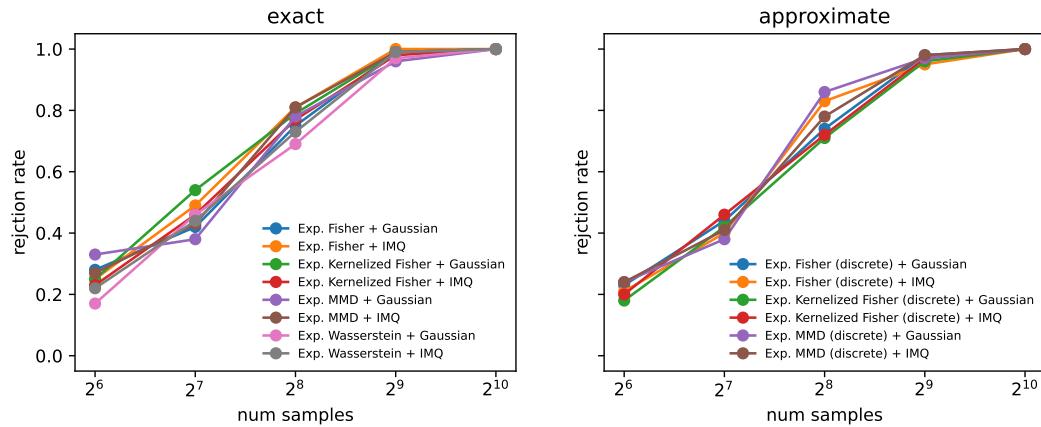


Figure VI.F.3: Rejection rate of the KCCSD for MGM ($\delta = 0.1, c = \mathbf{1}_d$).

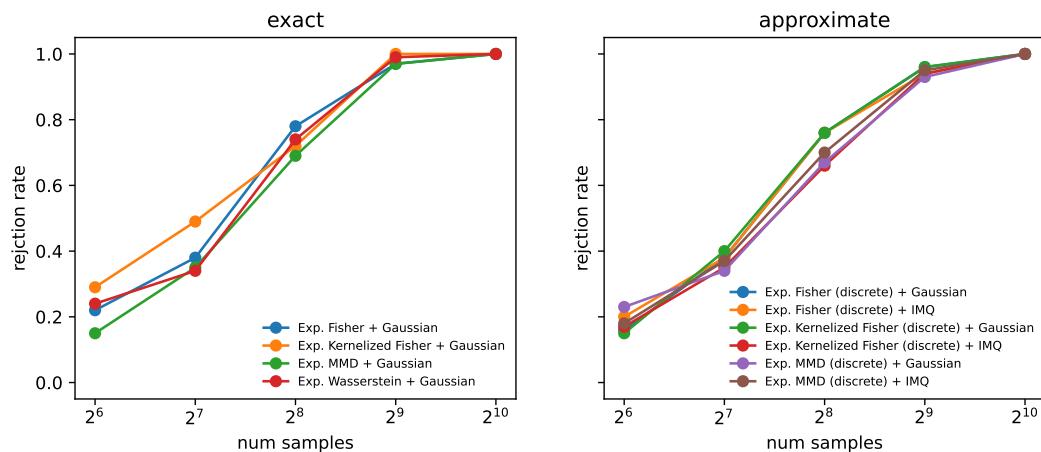


Figure VI.F.4: Rejection rate of the SKCE for MGM ($\delta = 0.1, c = \mathbf{1}_d$).

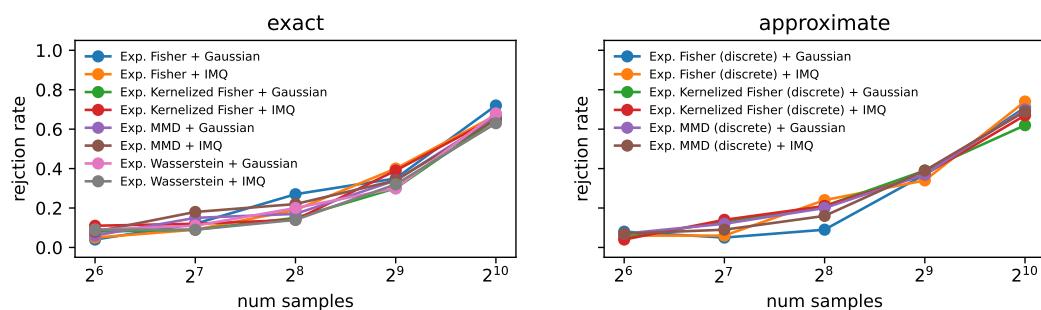


Figure VI.F.5: Rejection rate of the KCCSD for MGM ($\delta = 0.1, c = e_1$).

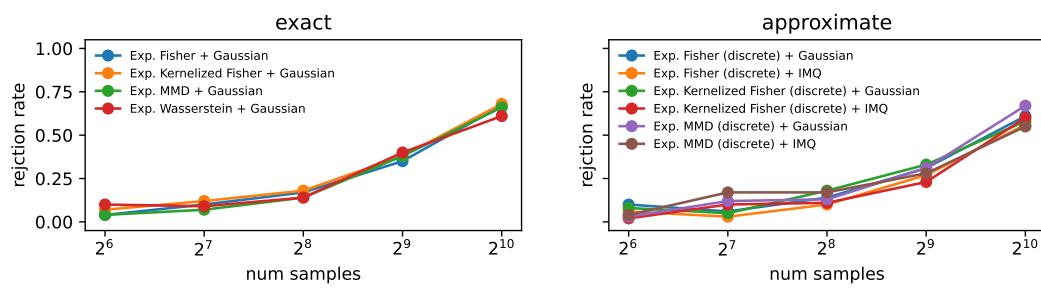


Figure VI.F.6: Rejection rate of the SKCE for MGM ($\delta = 0.1, c = e_1$).

VI.F.2 Linear Gaussian Model

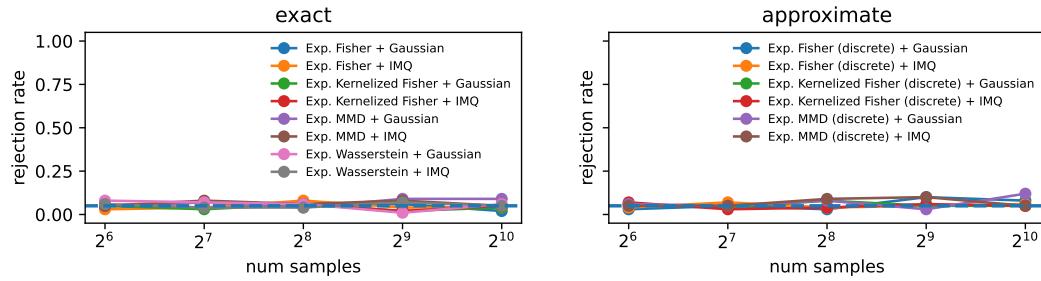


Figure VI.F.7: False rejection rate of the KCCSD for LGM ($\delta = 0$).

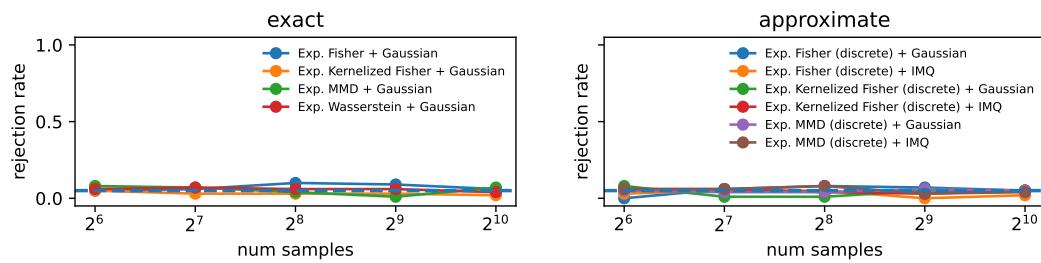


Figure VI.F.8: False rejection rate of the SKCE for LGM ($\delta = 0$).

VI.F.3 Heteroscedastic Gaussian Model

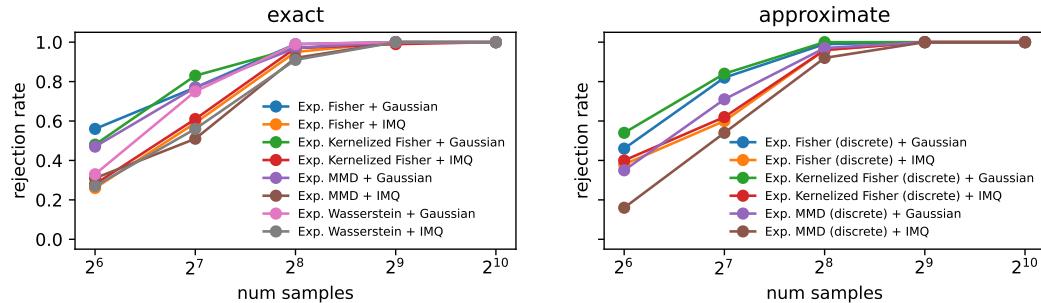
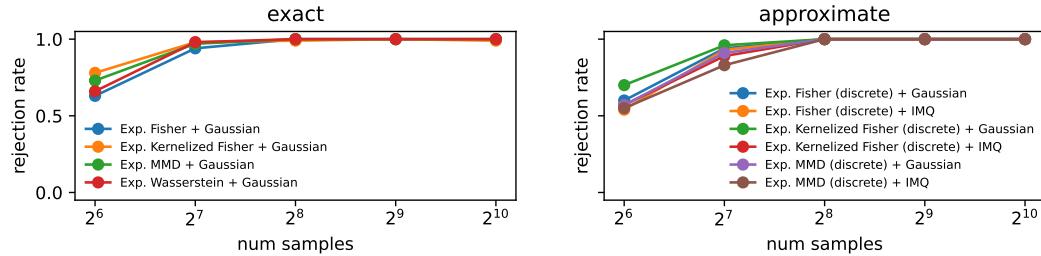
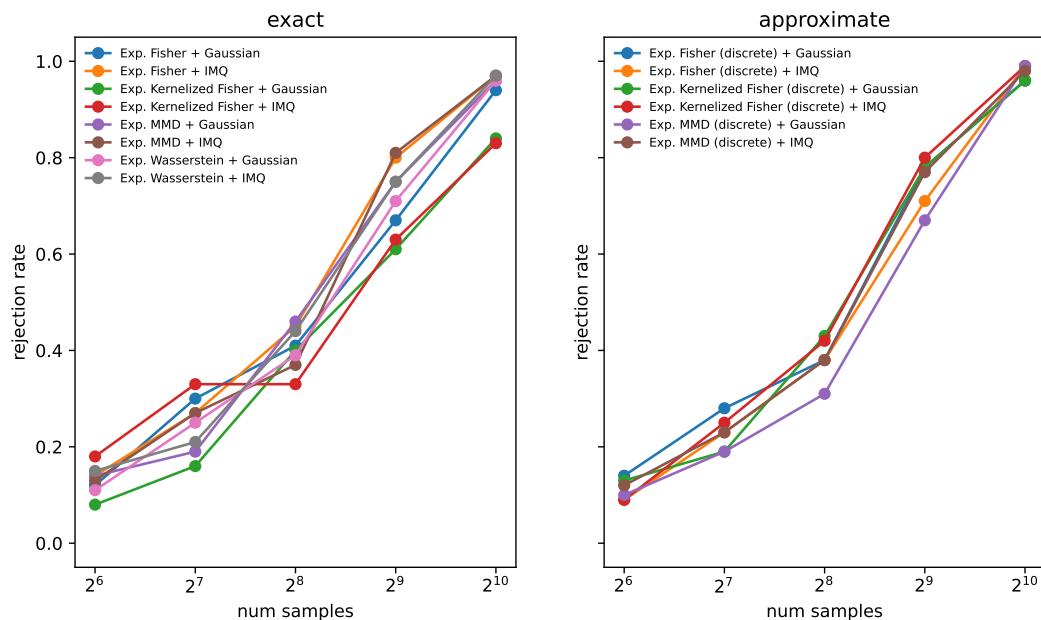
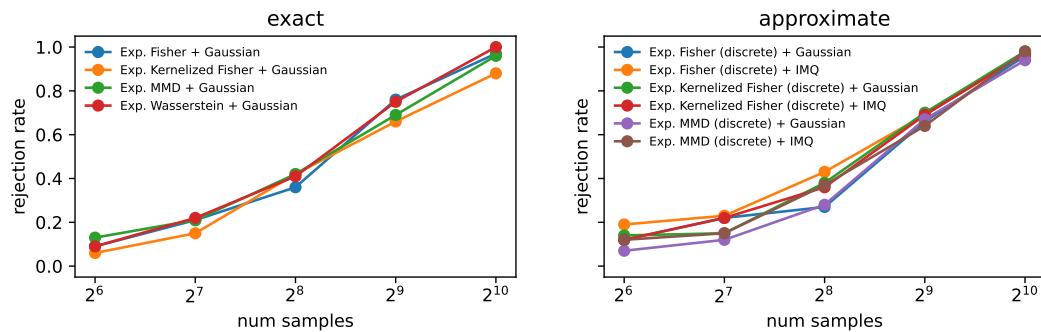


Figure VI.F.9: Rejection rate of the KCCSD for HGM ($\delta = 1$).

**Figure VI.F.10:** Rejection rate of the SKCE for HGM ($\delta = 1$).

VI.F.4 Quadratic Gaussian Model

**Figure VI.F.11:** Rejection rate of the KCCSD for QGM ($\delta = 1$).**Figure VI.F.12:** Rejection rate of the SKCE for QGM ($\delta = 1$).

CHAPTER VII

Kernel-based Evaluation of Conditional Biological Sequence Models

This Chapter is based on the following work:

Pierre Glaser, Steffanie Paul, Alissa M Hummer, Charlotte Deane, Debora Susan Marks, and Alan Nawzad Amin. Kernel-based evaluation of conditional biological sequence models. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=2dlmcTXfcY>

Abstract

We propose a set of kernel-based tools to evaluate the designs and tune the hyperparameters of conditional sequence models, with a focus on problems in computational biology. The backbone of our tools is a new measure of discrepancy between the true conditional distribution and the model’s estimate, called the Augmented Conditional Maximum Mean Discrepancy (ACMMD). Provided that the model can be sampled from, the ACMMD can be estimated unbiasedly from data to quantify absolute model fit, integrated within hypothesis tests, and used to evaluate model reliability. We demonstrate the utility of our approach by analyzing a popular protein design model, ProteinMPNN. We are able to reject the hypothesis that ProteinMPNN fits its data for various protein families, and tune the model’s temperature hyperparameter to achieve a better fit.

VII.1 Introduction

Conditional sequence models constitute one of the most prominent model classes of modern machine learning. Such models have allowed progress in longstanding problems in fields ranging from natural language generation to biomedical applications such as genomics and protein design. Abstracting away the precise nature of the data, the objective common to many of these problems can be summarized as the prediction of high-dimensional discrete-valued sequences, given some possibly

high-dimensional input information about the sequence. For example, in protein design, inverse folding models [51] seek to learn the conditional distribution of amino acid sequences (proteins) that are likely to fold to a given input protein backbone 3D geometry, or *structure*.

In such problems, it is crucial to evaluate the properties of the trained model. Model evaluation can help assess the risk of using the model’s predictions in the real world, such as performing in-vitro experiments (a time-intensive process), guide hyperparameter searches, and deepen one’s understanding of the model’s behavior. Two properties are particularly important to measure: the first one is model *accuracy*, which describes how well the model approximates the true conditional distribution of the target variable given the input. Models with high accuracy have learned the underlying structure of the data, suggesting a high potential value in deploying them in real-world applications. However, in practice, it is likely that models will not be perfectly accurate. Inaccurate models can still be useful as long as they fall back to conservative guesses (in the extreme case, the prior distribution) when they are uncertain. From a statistical perspective, this property is known as *reliability* [31, 254, 269], and will be the second property of interest in this work.

Given a set of real samples, the standard approach to evaluate models in protein design consists in using log-likelihoods or sequence recovery [51, 111, 72]. However, log-likelihoods cannot be used to evaluate reliability, and are only *relative* measures of accuracy: these methods can only be used to compare models and would not alert the practitioner for example if all models make very poor predictions. Instead, to assess how far a model is from being optimally accurate and consistent — and thus the potential value in improving it, by for example collecting more data or increasing its complexity, one should consider *absolute* rather than *relative* metrics, that is, metrics that not only allow one to compare models to each other, but also to evaluate a single model’s performance without any other point of comparison. For these metrics to have practical value, they should come with estimators computable from data samples. These estimators should be efficiently computable, recover the true metric as in the large sample size limit (i.e. be *consistent*), and preferably be

centered around the true value of the metric (i.e. be *unbiased*). Finally, to factor out the statistical error coming from estimating these metrics using a finite number of samples, these metrics should be integrable into hypothesis tests built to detect *statistically significant* mismatches between the model and the data.

Contributions In this work, we introduce a set of absolute evaluation metrics for measuring the accuracy and the reliability of conditional sequence models. Both our metrics are grounded in a new measure of divergence between conditional probability distributions, which we call the Augmented Conditional Maximum Mean Discrepancy (ACMMD), which extends the kernel-based conditional goodness-of-fit framework of Jitkrittum et al. [128], Pierre Glaser et al. [246], Widmann et al. [269] to the case of sequence-valued variables. We analyze the statistical properties of our proposed metrics, which can be estimated using samples from the data and the model. Under certain conditions, we show that the ACMMD is able to detect any mismatch between the model and the data. In addition, we integrate the ACMMD into hypothesis tests to detect such mismatches from the model and the data samples. We showcase the utility of our methods by using them in an in-depth analysis of a popular inverse folding model - ProteinMPNN [51]. Our results demonstrate the theoretical properties of our methods, while also providing insight as to how to gauge the certainty and applicability of ProteinMPNN for designing proteins of varying topologies and evolutionary families.

VII.2 Problem Setting

We consider the problem of predicting a discrete sequence-valued variable we are designing $Y \in \mathcal{Y}$, for example a biological sequence, conditionally on a variable $X \in \mathcal{X}$ at our disposal. The predicted sequence Y is allowed to have an arbitrary length, e.g. $\mathcal{Y} = \cup_{\ell=1}^{\infty} \mathcal{A}^{\ell}$, where \mathcal{A} is a finite set. In protein design, X could be the 3D structure of a protein (e.g. $\mathcal{X} = \cup_{\ell=1}^{\infty} \mathbb{R}^{3\ell}$) and Y the sequence of amino acids making up the protein, in which case \mathcal{A} is the set of amino acids. Given a large number of i.i.d measurements $\{X_i, Y_i\}_{i=1}^{N_T}$ from a distribution $\mathbb{P}(X, Y)$, for example pairs of sequences and structures from the Protein Data Bank [120], we train a *predictive*

model $Q| : x \mapsto Q_{|x}$ that takes in a value x and outputs a distribution on Y , $Q_{|x}(Y)$ that attempts to match the true conditional $\mathbb{P}(Y|X = x)$, denoted $\mathbb{P}_{|x}(Y)$ in this work. After training, we are interested in quantifying how accurately $Q|$ approximates $\mathbb{P}|$ on average across all values of x after training, using a held-out set of samples $\{X_i, Y_i\}_{i=1}^N \sim \mathbb{P}(X, Y)$. Quantifying the accuracy of $Q|$ is known as the *conditional goodness-of-fit* problem, and we address it in Section VII.3. Furthermore, we will also be interested in quantifying the reliability of $Q|$, a task which we address in Section VII.4.

VII.3 Conditional Goodness-of-Fit with ACMMD

In this section, we propose a metric that quantifies the accuracy of a predictive sequence model. We will show that this metric satisfies many desirable properties: first, it is absolute and able to detect any differences between conditional distributions. Second, it can be unbiasedly and efficiently estimated using samples from the model and the data distribution. Third, it can be used in hypothesis tests to detect statistically significant mismatches from such samples.

VII.3.1 The Augmented Conditional MMD

We now propose a method to quantitatively evaluate the conditional goodness-of-fit of $Q|$ to $\mathbb{P}|$. Our approach consists in constructing a *divergence* $D(\mathbb{P}|, Q|)$, between the conditional distribution of Y given X and the model $Q|$. By definition, this divergence should satisfy:

$$\begin{aligned} (i) \quad & D(\mathbb{P}|, Q|) \geq 0 \\ (ii) \quad & D(\mathbb{P}|, Q|) = 0 \iff \mathbb{P}_{|x} = Q_{|x}, \mathbb{P}(X)\text{-a.e.} \end{aligned} \tag{VII.1}$$

Combined, these two properties ensure that $D(\mathbb{P}|, Q|)$ is *absolute*, e.g. assigns the known value lowest value 0 to the best possible model, and is able to distinguish any mismatch between the model and the data, which is crucial to prevent blind spots in our evaluation. We borrow the idea of comparing $Q|$ with $\mathbb{P}|$ by comparing the joint $\mathbb{P}(X, Y)$ with a joint that keeps the same marginal $\mathbb{P}(X)$ but swaps $\mathbb{P}|$ with $Q|$. These two joint distributions are equal if and only if $Q|$ and $\mathbb{P}|$ match almost everywhere.

To compare these two distributions, we will use the Maximum Mean Discrepancy (MMD) [90] given by:

$$\text{MMD}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\substack{f \in \mathcal{H}_{\mathcal{Z}} \\ \|f\|_{\mathcal{H}_{\mathcal{Z}}} \leq 1}} \mathbb{E}_{\mathbb{Q}_1}[f(Z)] - \mathbb{E}_{\mathbb{Q}_2}[f(Z)]. \quad (\text{VII.2})$$

Here, \mathcal{Z} is some measurable space, \mathbb{Q}_1 and \mathbb{Q}_2 are probability measures on \mathcal{Z} , and $\mathcal{H}_{\mathcal{Z}}$ is a reproducing kernel Hilbert space (RKHS) of functions from \mathcal{Z} to \mathbb{R} with kernel $k_{\mathcal{Z}}$ [20]. Applying this general definition to the case at hand, we obtain a measure of accuracy for $Q|$, defined below.

Definition VII.3.1 (Augmented Conditional MMD). Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with law $\mathbb{P}_X \otimes \mathbb{P}|$. Let $Q|$ be a conditional probability from \mathcal{X} to \mathcal{Y} . We define the Augmented Conditional MMD (ACMMD) between $\mathbb{P}|$ and $Q|$ as:

$$\text{ACMMD}(\mathbb{P}|, Q|) := \text{MMD}(\mathbb{P}_X \otimes \mathbb{P}|, \mathbb{P}_X \otimes Q|) \quad (\text{VII.3})$$

where the MMD is evaluated with a user-specified kernel $k_{\mathcal{X} \times \mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$. Here, $\mathbb{P}_X \otimes \mathbb{P}|$ is defined by $(X, Y) \sim \mathbb{P}_X \otimes \mathbb{P}| \iff X \sim \mathbb{P}_X, (Y|X=x) \sim \mathbb{P}|_x$, and similarly for $\mathbb{P}_X \otimes Q|$.

Choice of kernel for ACMMD The ACMMD requires specifying a kernel on the joint space $\mathcal{X} \times \mathcal{Y}$. In this work, we will focus on the case where $k_{\mathcal{X} \times \mathcal{Y}}$ is the *tensor product* kernel $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ of two kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ on \mathcal{X} and \mathcal{Y} respectively:

$$k_{\mathcal{X} \times \mathcal{Y}}((x, y), (x', y')) = k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, y') \quad (\text{VII.4})$$

This choice is popular in practice, and the resulting ACMMD retains its desirable properties, as we show next.

The ACMMD is a divergence between conditional probabilities The ACMMD writes as divergence (which is symmetric, e.g. a distance) between joint distributions, while we seek to use it to compare conditional distributions. The following lemma shows that the same ACMMD can be formulated in alternative manner that highlights

its purpose as a conditional distribution comparator.

Lemma VII.3.2. *Under mild integrability conditions, we have:*

$$\text{ACMMD}(\mathbb{P}|, Q|) = \left\| T_{K_X}(\mu_{\mathbb{P}} - \mu_{Q|}) \right\|_{\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}}$$

Where $\mu_{\mathbb{P}|}$ and $\mu_{Q|}$ are the conditional mean embeddings [192] of $\mathbb{P}|$ and $Q|$, $K_X(x, x') := k_X(x, x')I_{\mathcal{H}_{\mathcal{Y}}}$ (here, $I_{\mathcal{H}_{\mathcal{Y}}}$ the identity operator) is an operator-valued kernel with associated vector-valued RKHS $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$, and T_{K_X} is its associated integral operator from $L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ to $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$. Moreover, if k_X and k_Y are C_0 -universal¹, then it holds that:

$$\text{ACMMD}(\mathbb{P}|, Q|) = 0 \iff \mathbb{P}_{|x} = Q_{|x}, \quad \mathbb{P}_X\text{-a.e.}$$

The complete statement (with the full set of assumptions, and the definition of integral operators) and its proof can be found in Section VII.A. Lemma VII.3.2 shows that the ACMMD can be understood as the result of a two-step procedure, given by (1) computing the conditional mean embedding $\mu_{\mathbb{P}|} : x \mapsto \mathbb{E}_{\mathbb{P}_{|x}}[k_Y(y, \cdot) | X = x]$ of $\mathbb{P}|$ (resp. of $Q|$), which is a function from \mathcal{X} to $\mathcal{H}_{\mathcal{Y}}$, and (2) embed the difference of these conditional mean embeddings into the *vector-valued* RKHS $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ with kernel K_X , before returning its associated RKHS norm. The second part of the lemma gives sufficient conditions for the ACMMD to discriminate between any non (\mathbb{P}_X -a.e) equal conditional distributions, fulfilling the requirements specified in Equation (VII.1): these conditions are to use *universal* kernels k_X and k_Y . Regarding k_Y , this requirement is not very restrictive, as many universal kernels on sequences have been shown to be universal [9]. The difficulty in finding a universal k_X will depend on the space \mathcal{X} (unspecified in this work) for the problem at hand.

Estimating the ACMMD from data Crucial to this work is the fact that if the model $Q|$ can be sampled from for any $x \in \mathcal{X}$, ACMMD² will admit tractable unbiased estimators. To see this, we first rewrite ACMMD² in a form that will make this

¹A kernel k is C_0 -universal if the associated RKHS \mathcal{H}_k is dense in $C_0(\mathcal{X})$, the space of continuous functions on \mathcal{X} vanishing at infinity [230]

property apparent.

Lemma VII.3.3. *Let $Z := (X, Y, \tilde{Y})$ the triplet of random variables with law ² $\mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$. Then, under the integrability assumptions of Lemma VII.3.2, we have that:*

$$\text{ACMMD}^2(\mathbb{P}_{|}, Q_{|}) = \mathbb{E}_{Z_1, Z_2} [h(Z_1, Z_2)]$$

where Z_1, Z_2 are two independent copies of Z and h is a symmetric function given by:

$$h(Z_1, Z_2) := k_{\mathcal{X}}(X_1, X_2)g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2))$$

$$g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) := k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) + k_{\mathcal{Y}}(Y_1, Y_2)$$

$$-k_{\mathcal{Y}}(\tilde{Y}_1, Y_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2)$$

Lemma VII.3.3, proved in Section VII.B.2, expresses ACMMD^2 as a double expectation given two independent samples of $(X, Y, \tilde{Y}) \sim \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$. Leveraging this fact, we can derive an unbiased and consistent estimator for ACMMD^2 .

Lemma VII.3.4. *Let $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \stackrel{i.i.d.}{\sim} \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$ be samples from the data and the model. Then an unbiased estimator $\widehat{\text{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ of $\text{ACMMD}^2(\mathbb{P}_{|}, Q_{|})$ is given by:*

$$\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j)) \quad (\text{VII.5})$$

Lemma VII.3.4, proved in Section VII.B.2, shows that it is possible to unbiasedly estimate ACMMD^2 even when the analytical model expectations are intractable, provided that one can sample from the model. This estimator takes the form of a U-statistics [221, Chapter 5] with symmetric probability kernel h , which are well-studied in the statistics literature. In particular, they provide a generic framework to obtain minimal-variance analogues of unbiased estimators [221, Chapter 5, p. 176]. $\widehat{\text{ACMMD}}^2$ is a *consistent* estimator of ACMMD^2 : under the integrability assumptions of Lemma VII.3.2 the strong law of large numbers applies [221, Section 5.4, Theorem A], and we have: $\widehat{\text{ACMMD}}^2(\mathbb{P}_{|}, Q_{|}) \xrightarrow[N \rightarrow \infty]{a.s.} \text{ACMMD}^2(\mathbb{P}_{|}, Q_{|})$. We

²Identifying $Q_{|}$ with its analogue Markov kernel $\tilde{Q}_{|}$ from $(\mathcal{X} \times \mathcal{Y}, \mathcal{X} \otimes \mathcal{Y})$ such that $\tilde{Q}_{|(x,y)}(dy') := Q_{|x}(dy')$.

provide a more detailed characterization of the asymptotic distribution of $\widehat{\text{ACMMD}}^2$ in Section VII.B.

VII.3.2 Testing Conditional Goodness-of-Fit with ACMMD

In the limit of infinitely many samples, a positive ACMMD means that the model and the data differ. However, in practice, when only a finite number of samples are available, our estimate $\widehat{\text{ACMMD}}^2$ is only a noisy version of the true ACMMD^2 , meaning we cannot conclude whether the model fits the data by directly inspecting its value. Instead, we need a procedure that accounts for the estimation noise; we achieve this by using the ACMMD as part of a hypothesis test deciding between two different hypotheses:

$$\begin{cases} H_0 : \text{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0 \\ H_1 : \text{ACMMD}(\mathbb{P}_{|}, Q_{|}) > 0 \end{cases}$$

In particular, we construct a test that takes as input a sample $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N$ from the data and the model and outputs a (binary) decision to reject (or not) the null hypothesis H_0 based on whether $\widehat{\text{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ exceeds a certain threshold. Because of the estimation noise arising from the use of finitely many samples, such a test cannot systematically output the right decision. Nonetheless, we build our test to ensure a *false rejection* (e.g. reject H_0 while $\mathbb{P}_{|} = Q_{|}$ a.e) rate of $\alpha \in (0, 1)$, a common practice in statistical testing [90]. To do so, we would like to set the rejection threshold $q_{1-\alpha}$ to be an estimate of the $1 - \alpha$ quantile of the distribution of $\widehat{\text{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ under H_0 . However, since $q_{1-\alpha}$ is not available in closed form, we instead compute an estimate $\widehat{q}_{1-\alpha}$ using the wild bootstrap procedure [13]. This procedure draws B samples $\{\widehat{\text{ACMMD}}_b^2\}_{b=1}^B$ of the form:

$$\widehat{\text{ACMMD}}_b^2 := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N}^N W_i^b W_j^b h(Z_i, Z_j) \quad (\text{VII.6})$$

where $\{W_i^b\}_{i=1}^{b=1\dots B}$ are i.i.d. Rademacher random variables independent of the data, from which we compute a quantile estimate $\widehat{q}_{1-\alpha}$ of this distribution of samples (see Section VII.C for a precise definition of $\widehat{q}_{1-\alpha}$). Importantly, this procedure guarantees an exact control of the false rejection rate at level α . We prove this

Algorithm 11 ACMMMD Conditional Goodness-of-fit Test

Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_| \otimes Q|$

Parameters: Level α , kernel $k_{\mathcal{X}}$, kernel $k_{\mathcal{Y}}$

// Estimate ACMMMD using Equation (VII.5)

$$\widehat{\text{ACMMMD}}^2 \leftarrow \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i < j}}^N h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j))$$

Sample $\{\widetilde{\text{ACMMMD}}_b^2\}_{b=1}^B$ using Equation (VII.6)

$$\hat{q}_{1-\alpha} \leftarrow \text{approx. } (1 - \alpha)\text{-quantile of } \{\widetilde{\text{ACMMMD}}_b^2\}_{b=1}^B$$

if $\widehat{\text{ACMMMD}}^2 \leq \hat{q}_{1-\alpha}$ **then**

Fail to reject H_0

else

Reject H_0

end if

fact in Section VII.C.3, where we cast the wild bootstrap procedure as a Monte-Carlo estimation of the distribution of $\widehat{\text{ACMMMD}}^2$ when $\mathbb{P}_| = Q|$, which is valid non-asymptotically. Our test, which we call the ACMMMD test, is summarized in Algorithm 11. To the best of our knowledge, this is the first conditional goodness-of-fit test that is applicable to sequence models.

VII.4 Assessing Reliability with ACMMMD

In practice, our model $Q|$ may not fit the data perfectly, and it is important to distinguish (at a given level of inaccuracy) models that remain consistent with their training data from ones that fail more drastically. In this section, we show how the ACMMMD can be used to evaluate model reliability, a statistical property capturing model and data consistency.

Problem Setting A model $Q|$ is said to be reliable [31, 254, 269] if the distribution of the target Y given that the model made a specific prediction q is this prediction q itself, e.g. if:

$$q = \mathbb{P}(Y \in \cdot \mid Q|_X = q) \quad \mathbb{P}(Q|_X)\text{-a.e.} \quad (\text{VII.7})$$

Here, $Q|_X \in \mathcal{P}(\mathcal{Y})$ (the space of probability distributions on \mathcal{Y}) is the random variable obtained by evaluating the model $Q|$ at a random value of the input variable X . Reliability differs from accuracy in that it does not require the model to learn all

the information between X and Y , but only to make truthful predictions on average — thus, by assessing reliability, one may be able to detect models that hallucinate non-realistic sequences (such as repeats of the same token) in regions of the input space where they are inaccurate, instead of making a conservative guess, such as falling back to the prior distribution. In particular, reliability can be used as an additional criterion to discriminate between models that are equally accurate. From a theoretical perspective, reliability and accuracy can be handled in a unified manner: indeed, Equation VII.7 shows that reliability is defined as an equality between the conditional distribution of Y given a model prediction q , $\mathbb{P}_{|q}^Q := \mathbb{P}(Y = \cdot | Q_{|X} = q)$ and a “model” of this conditional distribution mapping $q \in \mathcal{P}(\mathcal{Y})$ to itself, e.g. $Q_{|}^{\text{Rel}} : q \longmapsto Q_{|q}^{\text{Rel}} = q$. We thus propose to measure reliability using the ACMMD (a distance between conditional distributions) between $Q_{|}^{\text{Rel}}$ and $\mathbb{P}_{|}^Q$.

Definition VII.4.1 (ACMMD for Reliability). The Augmented Conditional MMD for reliability (ACMMD–Rel) between $\mathbb{P}_{|}$ and $Q_{|}$ as:

$$\begin{aligned} \text{ACMMD–Rel}(\mathbb{P}_{|}, Q_{|}) &:= \text{ACMMD}(\mathbb{P}_{|}^Q, Q_{|}^{\text{Rel}}) \\ &= \text{MMD}(\mathbb{P}_{|Q_{|X}} \otimes \mathbb{P}_{|}^Q, \mathbb{P}_{|Q_{|X}} \otimes Q_{|}) \end{aligned} \tag{VII.8}$$

where the ACMMD is evaluated with a user-specified kernel $k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ on $\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}$.

As for the ACMMD, we will restrict our attention to the case where $k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ is a tensor product kernel $k_{\mathcal{P}(\mathcal{Y})} \otimes k_{\mathcal{Y}}$ between a kernel on $\mathcal{P}(\mathcal{Y})$ and a kernel on \mathcal{Y} . Comparing the ACMMD–Rel with the ACMMD, we see that the former requires specifying a kernel *on the space of probability measures* on sequences $\mathcal{P}(\mathcal{Y})$ instead of a kernel on \mathcal{X} . Two important points must be addressed when working with such kernels. First, in order to have $\text{ACMMD–Rel}(\mathbb{P}_{|}, Q_{|}) = 0$ if and only if $Q_{|}$ is reliable, we must find universal kernels defined on $\mathcal{P}(\mathcal{Y})$. Second, as many kernels on probabilities are intractable, we must design an approximation strategy to estimate the ACMMD–Rel from data.

ACMMD–Rel can detect any pattern of unreliability Our first goal is to ensure that ACMMD–Rel can detect any pattern of unreliability. As ACMMD–Rel is a

specific instance of the ACMMMD, we can apply Lemma VII.3.2, stating that if the kernels $k_{\mathcal{P}(\mathcal{Y})}$ and $k_{\mathcal{Y}}$ are universal, then

$$\text{ACMMMD-Rel}(\mathbb{P}_|, Q_|) = 0 \iff Q_| \text{ is reliable.} \quad (\text{VII.9})$$

The task of finding a universal kernel $k_{\mathcal{Y}}$ on \mathcal{Y} was addressed in Section VII.3; thus, it remains to find a universal kernel $k_{\mathcal{P}(\mathcal{Y})}$ on $\mathcal{P}(\mathcal{Y})$. However, to the best of our knowledge, none of the existing kernels defined on probability distributions [33, 237, 239, 166, 246] have been shown to be universal when \mathcal{Y} is the space of arbitrary-length sequences. In the next proposition, we show that many such kernels can be constructed by following a simple recipe.

Proposition VII.4.1. *Let $k_{\mathcal{Y}}$ be a kernel on \mathcal{Y} vanishing at infinity (on $\mathcal{Y} \times \mathcal{Y}$). Suppose that $k_{\mathcal{Y}}$ has discrete masses, i.e. that $\delta_y \in \mathcal{H}_{\mathcal{Y}}$ for all sequences $y \in \mathcal{Y}$, where δ_y is the Dirac function at y , and let $\sigma > 0$. Then the kernel $k_{\mathcal{P}(\mathcal{Y})}$ on $\mathcal{P}(\mathcal{Y})$ defined as*

$$k_{\mathcal{P}(\mathcal{Y})}(q, q') := e^{-\frac{1}{2\sigma^2} \text{MMD}^2(q, q')}, \quad (\text{VII.10})$$

(where the MMD is computed in $\mathcal{H}_{\mathcal{Y}}$) is a C_0 -universal kernel on the space of probability distributions $\mathcal{P}(\mathcal{Y})$ (under the topology of convergence in distribution or Total Variation, which are identical, see [7]).

The proof, provided in Section VII.D.2, relies on an argument similar to prior work for universal kernels on probability measures [33], but tailored to the special case of sequences. Proposition VII.4.1 guarantees that any kernel on \mathcal{Y} vanishing at infinity with the discrete mass property [9] can be used to construct a universal kernel on $\mathcal{P}(\mathcal{Y})$. Kernels with discrete masses are studied in detail in [9]. In particular, the tilted Exponentiated Hamming Kernel $\frac{1}{|y||y'|} e^{-\lambda d_H(y, y')}$ (where $|y|$ is the length of the sequence y) is a kernel with discrete masses vanishing at infinity on $\mathcal{Y} \times \mathcal{Y}$, and can thus be used to construct a universal kernel on $\mathcal{P}(\mathcal{Y})$.

Estimating ACMMMD-Rel from data To estimate ACMMMD-Rel from the data $\{X_i, Y_i\}_{i=1}^N$ and samples from the model $\{\tilde{Y}_i \sim Q_{|X_i}\}_{i=1}^N$, one may try to use the general ACMMMD estimator proposed in Lemma VII.3.4, which, specialized to the

reliability setting, is given by:

$$\widehat{h}(Z_i, Z_j) := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h(Z_i, Z_j)$$

$$h(Z_i, Z_j) := k_{\mathcal{P}(\mathcal{Y})}(Q|_{X_i}, Q|_{X_j}) g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j))$$

This estimator requires evaluating $k_{\mathcal{P}(\mathcal{Y})}(Q|_{X_i}, Q|_{X_j})$ for pairs i, j . Unfortunately, exact evaluation of these quantities for the universal kernels proposed in Proposition VII.4.1 is in general impossible, as $\text{MMD}^2(Q|_{X_i}, Q|_{X_j})$ contains intractable expectations under $Q|_{X_i}$ and $Q|_{X_j}$. However, MMDs can be unbiasedly estimated using samples from $Q|_{X_i}$ and $Q|_{X_j}$ [90, 219]. Inspired by this fact, we propose the following estimator:

$$\widehat{\text{ACMMD-Rel}}^2 := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \widehat{h}(Z_i, Z_j) \quad (\text{VII.11})$$

$$\widehat{h}(Z_i, Z_j) := \widehat{k}_{ij} \times g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j))$$

Here, \widehat{k}_{ij} is an approximation of $k_{\mathcal{P}(\mathcal{Y})}(Q|_{X_i}, Q|_{X_j})$ obtained by drawing R samples $\{\tilde{Y}_i^r\}_{r=1}^R$ and $\{\tilde{Y}_j^r\}_{r=1}^R$ from $Q|_{X_i}$ and $Q|_{X_j}$, and replacing the $\text{MMD}^2(Q|_{X_i}, Q|_{X_j})$ term in $k_{\mathcal{P}}$ by an unbiased estimate $\widehat{\text{MMD}}_{ij}^2$ computed from these samples. The full estimation procedure is provided in Algorithm 12. This additional approximation step has several implications: first, unlike $\widehat{\text{ACMMD}}^2$, $\widehat{\text{ACMMD-Rel}}^2$ is not unbiased. However, the bias of this estimator can be controlled by increasing the number of samples R used to estimate the MMD. Moreover, we show in the next proposition that the estimator $\widehat{\text{ACMMD-Rel}}^2$ is still consistent provided that R is chosen appropriately.

Proposition VII.4.2. *Assume that $k_{\mathcal{Y}}$ is bounded. Then, if $R \equiv R(N)$, with $\lim_{N \rightarrow \infty} R(N) = +\infty$, $\widehat{\text{ACMMD-Rel}}^2$ converges in probability to ACMMD-Rel^2 as $N \rightarrow \infty$.*

Testing for reliability with ACMMD-Rel As an ACMMD, ACMMD-Rel has the potential to be used to test whether a model is reliable given some available data: to do so, one can use Algorithm 11, replacing $\widehat{\text{ACMMD}}^2$ by $\widehat{\text{ACMMD-Rel}}^2$,

Algorithm 12 Estimating ACMMD–Rel

Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_Y \otimes Q_Y$, model Q_Y
Parameters: kernel k_Y

```

for  $i$  in 1 to  $N$  do
     $[\tilde{Y}_i^r \sim Q_{|X_i}]$  for  $r$  in 1 to  $R$  ]
end for
for  $i, j$  in 1 to  $N$  do
    %Use, e.g. Gretton et al. [90, Equation 4]
     $\widehat{\text{MMD}}_{ij}^2 := \text{estimate\_mmd}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R)$ 
     $\hat{h}_{ij} := e^{-\frac{1}{2\sigma^2} \widehat{\text{MMD}}_{ij}^2} \times g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j))$ 
end for

return  $\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \hat{h}_{ij}$ 

```

and performing quantile estimation using the $\hat{h}(Z_i, Z_j)$ instead of the $h(Z_i, Z_j)$. A full description of the algorithm is provided in Appendix VII.D.3.1. An important question to answer is whether the approximation of using \hat{h} instead of h affects the false-rejection rate of the test. We show in the next proposition that this is not the case.

Proposition VII.4.3. *Assume that k_Y is bounded, and $k_{\mathcal{P}(Y)}$ is a kernel of the form of Equation VII.10. Then a reliability test using $\hat{h}(Z_i, Z_j)$ instead of $h(Z_i, Z_j)$ to estimate ACMMD–Rel and its $(1 - \alpha)$ -quantile under H_0 has a false-rejection rate of exactly α .*

VII.5 Related Work

Goodness-of-fit methods The *goodness-of-fit* problem is a well-studied problem in the statistics and machine learning literature, for which many methods were developed [42, 83, 86, 8, 17]. Impressively, these methods can operate directly from the model’s analytical form, without requiring access to samples from the model – which may be hard to generate. In these works, goodness-of-fit is defined as the problem of evaluating the fit of *unconditional* models to their data, which

is unlike the conditional goodness-of-fit problem we consider here. Evaluating conditional goodness-of-fit with kernels was recently studied in Jitkrittum et al. [128]. However, the proposed method requires the output space \mathcal{Y} to be a subset of \mathbb{R}^d , and is thus unsuitable for conditional sequence models. The use of conditional goodness-of-fit metrics to evaluate reliability was also done in Pierre Glaser et al. [246], in a method also limited to continuous output spaces. Finally, we note that ACMMD–Rel² recovers an existing calibration metric, the Squared Kernel Calibration Error (SKCE) of Widmann et al. [269]. However, the latter did not study the problems of universality, tractability and test validity in the case of sequence-valued outputs.

Deep Protein Design Models (Deep Learning–powered) conditional probability models have gained significant momentum in computational biology during the last decade. In particular, such models have revolutionized the protein design field [130, 19]. Inverse folding models are trained on protein structures and sequences in the protein data bank (PDB) [120, 111, 51]. They condition a sequence distribution on an input protein structure — thus learning what sequences would likely fold into that structure. The designs from these methods have been shown to be highly stable and retain function [234]. However, many of the leaps made using these models have used small, simple structural scaffolds (like loop-helix-loop motifs) [19, 265]. Protein engineers interested in leveraging these tools for novel scaffolds need to know how accurate and reliable the model is on average. If the model is too imprecise, one might wish to gather more data and train more bespoke models before using the method to design experiments.

VII.6 Experiments

We now investigate the behavior and utility of the ACMMD and ACMMD–Rel metrics and tests in practice. We start with a synthetic example showing that ACMMD is a natural measure of model distance. We then perform an extended analysis of a state-of-the-art inverse folding model, ProteinMPNN. We show that ACMMD can detect small perturbations in the model, and that it can be used to tune its temperature

parameter. Finally, we analyze the absolute performance of ProteinMPNN.

VII.6.1 A toy synthetic setting

We first study the behavior of ACMMMD and ACMMMD–Rel in a synthetic setting where the data distribution and the model are simple generative models on sequences. We set the input variable X to be a single scalar p drawn from some distribution \mathbb{P}_X with support in $(0.3, 0.5)$. Y is a sequence of arbitrary length with alphabet $\mathcal{A} = \{A, B, \text{STOP}\}$. We set the conditional distribution of Y given p to be:

$$p(y_n | y_{0:n-1}, x = p) = \begin{cases} A & \text{with probability } p \\ B & \text{with probability } p \\ \text{STOP} & \text{with probability } 1 - 2p \end{cases}$$

so long as $y_{n-1} \neq \text{STOP}$. The model distribution $Q|$ is the same as the data, except for the fact that the first factor $Q|_p(y_0)$ is perturbed by a parameter Δp :

$$Q|_p(y_0) = \begin{cases} A & \text{with probability } p - \Delta p \\ B & \text{with probability } p + \Delta p \\ \text{STOP} & \text{with probability } 1 - 2p \end{cases}$$

We set the kernel on \mathcal{Y} to be the exponentiated Hamming distance kernel $k_{\mathcal{Y}}(y, y') = e^{-d_H(y, y')}$, where $d_H(y, y')$ is the Hamming distance between y and y' , and $k_{\mathcal{X}}$ to be the Gaussian kernel $k_{\mathcal{X}}(p, p') = e^{-\frac{1}{2}(p-p')^2}$. With such choices, it is possible to show that:

$$\text{ACMMMD}(\mathbb{P}|, Q|) = C|\Delta p|$$

for some $C > 0$ that does not depend on Δp , and is computable in closed form for discrete priors on X . The proof and expression of C are given in Section VII.D.4. From this expression, we immediately see that ACMMMD is 0 only if $\Delta p = 0$, a manifestation of Lemma VII.3.2 which guarantees that the ACMMMD can detect any mismatch between the model and the data when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are universal. Moreover, in this case, the ACMMMD depends monotonically on the shift $|\Delta p|$. Since $|\Delta p|$

represents a natural measure of how different the model is from the data, this fact suggests that the ACMMMD is a natural, well-behaved measure of model distance. Additionally, we plot the average rejection rate of the ACMMMD test for various number of samples and shifts in Figure VII.6.1. The results for $\Delta p = 0$ confirm that our test has the correct specified type-I error rate (0.05). Moreover, we see that the power of the test increases with the number of samples, and the shift $|\Delta p|$.

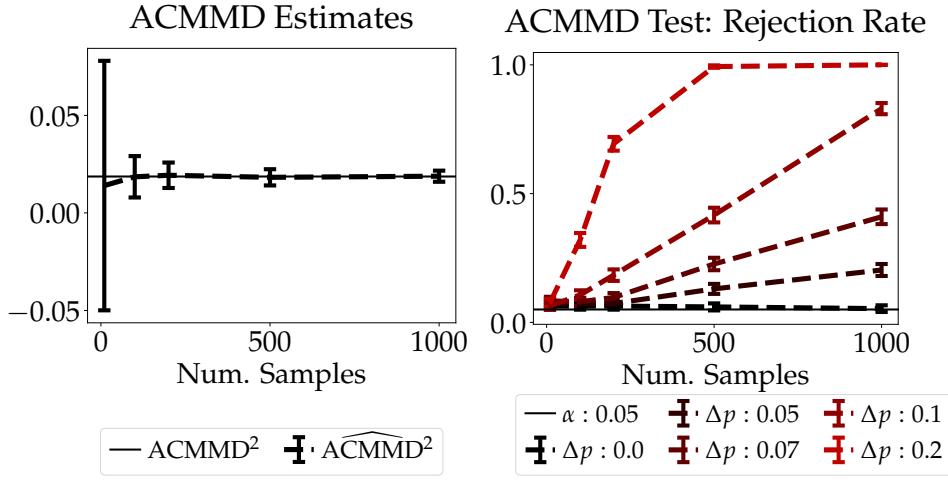


Figure VII.6.1: Left panel: ACMMMD estimates for a fixed shift value $\Delta p = 0.25$ and various number of samples in the synthetic example of Section VII.6.1. The analytic ACMMMD value is given by the horizontal line. Right panel: ACMMMD test average rejection rate for various number of samples and shifts in the same setting.

VII.6.2 ACMMMD Case Study: Inverse Folding Models

To demonstrate the utility of the ACMMMD measures and tests, we apply them to evaluate inverse folding models, a popular model framework used in protein design. Inverse folding models seek a distribution of amino acid sequences that are likely to fold into a given input three-dimensional structure, as discussed in Section VII.3. We focus our experiments on evaluating ProteinMPNN [51], a sampleable, commonly used model in this class. The sampling temperature T of ProteinMPNN can also be varied, letting the user control the trade-off between accuracy and diversity of the generated sequences.

Data We leveraged the CATH taxonomy to select a set of diverse (in sequence and structural topologies) protein structures to perform our ACMMMD test on. CATH

is a taxonomy of protein structures that categorizes proteins according to a hierarchy of structural organization [224]. We used the S60 redundancy filtered set which includes proteins that are at least 60% different in sequence identity from each other. Of these, we selected all single domain monomers (proteins where only one topological domain is found in the monomer), and removed any topologies that had fewer than 10 chains in its classification. This left us with 17,540 structures.

Choice of kernel Key to the performance of our metrics is the choice of the kernels $k_{\mathcal{X}}$, $k_{\mathcal{Y}}$ and $k_{\mathcal{P}(\mathcal{Y})}$. For $k_{\mathcal{Y}}$, we propose to use kernels that first embed each element – or *residue* – of a sequence y using an embedding function $\phi_{\mathcal{Y}} : \mathcal{A} \times \mathcal{Y} \mapsto \mathbb{R}^{d_{\mathcal{Y}}}$, and evaluating a euclidean kernel on $\mathbb{R}^{d_{\mathcal{Y}}}$ on the mean of the resulting embeddings, yielding a kernel of the form:

$$k_{\mathcal{Y}}(y, y') = k_{\mathbb{R}^{d_{\mathcal{Y}}}} \left(\frac{1}{|y|} \sum_{i=1}^{|y|} \phi_{\mathcal{Y}}(y_i, y), \frac{1}{|y'|} \sum_{i=1}^{|y'|} \phi_{\mathcal{Y}}(y'_i, y') \right)$$

where we noted $y = (y_1, \dots, y_{|y|})$. As the input space \mathcal{X} is also sequence-valued, we follow the same recipe to construct our a kernel $k_{\mathcal{X}}$, using an embedding function $\phi_{\mathcal{X}} : \mathbb{R}^3 \times \mathcal{X} \mapsto \mathbb{R}^{d_{\mathcal{X}}}$. Finally, for the kernel on $\mathcal{P}(\mathcal{Y})$, we will use a kernel of the form of Equation (VII.10), with kernel $k_{\mathcal{Y}}$ described above to compute the inner MMD. We set our embedding functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ to a pair of recent pre-trained neural networks that are commonly used for representation learning of protein sequences and structures: Gearnet [278] for structures, and ESM-2 [154] for sequences. Such two-step kernels allow us to instill the complex structure present in the distribution of protein structures and sequences within the ACMMD maximizing the performance and meaningfulness of our evaluation pipeline. Whether Proposition VII.4.1 holds for these kernels is an open question, but we find that they perform well in our experiments.

VII.6.2.1 The Discriminative power of ACMMD

We first propose to evaluate the behavior of the ACMMD and its associated test when comparing a known ground truth and a model distribution differing from the ground truth in a controlled manner. To this end, we set the ground truth to be a pre-trained

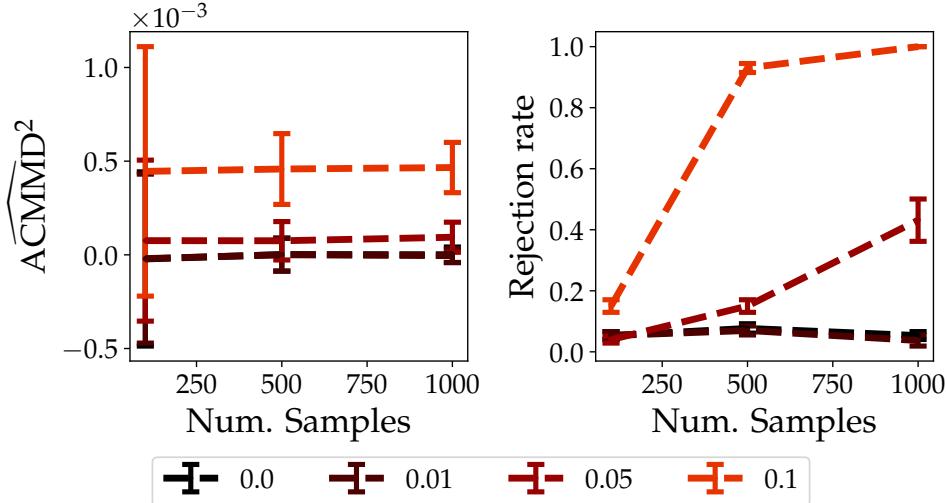


Figure VII.6.2: Values of $\widehat{\text{ACMMD}}^2$, (left) and of the average rejection rate of the ACMMMD test (right) in the setting described in VII.6.2.1. Each line corresponds to a different value for δT .

ProteinMPNN model $Q_{|}^T$ with temperature T , and the model to be the same model $Q_{|}^{T+\delta T}$ with temperature $T + \delta T$. As ProteinMPNN’s probability distribution is a continuous function of T , small changes in T result in small changes in the predicted distribution which will be hard to detect, translating into “low” values for $\widehat{\text{ACMMD}}^2$ relative to larger temperature changes. Conversely, we posit that large changes in T will result in large changes in the model distribution, and will be simpler to detect by the ACMMMD. To test these hypotheses, we performed an estimation of ACMMD^2 for a ground truth temperature $T = 0.1$ (the default in the ProteinMPNN documentation) and $\delta T \in \{0, 0.01, 0.05, 0.1\}$. We used the winged helix-like DNA binding domain superfamily (CATH ID: 1.10.10.10), and performed bootstrap sampling to produce dataset sizes ranging from 100 to 1000, and 100 different random seeds in order to obtain confidence intervals of our estimates.

The results are shown in Figure VII.6.2. As expected, the value of $\text{ACMMD}^2(Q_{|}^T, Q_{|}^{T+\delta T})$ robustly increases with increasing values of δT . Additionally, we performed the ACMMMD test of Section VII.3.2 with a target type-I error rate of $\alpha = 0.05$, and 100 permutations to estimate the $1 - \alpha$ quantile of the null distribution for the same values of N and δT , and computed the average rejection rate of the null hypothesis $H_0 : \text{ACMMD}^2(Q_{|}^T, Q_{|}^{T+\delta T}) > 0$, which, if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are universal, is equivalent

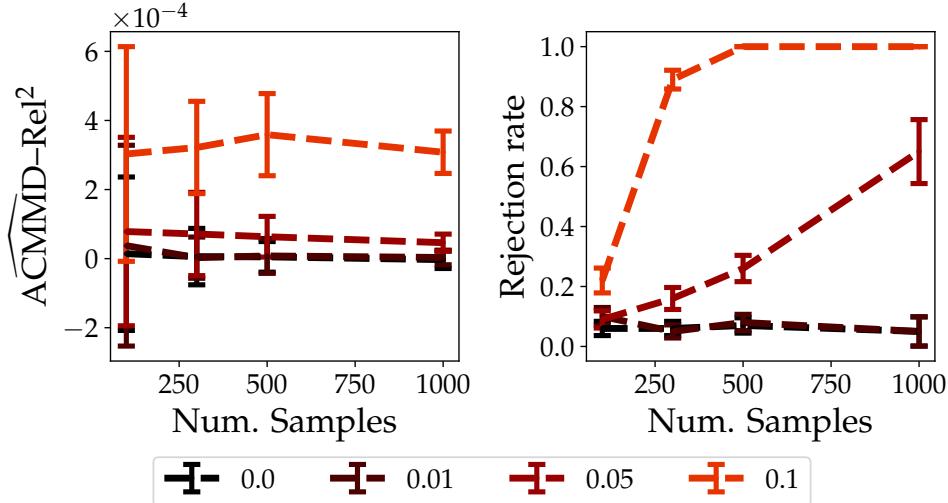


Figure VII.6.3: Values of $\widehat{\text{ACMMD-Rel}}^2$, (left) and of the average rejection rate of the ACMMMD-Rel test (right) in the setting described in Section VII.6.2.1. Each line corresponds to a different value for δT .

to $H_0 : \delta T = 0$. The results, shown in Figure VII.6.2 (right), empirically confirm that the ACMMMD test controls its type-I error rate and is able to detect differences in temperatures of an order relevant for ProteinMPNN. Similarly, we evaluate the behavior of the ACMMMD-Rel, which is used to assess the (lack of) reliability of between model $Q_{|}^{T+\delta T}$ w.r.t the data $\mathbb{P}_{|X} \otimes Q_{|}^T$, the assumption being that $Q_{|}^{T+\delta T}$ is not reliable when $\delta T \neq 0$. The results are shown in Figure VII.6.3, and exhibits similar behavior.

VII.6.2.2 Evaluation of ProteinMPNN on the CATH dataset

Now that we have confirmed the discriminative power of the ACMMMD on semi-synthetic data, we use our tests to evaluate ProteinMPNN against real-world protein structures and sequences from the CATH dataset. We perform a whole-data evaluation, using samples of 5000 proteins across all families in the dataset. Then we perform a fine-grain evaluation on a subset of CATH superfamilies.

Whole-data Evaluation We first study the deviation of ProteinMPNN from the true data by computing $\widehat{\text{ACMMD}}^2$ and estimating its mean and variance by bootstrapping over 10 random seeds. We find that ProteinMPNN with no temperature adjustment ($T = 1.0$) has an $\widehat{\text{ACMMD}}^2$ value of 0.0916 (and a p-value < 0.01). Comparing this

to the criterion values obtained on similar dataset sizes in the toy data experiments demonstrates that the model does not fit the test data. This suggests that there is still much room for improvement on solving the inverse folding problem.

On optimal temperature choices for ProteinMPNN Practitioners vary the sampling temperature as a heuristic method for sampling more certain sequences from ProteinMPNN; lower temperature settings have been found to generate sequences with fewer unrealistic artifacts (e.g. runs of alanines) which fold to more stable structures [234]. However, the relationship between sampling temperature, model reliability, and design accuracy has not been fully established. To thoroughly evaluate this, designs from different sampling temperatures conditioned on a diverse set of backbone structures would need to be experimentally characterized, which is resource intensive in practice. We leverage the ACMMMD to understand at what sampling temperature ProteinMPNN best fits the data, which gives insight as to what temperature is optimum, by computing $\widehat{\text{ACMMD}}^2$ and $\widehat{\text{ACMMD-Rel}}^2$ for varying temperature values across 10 seeds for each temperature value. The results are shown in Figure VII.6.4.

First, we observe that reducing the temperature below 1.0 improves both the model’s goodness-of-fit and its reliability. This corroborates the empirical design success of lowering the sampling temperature, suggesting that greater model fit may increase the quality of samples from the model. The decrease in reliability at higher temperature shows that even though increasing the temperature increases the diversity of the model’s predictions, this diversity does not necessarily capture the one of the data distribution, as for instance the prior would. The optimal temperature from the perspective of goodness-of-fit is 0.4 (which lies outside the suggested temperature range of 0.1-0.3 in the ProteinMPNN documentation [51]). However, we notice that model reliability continues to improve with even lower sampling temperatures while accuracy slightly increases, suggesting a trade-off between reliability and accuracy. Further experiments will determine how this trade-off manifests in the quality of designs from low-temperature settings.

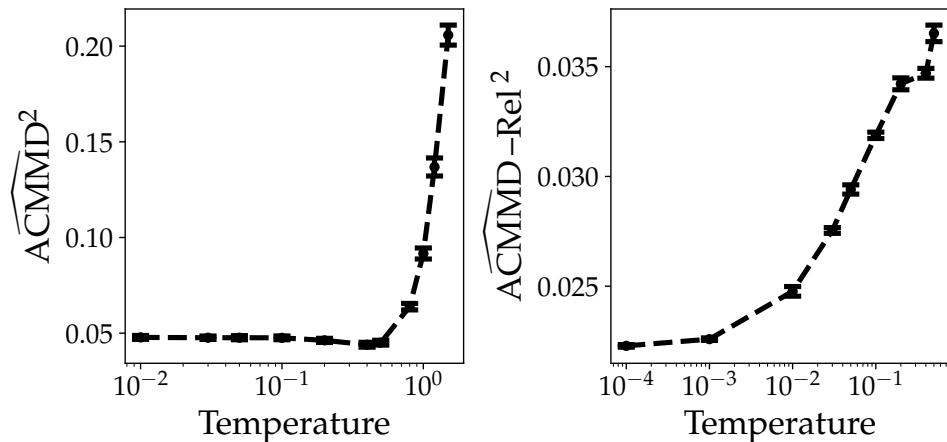


Figure VII.6.4: Evolution of $\widehat{\text{ACMMD}}^2$ (left) and $\widehat{\text{ACMMD-Rel}}^2$ (right) between a pre-trained ProteinMPNN model and the CATH S60 reference dataset, for varying temperature values

Structural superfamily evaluation The (H)omologous superfamily tier within the CAT(H) hierarchy groups proteins with the same similar folds and sequence identity. While ProteinMPNN has shown great performance in designing particular structural scaffolds, a practitioner aiming to leverage this model on a yet untested structural family may want some insight as to how well ProteinMPNN may fit the distribution of proteins they are interested in. Thus, we performed ACMMMD evaluation separately on individual superfamilies contained in our dataset to gain insights on what types of structures ProteinMPNN does or does not fit well. We filtered the superfamilies for groupings with at least 500 proteins under a length of 100, yielding 11 families. The results are shown in Figure VII.6.5. We find that the model fit varies across families and the fit ranking is largely maintained at different temperatures. With no temperature adjustment ($T = 1.0$) the best fit superfamily (lowest $\widehat{\text{ACMMD}}^2$) is the Homeodomain-like proteins (CATH ID: 1.10.10.60). These structures are largely dominated by helical bundles - a class of proteins that ProteinMPNN has demonstrated success on designing [51, 265, 19]. While the Immunoglobulin superfamily has the highest fit at lower sampling temperatures, we note that most of an immunoglobulin structure consists of the beta sandwich of the framework, while, for antibody design, engineers are often most interested in the unstructured complementarity determining regions (CDRs) of antibodies [143, 155, 122]. As the criterion

is calculated across the entire sequence, this may not reflect that ProteinMPNN has learned the distribution of CDR loops well. Further work will extend these tests to focus on subsequences of a domain to answer specific questions of model fit on regions of interest.

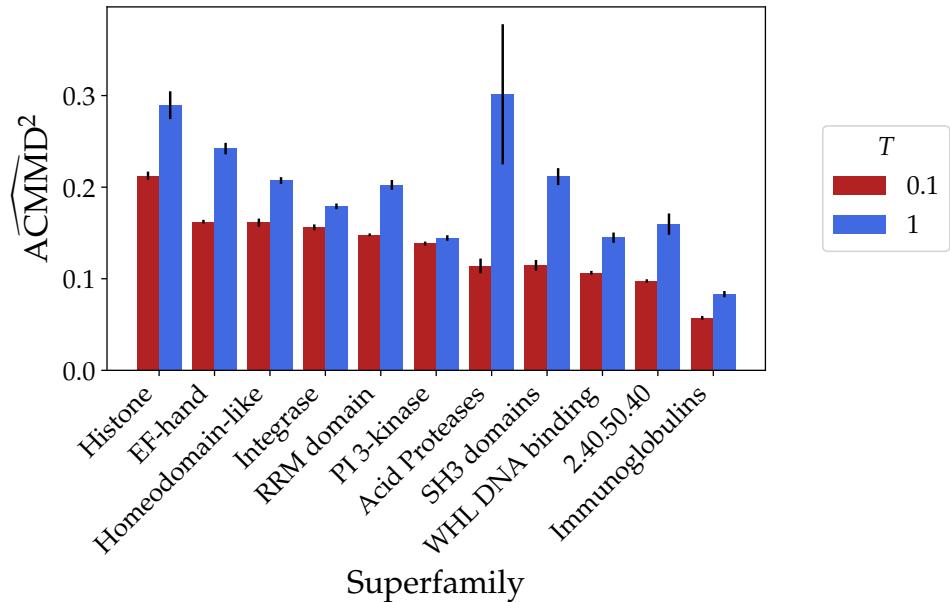


Figure VII.6.5: Estimated value $\widehat{\text{ACMMD}}^2$ between ProteinMPNN and the CATH S60 reference dataset on a subset of 10 superfamilies for two different temperatures $T = 1.0$ and $T = 0.1$.

VII.7 Discussion

Advancing the computational evaluation of conditional sequence models is crucial for accelerating the development of these methods for protein engineering. Given the limitations of current evaluation methods, and leveraging recent advancements in kernel methods for designing tests of goodness-of-fit and calibration, we propose a criterion and its associated test to principledly evaluate protein sequence models for how well they have learned input-conditioned sequence distributions. We discuss the statistical properties of our metrics and develop testing frameworks from them. Finally, we leverage them to investigate the performance of inverse folding models under default and temperature-adjusted settings. We develop novel insights on ideal temperature settings for ProteinMPNN and discuss the trade-off between design

accuracy and model calibration that our tests demonstrate for lower temperatures. Future work can perform a more fine-grained evaluation, for example investigating which structures in particular cause the model to make unreliable predictions and what features of the model’s predictions do not match the data through the use of witness functions, a by-product of MMDs [158]. We also note that protein engineering goals may differ from pure modeling goals, and whether performance under our metrics reflect experimental design success rates requires further investigation to determine. Yet, barring orthogonal *in silico* validation data or experimental testing, our methods offer a powerful framework to test conditional sequence models for desirable statistical properties.

Impact Statement

The tools developed in this work assess the quality of sequences predictors. As such, they have the potential to influence various procedures in protein design, and, on longer timescales, healthcare. However, the conclusions that they provide are only statistical: while they are guaranteed to hold on average, they will not hold every time. Such tools should thus be used with caution, and in conjunction with external help from domain experts to ensure that the real-world actions they will influence remain beneficial to society.

Appendix

Supplementary Material of the paper *Kernel-Based Evaluation of Conditional Biological Sequence Models*

VII.A Proof of Lemma VII.3.2

Let us first re-state the lemma in its complete form.

Lemma (Complete form of Lemma VII.3.2). *Assume that \mathcal{X} is locally-compact and second countable. Moreover, assume that $k_{\mathcal{X} \times \mathcal{Y}} = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$, and that $k_{\mathcal{X}}, k_{\mathcal{Y}}$ satisfy the integrability conditions $\mathbb{E}[k_{\mathcal{X}}(X, X)k_{\mathcal{Y}}(Y, Y)] < +\infty$ and $\mathbb{E}[k_{\mathcal{Y}}(Y, Y)] < +\infty$ (and similarly for \tilde{Y}). Then,*

$$\text{ACMMD}^2(\mathbb{P}_{|}, Q_{|}) = \left\| T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) \right\|_{\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}}^2$$

Where $\mu_{\mathbb{P}_{|}}$ and $\mu_{Q_{|}}$ are the conditional mean embeddings [192] of $\mathbb{P}_{|}$ and $Q_{|}$, given by: $\mu_{\mathbb{P}_{|}} : x \mapsto \mathbb{E}_{y \sim \mathbb{P}_{|x}} k_{\mathcal{Y}}(y, \cdot)$ (and similarly for $Q_{|}$), $K_{\mathcal{X}}(x, x') := k_{\mathcal{X}}(x, x')I_{\mathcal{H}_{\mathcal{Y}}}$ is an operator-valued kernel with associated vector-valued RKHS $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$, and $T_{K_{\mathcal{X}}}$ is its associated integral operator from $L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}})$ to $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$, defined as

$$T_{K_{\mathcal{X}}} f(x) = \int_{\mathcal{X}} K_{\mathcal{X}}(x, x') f(x') \mathbb{P}_X(dx') \in \mathcal{H}_{\mathcal{Y}}$$

for all $f \in L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ and $x \in \mathcal{X}$. Moreover, if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are C_0 -universal³

$$\text{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0 \iff \mu_{\mathbb{P}_{|x}} = \mu_{Q_{|x}}, \quad \mathbb{P}_X\text{-a.e.}$$

Proof. Let us introduce the notations used in this proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the sample space, $X(\omega), Y(\omega), \tilde{Y}(\omega)$ being random variables on Ω corresponding to the input, target and the model. When clear, we will identify the measure \mathbb{P} and the push-forwards $Y_{\#}\mathbb{P}, \tilde{Y}_{\#}\mathbb{P}$ and drop the dependence of Y, \tilde{Y} on ω . Given $x \in \mathcal{X}$, we write K_x the linear operator from $\mathcal{H}_{\mathcal{Y}}$ to $\mathcal{L}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$, the space of linear operators from \mathcal{X} to $\mathcal{H}_{\mathcal{Y}}$, such that $(K_x f)(x') = K_{\mathcal{X}}(x, x')f \in \mathcal{H}_{\mathcal{Y}}$ for all $f \in \mathcal{H}_{\mathcal{Y}}$. When no confusion is possible, we may identify the notations $k_{\mathcal{Y}}(y, \cdot)$ and k_y .

³A kernel k is C_0 -universal if the associated RKHS \mathcal{H}_k is dense in $C_0(\mathcal{X})$, the space of continuous functions on \mathcal{X} vanishing at infinity [230]

The existence of the conditional mean embeddings $\mu_{\mathbb{P}_|}$ and $\mu_{Q|}$ is guaranteed by [192, Definition 3.1] under the integrability assumption $\int k_{\mathcal{Y}}(y, y) d\mathbb{P}(y) < +\infty$ and $\int k_{\mathcal{Y}}(\tilde{y}, \tilde{y}) d\mathbb{P}(\tilde{y}) < +\infty$. The second integrability assumption guarantees the existence of the mean embedding $\mu_{\mathbb{P}_X \otimes \mathbb{P}_|}$, defined as:

$$\int k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot) d(\mathbb{P}_X \otimes \mathbb{P}_|)(x, y) \in \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$$

by [90, Lemma 3] (and respectively for $Q|$). Here, $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is the tensor product Hilbert space of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$, with kernel $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$. We actually prove a stronger form of the lemma, given by removing the norm from both hands of the equality and replacing it with a suitable isometric isomorphism $\phi : \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$

$$\phi\left(\int k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot) d(\mathbb{P}_X \otimes \mathbb{P}_|)(x, y)\right) = \int K_x \mu_{\mathbb{P}_|} d\mathbb{P}_X(x)$$

This isometric isomorphism is shown to exist in the “Currying lemma” of Carmeli et al. [33, Example 6] regarding tensor product kernels (note that both \mathcal{X} – by assumption – and \mathcal{Y} are locally compact and second-countable). This lemma shows that the mapping:

$$\phi : \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}} \rightarrow \mathcal{F}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$$

$$f \otimes g \mapsto \phi(f \otimes g) = (x \in \mathcal{X} \mapsto f(x)g \in \mathcal{H}_{\mathcal{Y}})$$

is an isometric isomorphism between $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ and $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$. This lemma gives both a representation formula for elements of $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$, and a way to formalize the currying operation, (e.g. the transformation of a function of two variables into a higher-order function of one variable and returning a function of one variable) on tensor-product spaces, since $(f \otimes g)(x, y) = (\phi(f \otimes g)(x))(y)$. We refer to Carmeli et al. [33, Example 6] for a proof. Proceeding with the proof of Lemma VII.3.2, when f and g are kernel functions $k_{\mathcal{X}}(x, \cdot)$ and $k_{\mathcal{Y}}(y, \cdot)$, the right-hand side of the

equality can be related to K_x as

$$\begin{aligned}\phi(k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))(x') &= k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, \cdot) \\ &= K_{x'}^* K_x k_{\mathcal{Y}}(y, \cdot) \\ &= K_x k_y(x')\end{aligned}$$

where the second to last equality follows from the reproducing property of $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$. Since ϕ is linear and unitary, it commutes with the mean embedding operation: [57, Theorem 36], yielding:

$$\begin{aligned}\phi\left(\int (k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))d(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y)\right) &= \int \phi(k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))d(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y) \\ &= \int K_x k_y d(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y)\end{aligned}$$

To complete the proof, it remains to relate the right-hand side to the conditional mean embedding $\mu_{\mathbb{P}_{|}}$, using

$$\begin{aligned}\int K_x k_y d(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y) &= \iint K_x k_y d\mathbb{P}_X(x) d\mathbb{P}_{|x}(y) \\ &= \int K_x \int k_y d\mathbb{P}_{|x}(y) d\mathbb{P}_X(x) \\ &= \int K_x \mu_{\mathbb{P}_{|}}(x) d\mathbb{P}_X(x)\end{aligned}$$

as K_x is a bounded linear operator. We thus have that:

$$\phi\left(\int k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot)d(\mathbb{P}_{|X} \otimes \mathbb{P}_{|})(x, y)\right) = \int K_x \mu_{\mathbb{P}_{|}} d\mathbb{P}_X(x)$$

Combining this with the analogue of this result holding for $\mu_{Q_{|}}$ allows to show the stronger form of Lemma VII.3.2. Let us now prove the second part of the lemma. Assume $\text{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$, meaning

$$T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) = 0$$

By Carmeli et al. [33, Theorem 2] $K_{\mathcal{X}}$ is a C_0 -universal operator-valued kernel, the operator $T_{K_{\mathcal{X}}}$ is injective. This implies that the conditional mean embeddings of

$\mathbb{P}_{|x}$ and $Q_{|x}$ are equal \mathbb{P}_X -almost everywhere. By Park and Muandet [192, Theorem 5.2] applied to the case where the marginals are equal, and since $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is C_0 -universal, this implies that $\mathbb{P}_{|x} = Q_{|x}$, \mathbb{P}_X -almost everywhere, and in summary, $\text{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$ implies $\mathbb{P}_{|x} = Q_{|x}$, \mathbb{P}_X -almost everywhere. To prove the reverse direction, assume that $\mathbb{P}_{|x} = Q_{|x}$, \mathbb{P}_X -almost everywhere. Since Park and Muandet [192, Theorem 5.2] also prove the reverse direction of the statement relied upon in the previous argument, we have that conversely $\mu_{\mathbb{P}_{|}}(x) = \mu_{Q_{|}}(x)$, \mathbb{P}_X -almost everywhere. By linearity of $T_{K_{\mathcal{X}}}$, we thus have that $T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) = 0$, and therefore $\text{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$. \square

VII.B Asymptotic distribution of $\widehat{\text{ACMMD}}^2$

As discussed in the main text, it is possible to characterize the asymptotic distribution of $N\widehat{\text{ACMMD}}^2$. When $\mathbb{P}_{|} = Q_{|}$, and $\sqrt{N}(\widehat{\text{ACMMD}}^2 - \text{ACMMD}^2)$ when $\mathbb{P}_{|} \neq Q_{|}$. This characterization is given in the next lemma.

Lemma VII.B.1. *Assume that the integrability assumptions of Lemma VII.3.2 hold, and that $\mathbb{E}_{Z_1, Z_2} h(Z_1, Z_2)^2 < +\infty$, and that*

- if $\mathbb{P}_{|x} = Q_{|x}$ $\mathbb{P}(X)$ -a.s, then $\mathbb{E}[\widehat{\text{ACMMD}}] = 0$ and

$$N\widehat{\text{ACMMD}}^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1)$$

where $\{\chi_{1j}^2\}_{j=1}^{\infty}$ are independent random χ_1^2 variables, and λ_j are the eigenvalues of the operator defined as:

$$\phi \longmapsto \int h(z, \cdot) \phi(z) d\mathbb{P}(z)$$

- Assume moreover that $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are C_0 -universal kernels, and that $\sigma_h^2 = 4\mathbb{V}_{Z_2}[\mathbb{E}_{Z_1} h(Z_1, Z_2)] > 0$. Then

$$\sqrt{N}(\widehat{\text{ACMMD}}^2 - \text{ACMMD}^2) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2)$$

Proof. Since $\mathbb{E}_{Z_1, Z_2} h(Z_1, Z_2)^2 < +\infty$ we have that $\mathbb{V}_{Z_1, Z_2} \mathbb{E}_{Z_1, Z_2} h(Z_1, Z_2)^2 < +\infty$. Let us define, as in Serfling [221, Section 5.1.5], the function $h(z) = \mathbb{E}_{Z_2} h(z, Z_2)$, and define $\zeta := \mathbb{V}_z h$. For the first point, we will show that if $\mathbb{P}_| = Q|$, \mathbb{P} -a.s, then $\zeta = 0$, and the result will follow from Serfling [221, Section 5.5.2]. Indeed, noting $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$,

$$\begin{aligned} h(z) &= \mathbb{E}_{Z_2} h(z, Z_2) \\ &= \mathbb{E}_{Z_2} \langle k((x, y), \cdot) - k((x, y'), \cdot), k((X_2, Y_2), \cdot) - k(X_2, \tilde{Y}_2) \rangle \\ &= \langle k((x, y), \cdot), -k((x, y'), \cdot), \mathbb{E}_{Z_2} [k((X_2, Y_2), \cdot) - k(X_2, \tilde{Y}_2), \cdot] \rangle \end{aligned}$$

Where we exchanged the order of integration and inner product, which is possible since $h \mapsto \langle k((x, y), \cdot) - k((x, \tilde{y}), \cdot), h \rangle$ is a bounded linear functional for all (x, y, \tilde{y}) . Now,

$$\mathbb{E}_{Z_2} k((X_2, Y_2), \cdot) - k((X_2, \tilde{Y}_2), \cdot) = \mathbb{E}_{\mathbb{P}_X} [\mathbb{E}_{\mathbb{P}|} k((X_2, Y_2), \cdot) - \mathbb{E}_{Q|} k((X_2, \tilde{Y}_2), \cdot)] = 0$$

since $\mathbb{P}_{|x} = Q_{|x}$ \mathbb{P}_X -a.s. Thus, $h(z)$ is a constant function, and $\zeta = 0$. The second case follows by assumption from Serfling [221, Section 5.1.1]. \square

VII.B.1 Proof of Lemma VII.3.3

Let $\mathcal{H}_{\mathcal{X} \times \mathcal{Y}} := \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ be the tensor-product RKHS of functions from $\mathcal{X} \times \mathcal{Y}$ with kernel $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$. The result can be obtained by applying a “coupling” argument, and starting from the following object:

$$\begin{aligned} \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q|} &:= \int (k((x(\omega), y(\omega)), \cdot) - k((x(\omega), \tilde{y}(\omega)), \cdot)) d\mathbb{P}(\omega) \\ &= \mathbb{E}_{x, y, \tilde{y}} [k((x, y), \cdot) - k((x, \tilde{y}), \cdot)] \end{aligned} \tag{VII.12}$$

We first show that $\mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q|}$ is a well-defined element of $\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}$. Indeed, the following operator

$$T : f \in \mathcal{H} \mapsto \mathbb{E}_z f((x, y)) - f((x, \tilde{y}))$$

satisfies

$$\begin{aligned} |Tf| &\leq \mathbb{E} [|f(x, y)| + |f(x, \tilde{y})|] \\ &\leq \|f\|_{\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}} (\mathbb{E} \sqrt{k((x, y), (x, y))} \\ &\quad + \mathbb{E} \sqrt{k((x, \tilde{y}), (x, \tilde{y}))}) \end{aligned}$$

and is bounded thanks to the integrability assumptions of Lemma VII.3.2. Applying the same argument as [90, Lemma 3], it follows that the object in Equation (VII.12) is well-defined and belongs to $\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}$. Furthermore, by linearity of integration, we have that:

$$\begin{aligned} &\int (k((x(\omega), y(\omega)), \cdot) - k((x(\omega), \tilde{y}(\omega)), \cdot)) d\mathbb{P}(\omega) \\ &= \int k((x(\omega), y(\omega)), \cdot) d\mathbb{P}(\omega) - \int k((x(\omega'), \tilde{y}(\omega')), \cdot) d\mathbb{P}(\omega') \\ &= \mu_{\mathbb{P}_X \otimes \mathbb{P}_|} - \mu_{\mathbb{P}_X \otimes Q_|} \end{aligned}$$

To conclude, note that:

$$\begin{aligned} \text{ACMMD}(\mathbb{P}_|, Q_|)^2 &= \left\| \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|} \right\|_{\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}}^2 \\ &= \left\langle \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|}, \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|} \right\rangle_{\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}} \\ &= \left\langle \mathbb{E}_{x, y, \tilde{y}} [k((x, y), \cdot) - k((x, \tilde{y}), \cdot)], \mathbb{E}_{x, y, \tilde{y}} [k((x, y), \cdot) - k((x, \tilde{y}), \cdot)] \right\rangle_{\mathcal{H}_{\mathcal{X} \times \mathcal{Y}}} \\ &= \mathbb{E}_{x_1, y_1, \tilde{y}_1} \mathbb{E}_{x_2, y_2, \tilde{y}_2} h((x_1, y_1, \tilde{y}_1), (x_2, y_2, \tilde{y}_2)) \end{aligned}$$

Where the last equality was obtained by exchanging the order of integration and dot product, possible thanks to the integrability assumptions of Lemma VII.3.2, by using the bilinearity of the inner product and the reproducing property of the kernel k . The symmetry of h in (Z_1, Z_2) follows from the symmetry of $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$.

VII.B.2 Proof of Lemma VII.3.4

Proof. The proof of the unbiasedness of $\widehat{\text{ACMMD}}^2$ follows by linearity of the expectation, and that each $h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j))$ is an unbiased estimator of $\text{ACMMD}^2(\mathbb{P}_|, Q_|)$. \square

VII.C Type-I error control of the ACMMMD test

The goal of this section is to show that the ACMMMD test is guaranteed to control its type-I error rate at level α .

VII.C.1 Quantile estimation and Decision Rule

We first fully specify the way we compute our quantile estimate $\hat{q}_{1-\alpha}$. Let $b_\alpha := \lceil (1 - \alpha)(B + 1) \rceil$. Given B bootstrap samples $\{\widetilde{\text{ACMMMD}}_b^2\}_{b=1}^B$ and an $\widehat{\text{ACMMMD}}^2$ estimate, we order them in increasing order in a sequence of size $B + 1$, with ties broken arbitrarily. Let $m = \min\{b \in \llbracket 1, B + 1 \rrbracket \mid \widetilde{\text{ACMMMD}}_b^2 = \widetilde{\text{ACMMMD}}_{b_\alpha}^2\}$, and $M = \max\{b \in \llbracket 1, B + 1 \rrbracket \mid \widetilde{\text{ACMMMD}}_b^2 = \widetilde{\text{ACMMMD}}_{b_\alpha}^2\}$. We set $\hat{q}_{1-\alpha}$ to be the $(m - 1)$ -th element with probability $(b_\alpha - (1 - \alpha)(B + 1))/(M - m + 1)$ (with the convention that the 0-th element is $-\infty$), and the b_α -th element otherwise. The decision rule is then to reject the null hypothesis if $\widehat{\text{ACMMMD}}^2 > q_{1-\alpha}$.

VII.C.2 Wild-bootstrap and permutation-based approaches are equivalent in the ACMMMD test

To show that the ACMMMD test is guaranteed to control its type-I error rate at level α , we show that the use of a wild bootstrap procedure in the ACMMMD test can be cast as a computationally efficient way to approximate the quantiles of the random variable $\widehat{\text{ACMMMD}}^2$ when $\mathbb{P}_{|x} = Q_{|x} \mathbb{P}_X$ -a.e.

Lemma VII.C.1. *Let $\{Z_i\}_{i=1}^N$ be i.i.d realizations of $\mathbb{P}_X \otimes \mathbb{P}_| \otimes Q_|$, and let $\{W_i^b\}_{i=1}^{B=1..N}$ be i.i.d. Rademacher random variables independent of the data. Given a function $\sigma : \llbracket 1, N \rrbracket \mapsto \{-1, 1\}$, define $\{Z_i^\sigma\}_{i=1}^N := \{X_i, Y_i^\sigma, \tilde{Y}_i^\sigma\}_{i=1}^N$, where $(Y_i^\sigma, \tilde{Y}_i^\sigma) = (Y_i, \tilde{Y}_i)$ if $\sigma(i) = 1$, and (\tilde{Y}_i, Y_i) otherwise. Then we have:*

$$\widetilde{\text{ACMMMD}}_b^2 = \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i < j}}^N h(Z_i^{\sigma_b}, Z_j^{\sigma_b}) := \widehat{\text{ACMMMD}}_{\sigma_b}^2$$

for $\sigma_b(i) := W_i^b$.

The W_i^b should be understood as elements of a random swap σ_b , which for each i , swaps Y_i and \tilde{Y}_i with probability $1/2$.

Proof. Without loss of generality, we fix $i = 1$ and $j = 2$, and fix b , dropping the b index. Note that $h(Z_1, Z_2)$ and $h(Z_1^\sigma, Z_2^\sigma)$ share the same $k_{\mathcal{X}}(X_1, X_2)$. The only differing term is

$$g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) := k_{\mathcal{Y}}(Y_1, Y_2) + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2),$$

and we only need to show that $W_1 W_2 g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) = g((Y_1^\sigma, \tilde{Y}_1^\sigma), (Y_2^\sigma, \tilde{Y}_2^\sigma))$.

Case $W^1 = W^2 = 1$ In that case, $Z_1 = Z_1^\sigma$ and $Z_2 = Z_2^\sigma$, and $W_1 W_2 h(Z_1, Z_2) = h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$, by definition of σ .

Case $W_1 = W_2 = -1$ In that case, we have:

$$g((Y_1^\sigma, \tilde{Y}_1^\sigma), (Y_2^\sigma, \tilde{Y}_2^\sigma)) = k(\tilde{Y}_1, \tilde{Y}_2) + k(Y_1, Y_2)k(\tilde{Y}_1, Y_2) - k(Y_1, \tilde{Y}_2) = h(Z_1, Z_2)$$

implying again $W_1 W_2 h(Z_1, Z_2) = h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$.

Case $W_1 = 1$ and $W_2 = -1$ In that case, we have:

$$\begin{aligned} h(Z_1^\sigma, Z_2^\sigma) &= k((X_1, Y_1), (X_2, \tilde{Y}_2)) + k((X_1, \tilde{Y}_1), (X_2, Y_2)) - k((X_1, Y_1), (X_2, Y_2)) \\ &\quad - k((X_1, \tilde{Y}_1), (X_2, \tilde{Y}_2)) \\ &= -h(Z_1, Z_2) \end{aligned}$$

meaning again $W_1 W_2 h(Z_1, Z_2) = -h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$, and the last case is proved similarly. \square

VII.C.3 Level of the ACMMMD test

We now show that the ACMMMD test has the desired type-I error rate.

Lemma VII.C.2. *Assume that $\mathbb{P}_{|x} = Q_{|x}$ \mathbb{P}_X -a.s. Then the probability that the ACMMMD test rejects the null hypothesis is exactly α .*

The proof consists in 2 steps. First, we show that the decision rule is equivalent to a simpler one. Then, we analyze the latter decision rule.

An equivalent decision rule This decision rule is equivalent to the one rejecting H_0 if the position Q (with ties broken uniformly at random) of $\widehat{\text{ACMMMD}}^2$ in that

sequence satisfies $Q > b_\alpha$, accepting it if $Q < b_\alpha$, and rejecting it with probability $b_\alpha - (1 - \alpha)(B + 1)$ if $Q = b_\alpha$: Indeed, $Q > M \iff \widehat{\text{ACMMMD}}^2 > q_{1-\alpha}$ (we always reject), $Q < m \iff \widehat{\text{ACMMMD}}^2 \leq q_{1-\alpha}$ (we never reject), and for both rules, when the random position Q is in $\llbracket m, M \rrbracket$, the null is rejected with probability $(b_\alpha - \alpha(B + 1))/(M - m + 1)$.

Analysis of the decision rule We derive the type-I error of our decision rule by analyzing the equivalent, latter one. Our analysis follows a similar argument, in flavor, as Domingo-Enrich et al. [59, Appendix C]. Now, recall that from Lemma VII.C.1, the wild bootstrap quantile estimation are draws of $\widehat{\text{ACMMMD}}$ on swapped samples Z^σ , e.g. $\{X_i, Y_i^\sigma, \tilde{Y}_i^\sigma\}_{i=1}^N$ parameterized by $\sigma : \llbracket 1, N \rrbracket \mapsto \{-1, 1\}$ where $Y_i^\sigma = Y_i$ if $\sigma(i) := w_i$ and $Y_i^\sigma = \tilde{Y}_i$ otherwise:

$$(\widetilde{\text{ACMMMD}}_b^2)_{b=1}^B = (\widehat{\text{ACMMMD}}_{\sigma_b}^2)_{b=1}^B$$

using the notation of Lemma VII.C.1. Note that σ is a random swap operator such that $\sigma(i) = 1$ with probability 0.5, and $\sigma(i) = -1$ with probability 0.5. If $\mathbb{P}_{|x} = Q_{|x}$ a.e., then since the B swap maps $\sigma_1, \dots, \sigma_B$ are i.i.d. let us note $\widehat{\text{ACMMMD}}_{\sigma_0}^2 = \widehat{\text{ACMMMD}}^2$, e.g. $\sigma_0(i) = 1$. Then the random sequence $(\widehat{\text{ACMMMD}}_{\sigma_b}^2)_{b=0}^B$ is exchangeable. Since Q is the position of $\widehat{\text{ACMMMD}}^2$ within that sorted sequence, and that all positions are equally likely under exchangeability, we have:

$$\begin{aligned}\mathbb{P}[Q < m] &= 1/(B + 1) \\ \mathbb{P}[Q > b_\alpha] &= (B + 1 - b_\alpha)/(B + 1) \\ \mathbb{P}[Q < b_\alpha] &= (b_\alpha - 1)/(B + 1)\end{aligned}$$

Noting $\Delta((X^i, Y^i, \tilde{Y}^i)_{i=1}^N)$ the event that the null hypothesis is rejected, we have:

$$\begin{aligned}\mathbb{P}[\Delta((X^i, Y^i, \tilde{Y}^i)_{i=1}^N)] &= \mathbb{P}[Q > b_\alpha] + \mathbb{P}[Q = b_\alpha] \mathbb{P}[\text{Reject} | Q = b_\alpha] \\ &= (B + 1 - b_\alpha)/(B + 1) + (b_\alpha - (1 - \alpha)(B + 1))/(B + 1) \\ &= \alpha,\end{aligned}$$

thus showing that the ACMMD test has the desired type-I error rate. \square

VII.D Proofs related to ACMMD–Rel

VII.D.1 Differences between the SKCE U-statistics and the ACMMD U-statistic

We recall the definition of the SKCE U-statistics estimator from [269, Lemma 2]:

$$\widehat{\text{SKCE}} = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} G((Q|_{X_i}, Y_i), (Q|_{X_j}, Y_j)) \quad (\text{VII.13})$$

where

$$\begin{aligned} G((q, y), (q', y')) \\ := k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}((q, y), (q', y')) - \mathbb{E}_{Y \sim q} k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}((q, Y), (q', y')) \\ - \mathbb{E}_{Y' \sim q'} k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}((q, y), (q', Y')) + \mathbb{E}_{Y \sim q} \mathbb{E}_{Y' \sim q'} k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}((q, Y), (q', Y')) \\ = k_{\mathcal{P}(\mathcal{Y})}(q, q') \times (k_{\mathcal{Y}}(y, y') - \mathbb{E}_{Y \sim q} k_{\mathcal{Y}}(Y, y') - \mathbb{E}_{Y' \sim q'} k_{\mathcal{Y}}(y, Y') \\ + \mathbb{E}_{Y \sim q} \mathbb{E}_{Y' \sim q'} k_{\mathcal{Y}}(Y, Y')) \end{aligned} \quad (\text{VII.14})$$

Where the second equality holds when focusing on tensor product kernels. Comparing Equation (VII.13) and Equation (VII.14) with the expression of the ACMMD U-statistics estimator given in Equation (VII.5) and Lemma VII.3.3, we see that the SKCE population criterion equals the ACMMD. However, the SKCE U-statistics estimator is different from the ACMMD U-statistics estimator: while the ACMMD U-statistics only requires samples the conditional distributions $Q|_X$, the SKCE U-statistics contains expectations over the conditional distributions $Q|_X$, which are rarely available in practice.

VII.D.2 Proof of Proposition VII.4.1

Proof. We will show that the image of $q \mapsto \mu_q$ is compact, and the result will follow from [39]. Let $\mathcal{M}(\mathcal{Y})$ the Banach space of measures of sequences endowed with the

total variation norm:

$$\|q\|_{\text{TV}} := q_+(\mathcal{Y}) + q_-(\mathcal{Y})$$

We recall that by the Riesz-Markov theorem, $(\mathcal{M}(\mathcal{Y}), \|\cdot\|_{\text{TV}})$ can be identified with the topological dual of $C_0(\mathcal{Y})$, $(C_0(\mathcal{S})^*, \|\cdot\|_{\text{op}})$ through an isometric isomorphism $q \in \mathcal{M}(\mathcal{Y}) \mapsto \tilde{q} \in C_0(\mathcal{Y})^*$, and for which the following holds:

$$\tilde{q}(f) = \int_{\mathcal{Y}} f d\tilde{q}, \quad \forall f \in C_0(\mathcal{Y}).$$

Let $B := \{q \in \mathcal{M}(\mathcal{Y}) \mid \|q\|_{\text{TV}} \leq 1\}$. As a unit ball, by the Banach-Alaoglu theorem, B is compact under the weak- \star topology and contains all distributions on sequences. We will show that $q \mapsto \mu_q$ is continuous on B and the result will follow. Given that this mapping is linear, it is sufficient to show continuity at 0. Moreover, since $\{B(0_{\mathcal{H}}, r)\}_{r>0}$ is a neighborhood basis of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, it suffices to show that there is a neighborhood \mathcal{V} of the null measure in the weak- \star topology such that $\|\int k_{\mathcal{Y}}(y, \cdot) dq(y)\|_{\mathcal{H}}^2 = \int k_{\mathcal{Y}}(y, y') d(q \otimes q)(y, y') < 1$ for all q in \mathcal{V} . Since the family

$$\{q \in \mathcal{M}(\mathcal{Y}), \int f_i(x) dq(x) < \varepsilon, i \in 1, \dots, k, f_i \in C_0(\mathcal{Y})\}$$

form a neighborhood basis of the weak- \star topology, we can consider candidates of this form for \mathcal{V} . In particular, let us set $\{f_i\} = \{x \mapsto \sqrt{k_{\mathcal{Y}}(x, x)} \in C_0(\mathcal{Y})\}$, since $k_{\mathcal{Y}} \in C_0(\mathcal{Y} \times \mathcal{Y})$, and $\varepsilon = 0.5$. On this neighborhood, we have:

$$\int k(y, y') dq(y) dq(y') \leq \int \sqrt{k(y, y) \times k(y', y')} dq(y) dq(y') \leq 0.5^2 < 1, \quad \forall q \in \mathcal{V}$$

showing the continuity of the map in question. As a consequence, the image of $B_{\mathcal{M}}(S)(0, 1)$ by the map $q \mapsto \mu_q$ is compact, implying from [39] that the kernel

$$\tilde{k}(f, g) := \exp\left(-\frac{1}{2\sigma^2} \|f - g\|_{\mathcal{H}}^2\right)$$

is universal on that set. Thus, we have shown that \tilde{k} is universal on \mathcal{H} under the strong topology (e.g. the norm topology in \mathcal{H}). This is equivalent to the TV topology of

$\mathcal{P}(S)$ since k has discrete masses by proposition 9 of [9], and thus $k_{\mathcal{P}(\mathcal{Y})}$ is universal on $\mathcal{P}(\mathcal{Y})$. \square

VII.D.3 Proofs regarding the impact of approximate kernels

To prove the convergence of the ACMMD–Rel estimator and the validity of its test, we rely on an augmented U-statistics formulation. Let:

$$U := (Q|_X, \tilde{Y}^1, \dots, \tilde{Y}^R, \tilde{Y}, Y) \sim \mathbb{P}_{Q|_X} \otimes \mathbb{Q}_|^{\otimes r} \otimes \mathbb{Q}_| \otimes \mathbb{P}_|^Q := \mathbb{U}$$

U is the random variable which, for each model $Q|_X$, concatenates the synthetic samples $(\tilde{Y}^1, \dots, \tilde{Y}^R)$ used to perform the kernel approximation $\hat{k}_{\mathcal{P}(\mathcal{Y})}(q, q')$, the synthetic sample \tilde{Y} used to evaluate h , and \tilde{Y} , a sample from $\mathbb{P}_|^Q$ the conditional distribution of Y given $Q|_X$. Then, given N realizations $\{U_i\}_{i=1}^N$ of U , the estimator $\widehat{\text{ACMMD–Rel}}^2$ can be written as a U-statistics on the $\{U_i\}_{i=1}^N$

$$\widehat{\text{ACMMD–Rel}}^2 = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h_a(U_i, U_j)$$

where

$$h_a(U_i, U_j) := \hat{k}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R) \times (k_{\mathcal{Y}}(Y_i, Y_j) + k_{\mathcal{Y}}(\tilde{Y}_i, \tilde{Y}_j) - k_{\mathcal{Y}}(Y_i, \tilde{Y}_j) - k_{\mathcal{Y}}(\tilde{Y}_i, Y_j)) \quad (\text{VII.15})$$

We also will note

$$\text{ACMMD}_a^2 := \mathbb{E}_{U_1, U_2 \sim \mathbb{U} \otimes \mathbb{U}} h_a(U_1, U_2)$$

VII.D.3.1 Proof of Proposition VII.4.3

With this formalism, we now prove that the ACMMD–Rel test has the specified type-I error rate of $\alpha \in (0, 1)$, e.g. rejects H_0 when $\mathbb{P}|_x = Q|_x$ with probability α . Indeed, straightforward adaptations of the arguments in Section VII.C.2 show that

Algorithm 13 ACMMMD–Rel Conditional Goodness of fit Test

Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_Y \otimes Q_{|Y}$

Parameters: Level α , kernel k_X , kernel k_Y

// Estimate ACMMMD–Rel using Algorithm 12 and collect the $\hat{h}(Z_i, Z_j)$ of Equation (VII.11)

$\widehat{\text{ACMMMD–Rel}}^2, \{h(Z_i, Z_j)\}_{1 \leq i < j \leq N} \leftarrow \text{estimate_acmmmd_rel}(\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N, Q_{|X_i})$

$[W_i^b \sim \text{Rademacher} \text{ for } i \in 1, \dots, N \text{ for } b \in 1, \dots, B]$

$[\widetilde{\text{ACMMMD}}_b^2 \leftarrow \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} W_i^b W_j^b \hat{h}(Z_i, Z_j), \text{ for } b \text{ in } 1, \dots, B]$

// See Section VII.C.1 for how to compute $\hat{q}_{1-\alpha}$

$\hat{q}_{1-\alpha} \leftarrow \text{approx. } (1 - \alpha)\text{-quantile of } \{\widetilde{\text{ACMMMD}}_b^2\}_{b=1}^B$

if $\widetilde{\text{ACMMMD}}^2 \leq \hat{q}_{1-\alpha}$ **then**

Fail to reject H_0

else

Reject H_0

end if

that doing a wild bootstrap using the $\hat{h}(Z_i, Z_j)$ is equivalent to estimating

$$\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} h_a(U_i^\sigma, U_j^\sigma) := \widehat{\text{ACMMMD–Rel}}_{\sigma_b}^2$$

where $U_i^\sigma := (Q_{|X_i}, \tilde{Y}_i^1, \dots, \tilde{Y}_i^R, \tilde{Y}_i^\sigma, Y_i^\sigma)$, where $(Y_i^\sigma, \tilde{Y}_i^\sigma) = (Y_i, \tilde{Y}_i)$ if $\sigma(i) = 1$ and (\tilde{Y}_i, Y_i) otherwise. The same argument to show that ACMMMD test has the desired type-I error rate follows in this case too: Under $\mathbb{P}_{|x} = Q_{|x}$, the sequence $\{\widehat{\text{ACMMMD–Rel}}_{\sigma_b}^2\}_{b=0}^B$ is exchangeable (noting $\widehat{\text{ACMMMD–Rel}}_{\sigma_0}^2 = \widehat{\text{ACMMMD–Rel}}^2$, e.g. $\sigma_0(i) = 1$ for all $1 \leq i \leq N$), and we can repeat the derivations of the proof of Lemma VII.C.2 to show that the ACMMMD–Rel test has the desired type-I error rate.

VII.D.3.2 Proof of Proposition VII.4.2

We prove a slightly more general version of the proposition, for kernels of the form $\phi(d(q, q')^2)$, where ϕ is a Lipschitz function and d is a distance on $\mathcal{P}(\mathcal{Y})$. Setting $\phi = e^{-\frac{\cdot}{\sigma^2}}$, we recover the kernels of Proposition VII.4.2, which include the exponentiated MMD kernel.

Proposition VII.D.1. *Assume that k_Y and $k_{\mathcal{P}(\mathcal{Y})}$ is a kernel of the form $k_{\mathcal{P}(\mathcal{Y})}(q, q') = \phi(d(q, q'))$, for a Lipschitz function ϕ and a function $d(q, q')$ admitting an unbiased*

estimator of the form $\widehat{d}(\{y_1^r\}_{r=1}^R, \{y_2^r\}_{r=1}^R)$ where $\{y_1^r\}_{r=1}^R$ and $\{y_2^r\}_{r=1}^R$ are R i.i.d samples of q and q' respectively, with variance $O(\frac{1}{R})$ (the bound in uniform in q and q'). Then, assuming $R \equiv R(N)$, with $\lim_{N \rightarrow \infty} R(N) = +\infty$, $\widehat{\text{ACMMD}}\text{--Rel}^2$ converges in probability to $\text{ACMMD}\text{--Rel}^2$ as $N \rightarrow \infty$.

Proof. As discussed above, the estimator $\widehat{\text{ACMMD}}\text{--Rel}^2$ can be written as a U-statistics on the $\{U_i\}_{i=1}^N$, where $U_i = (Q|_{X_i}, \tilde{Y}_i^1, \dots, \tilde{Y}_i^R, \tilde{Y}_i, Y_i)$, and using the kernel h_a defined as (accounting approximating $k_{\mathcal{Y}}$ through d directly)

$$\begin{aligned} h_a(U_i, U_j) &:= \phi(\widehat{d}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R)) \\ &\quad \times (k_{\mathcal{Y}}(Y_i, Y_j) + k_{\mathcal{Y}}(\tilde{Y}_i, \tilde{Y}_j) - k_{\mathcal{Y}}(Y_i, \tilde{Y}_j) - k_{\mathcal{Y}}(\tilde{Y}_i, Y_j)) \end{aligned} \quad (\text{VII.16})$$

e.g.

$$\widehat{\text{ACMMD}}\text{--Rel}^2 = \frac{2}{N(N-1)} \sum_{i < j} h_a(U_i, U_j)$$

To study the convergence in probability of $\widehat{\text{ACMMD}}_a^2$ to ACMMD_a^2 , we use finite-sample bounds on U -statisticcs Hoeffding [110]:

$$P \left(|\widehat{\text{ACMMD}}_a^2 - \text{ACMMD}_a^2| > \|h_a\|_{\infty} \sqrt{\frac{\log(2/\delta)}{2[N/2]}} \right) \leq \delta$$

for all $\delta > 0$, where, by assumption, $k_{\mathcal{Y}}$ bounded, and $\widehat{k}_{\mathcal{P}(\mathcal{Y})}$ is of the form $\phi(\widehat{d}(\{y_1^r\}_{r=1}^R, \{y_2^r\}_{r=1}^R))$ for some bounded function ϕ , implying that h_a is bounded.

To show the dependence in R , we bound the difference ACMMD_a^2 and ACMMD^2 .

$$\begin{aligned}
& |\text{ACMMD}_a^2 - \text{ACMMD}^2| \\
&= \mathbb{E}_{\mathbb{U}, \mathbb{U}} \left[(\widehat{k}_{\mathcal{P}(\mathcal{Y})}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_1}, Q_{X_2})) \times (k_{\mathcal{Y}}(Y_1, Y_2) \right. \\
&\quad \left. + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2)) \right] \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty |\mathbb{E}_{\mathbb{P}_{Q_1} \otimes \mathbb{Q}^{\otimes r} \times \mathbb{P}_{Q_2} \otimes \mathbb{Q}^{\otimes r}} \widehat{k}_{\mathcal{P}(\mathcal{Y})}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_1}, Q_{X_2})| \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \|\phi\|_{\text{Lip}} \mathbb{E}_{\mathbb{P}_{Q_1} \otimes \mathbb{Q}^{\otimes r} \times \mathbb{P}_{Q_2} \otimes \mathbb{Q}^{\otimes r}} |\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - d(Q_{|X_1}, Q_{X_2})| \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \|\phi\|_{\text{Lip}} \times \\
&\quad \mathbb{E}_{\mathbb{P}_{Q_1} \times \mathbb{P}_{Q_2}} \left[\mathbb{E}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} |\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - d(Q_{|X_1}, Q_{X_2})| \mid Q_{|X_1}, Q_{X_2} \right] \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \|\phi\|_{\text{Lip}} \mathbb{E}_{\mathbb{P}_{Q_1} \times \mathbb{P}_{Q_2}} \left[\sqrt{\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} \widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R)} \mid Q_{|X_1}, Q_{X_2} \right]
\end{aligned}$$

Where the last inequality follows from Jensen's inequality and the unbiasedness of \widehat{d} . The result follows by applying the assumption on the variance of \widehat{d} (a bound which we assume is uniform in $Q_{|X_1}, Q_{|X_2}$). \square

The term $\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} [\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) | Q_{|X_1}, Q_{|X_2}]$ can be more precisely characterized depending on \widehat{d} . For instance, we have, when \widehat{d} is a U-statistics (for instance, using the MMD estimator of Gretton et al. [90, Lemma 6, Equation 4]), that [221, section 5.2.1] $\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} \widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) < \zeta(Q_1, Q_{|X_2})/R$, where $\zeta(Q_{|X_1}, Q_{|X_2}) := \mathbb{V}_{(Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2) \sim Q_{|X_1} \otimes Q_{|X_2}} (\tilde{h}((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)))$ and $\tilde{h}((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) = k_{\mathcal{Y}}(Y_1, Y_2) + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2)$, which is uniformly bounded by $4 \|k_{\mathcal{Y}}\|_\infty$ for bounded kernels. Putting the two parts together, we thus have that:

$$P \left(\left\{ \widehat{\text{ACMMD}}_a^2 - \text{ACMMD}^2 \right\} > 4 \|k_{\mathcal{Y}}\|_\infty \sqrt{\frac{\log(2/\delta)}{2[N/2]}} + \frac{16 \|k_{\mathcal{Y}}\|_\infty^2 \|\phi\|_{\text{Lip}}}{\sqrt{R}} \right) \leq \delta$$

for all $\delta > 0$, showing the convergence in probability of $\widehat{\text{ACMMD}}_a$ to ACMMD .

VII.D.4 Additional Details for ACMMMD and ACMMMD–Rel in the synthetic example

VII.D.4.1 Derivations of ACMMMD in the synthetic example

We first prove that ACMMMD is proportional to Δp .

Lemma VII.D.1. *In the setting described in Section VII.6.1, we have*

$$\text{ACMMMD}^2(\mathbb{P}_|, \mathbb{Q}_|) = C \times \Delta p^2$$

for

$$C := \iint k_{\mathcal{X}}(p, p') 2(1 - e^{-\lambda}) \frac{(1 - 2p)(1 - 2p')}{1 - 2pp'(1 + e^{-\lambda})} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right) d\mathbb{P}_X(p) d\mathbb{P}_X(p') \quad (\text{VII.17})$$

Proof. Recall that we have

$$\begin{aligned} \text{ACMMMD}^2 &= \text{MMD}^2(\mathbb{P}_X \otimes \mathbb{P}_|, \mathbb{P}_X \otimes \mathbb{Q}_|)^2 \\ &= \int k_{\mathcal{X}}(p, p') (T_{11} + T_{22} - 2T_{12}) d\mathbb{P}_X(p) d\mathbb{P}_X(p') \end{aligned}$$

where

$$T_{12} = \int k_{\mathcal{Y}}(y, y') p(y|p) q(y'|p') d(y) d(y')$$

and T_{22} and T_{11} are defined similarly. For a sequence y , we define the function len given by $\text{len}(y) := \min \{i \in \mathbb{N} | y_i = \text{STOP}\}$, which intuitively returns the length of the sequence.

Computing T_{ij} As we will see, a lot of the computations are agnostic to whether we are computing T_{11} , T_{22} or T_{12} . Note that the exponentiated hamming distance kernel on \mathcal{Y} writes as a product

$$k_{\mathcal{Y}}(y, y') = e^{-\lambda d_H(y, y')} = e^{-\lambda_y \sum_{i=0}^{\infty} \delta(y_i \neq y'_i)} = \prod_{i=0}^{\infty} e^{-\lambda \delta(y_i \neq y'_i)} = \prod_{i=0}^{\max(\text{len}(y), \text{len}(y'))} e^{-\lambda \delta(y_i \neq y'_i)}$$

let us define the following events

$$\begin{aligned} F(m) &:= \left\{ \min(\text{len}(y), \text{len}(y')) = m \right\} \\ G(m, \delta m) &:= \left\{ \max(\text{len}(y), \text{len}(y')) = m + \delta m \right\} \end{aligned}$$

which we further break down as

$$\begin{aligned} F_1(m) &= \{\text{len}(y) = m\} \cap \{\text{len}(y') > m\} \\ F_2(m) &= \{\text{len}(y) > m\} \cap \{\text{len}(y') = m\} \\ F_3(m) &= \{\text{len}(y) = m\} \cap \{\text{len}(y') = m\} \\ \implies F(m) &= F_1(m) \cup F_2(m) \cup F_3(m) \end{aligned}$$

For which the following probabilities hold:

$$\begin{aligned} P(F_1(m)) &= P(\text{len}(y) = m) \times P(\text{len}(y') > m) \\ &= ((2p)^m(1 - 2p))(2p')^{m+1} \\ P(F_2(m)) &= P(\text{len}(y') = m) \times P(\text{len}(y) > m) \\ &= ((2p')^m(1 - 2p'))(2p)^{m+1} \\ P(F_3(m)) &= P(\text{len}(y') = m) \times P(\text{len}(y) = m) \\ &= ((2p')^m \times (1 - 2p')) \times ((2p)^m \times (1 - 2p)) \\ P(G(m, \delta m)|F_1(m)) &= P(\text{len}(y') = m + \delta m | \text{len}(y) = m, \text{len}(y') > m) \\ &= (2p')^{\delta m - 1} \times (1 - 2p')\delta_{(\delta m \geq 1)} \\ P(G(m, \delta m)|F_2(m)) &= P(\text{len}(y) = m + \delta m | \text{len}(y') = m, \text{len}(y) > m) \\ &= (2p)^{\delta m - 1} \times (1 - 2p)\delta_{(\delta m \geq 1)} \\ P(G(m, \delta m)|F_3(m)) &= \delta(\delta m = 0) \end{aligned}$$

Let us note

$$E(m, \delta m, i) := F_i(m) \cap G(m, \delta m)$$

We have that $E(m, \delta_m, i) \cap E(m', \delta_{m'}, j) = \emptyset$ if $(m, \delta_m, i) \neq (m', \delta_{m'}, j)$.

$$\Omega = \bigcup_{m=0}^{+\infty} \bigcup_{i=1}^3 \bigcup_{\delta_m=0}^{+\infty} E(m, \delta_m, i)$$

Using the law of total probability, we have that Thus, using the law of total probability:

$$\begin{aligned} T_{ij}(p, p') &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \sum_{\delta_m=0}^{+\infty} \mathbb{P}(E(m, \delta_m, i)) \mathbb{E}(e^{-\lambda d_H(y, y')} | E(m, \delta_m, i)) \\ &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \sum_{\delta_m=0}^{+\infty} \mathbb{P}(F_i(m) \cap G(m, \delta_m)) \mathbb{E}(e^{-\lambda d_H(y, y')} | E(m, \delta_m, i)) \\ &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \mathbb{P}(F_i(m)) \sum_{\delta_m=0}^{+\infty} P(G(m, \delta_m) | F_i(m)) \mathbb{E}(e^{-\lambda d_H(y, y')} | E(m, \delta_m, i)) \\ &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \mathbb{P}(F_i(m)) \mathbb{E}(e^{-\lambda d_H(y_{:m}, y'_{:m})} | F_i(m)) \\ &\quad \times \sum_{\delta_m=0}^{+\infty} P(G(m, \delta_m) | F_i(m)) \mathbb{E}(e^{-\lambda d_H(y_{m+1:m+\max(\delta_m, 1)}, y'_{m+1:m+\max(\delta_m, 1)})} | E(m, \delta_m, i), p, p') \\ &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \mathbb{P}(F_i(m)) \mathbb{E}(e^{-\lambda d_H(y_{:m}, y'_{:m})} | F_i(m)) \\ &\quad \times \sum_{\delta_m=0}^{+\infty} P(G(m, \delta_m) | F_i(m)) e^{-\lambda(\max(0, \delta_m - 1) + \delta(m > 0))} \\ &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \mathbb{P}(F_i(m)) \mathbb{E}(e^{-\lambda d_H(y_{:m}, y'_{:m})} | F_i(m)) \left(\prod_{i=1}^{\max(m-1, 1)} \mathbb{E}(e^{-\lambda \delta(y_i \neq y'_i)} | F_i(m)) \right)^{\delta(m \geq 2)} \\ &\quad \times \sum_{\delta_m=0}^{+\infty} P(G(m, \delta_m) | F_i(m)) e^{-\lambda(\max(0, \delta_m - 1) + \delta(m > 0))} \end{aligned}$$

where we break down the factorized hamming distance over the sequence into the sum of the hamming distances over each coordinate, and made use of the fact that

$$d_H(y_{m:m+\delta_m}, y'_{m:m+\delta_m}) = \max(0, \delta_m - 1) + \delta(m > 0)$$

conditioned on $F_i(m)$ and $G(m, \delta m)$. The disjunction of cases is necessary in order to not count the term 0^{th} term twice in the event when $m = 0$. This representation is convenient since whenever $m \geq 2$, for any $1 \leq i \leq m - 1$,

$$P(\delta(y_i, y'_i) = 1 | F_i(m)) = \frac{(pp') + (pp')}{(p+p) \times (p'+p')} = \frac{1}{2} = P(\delta(y_i, y'_i) = 0 | F_i(m))$$

meaning we have

$$\begin{aligned} T_{ij}(p, p') &= \sum_{m=0}^{+\infty} \sum_{i=1}^3 \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_i(m)) \mathbb{P}(F_i(m)) \left(\frac{1 + e^{-\lambda}}{2} \right)^{\max(m-1, 0)} \\ &\quad \times \sum_{\delta m=0}^{+\infty} P(G(m, \delta m) | F_i(m)) e^{-\lambda(\max(0, \delta m - 1) + \delta(m > 0))} \end{aligned}$$

Inserting the relevant event probabilities into the expression for T_{ij} , we have

$$\begin{aligned}
T_{ij}(p, p') &= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2} \right)^{\max(m-1,0)} \times \left(\right. \\
&\quad \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_1(m)) (2p)^m \times (1-2p) (2p')^{m+1} \times (1-2p') e^{-\lambda \delta(m>0)} \\
&\quad \times \sum_{\delta m=1}^{+\infty} e^{-\lambda(\delta m-1)} (2p')^{\delta m-1} \\
&\quad + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_2(m)) (2p')^m \times (1-2p') (2p)^{m+1} \times (1-2p) e^{-\lambda \delta(m>0)} \\
&\quad \times \sum_{\delta m=1}^{+\infty} e^{-\lambda(\delta m-1)} (2p)^{\delta m-1} \\
&\quad \left. + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_3(m)) (2p')^m \times (1-2p') (2p)^m (1-2p) \right) \\
&= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2} \right)^{\max(m-1,0)} \times \left(\right. \\
&\quad \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_1(m)) (2p)^m \times (1-2p) (2p')^{m+1} \times (1-2p') e^{-\lambda \delta(m>0)} \\
&\quad \sum_{\delta m=0}^{+\infty} e^{-\lambda \delta m} (2p')^{\delta m} \\
&\quad + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_2(m)) (2p')^m \times (1-2p') (2p)^{m+1} \times (1-2p) e^{-\lambda \delta(m>0)} \\
&\quad \sum_{\delta m=0}^{+\infty} e^{-\lambda \delta m} (2p)^{\delta m} \\
&\quad \left. + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_3(m)) (2p')^m \times (1-2p') (2p)^m (1-2p) \right) \\
&= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2} \right)^{\max(m-1,0)} \times \left(\right. \\
&\quad \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_1(m)) (2p)^m \times (1-2p) (2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda \delta(m>0)}}{1-2p' e^{-\lambda}} \\
&\quad + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_2(m)) (2p')^m \times (1-2p') (2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda \delta(m>0)}}{1-2p e^{-\lambda}} \\
&\quad \left. + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_3(m)) (2p')^m \times (1-2p') (2p)^m (1-2p) \right)
\end{aligned}$$

Now, some simplifications arise when $m \geq 1$. Indeed, in that case, $\mathbb{E}(e^{-\lambda \delta(y_0, y'_0)} | F_i(m))$ is independent of i . Noting $T_{ij}^1(p, p')$ the sum of the terms for $m \geq 1$, we thus have

$$\begin{aligned} T_{ij}^1(p, p') &= \sum_{m=1}^{+\infty} \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F(m)) \left(\frac{1+e^{-\lambda}}{2} \right)^{m-1} \\ &\quad \times \left((2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda}}{1-2p'e^{-\lambda}} \right. \\ &\quad + (2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda}}{1-2pe^{-\lambda}} \\ &\quad \left. + (2p')^m \times (1-2p')(2p)^m (1-2p) \right) \end{aligned}$$

Noting A_{ij} the term $\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F(m))$, which is constant for all $m \geq 1$

$$\begin{aligned} T_{ij}^1(p, p') &= A_{ij} \sum_{m=1}^{+\infty} \left(\frac{1+e^{-\lambda}}{2} \right)^{m-1} \times \left((2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda}}{1-2p'e^{-\lambda}} \right. \\ &\quad + (2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda}}{1-2pe^{-\lambda}} \\ &\quad \left. + (2p')^m \times (1-2p')(2p)^m (1-2p) \right) \\ &= A_{ij} (1-2p)(1-2p') 4pp' \left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1 \right) \\ &\quad \times \sum_{m=0}^{+\infty} (4pp'(1+e^{-\lambda})/2)^m \\ &= A_{ij} \times \frac{(1-2p)(1-2p') 4pp'}{1-4pp'(1+e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1 \right) \\ &= C \times A_{ij} \end{aligned}$$

where

$$C(p, p') = \frac{(1-2p)(1-2p') 4pp'}{1-2pp'(1+e^{-\lambda})} \left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1 \right)$$

is a constant that does not depend on i, j . We compute the $m = 0$ sum, noted $T_{ij}^0(p, p')$.

We have

$$\begin{aligned} T_{ij}^0(p, p') &= \left(\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_1(0)) \times (1 - 2p)(2p') \times (1 - 2p') \times \frac{1}{1 - 2p'e^{-\lambda}} \right. \\ &\quad + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_2(0)) \times (1 - 2p')(2p) \times (1 - 2p) \times \frac{1}{1 - 2pe^{-\lambda}} \\ &\quad \left. + \mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_3(0)) \times (1 - 2p')(1 - 2p) \right) \end{aligned}$$

And we need to compute the terms $\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_i(0))$ individually.

$i = 1, i = 2$ For $i = 1$, we must have $y_0 \neq y'_0$, since $y_0 = \text{STOP}$, and $\text{len}(y') > 0$. Thus, $\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_1(0)) = e^{-\lambda}$. Similarly, $\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_2(0)) = e^{-\lambda}$.

$i = 3$ In that case, we must have $y_0 = y'_0 = \text{STOP}$, since $\text{len}(y) = \text{len}(y') = 0$. Thus, $\mathbb{E}(e^{-\lambda \delta(y_0 \neq y'_0)} | F_3(0)) = 1$.

Putting this together, we have

$$T_{ij}^0(p, p') = (1 - 2p)(1 - 2p') \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right)$$

With that notation, we have:

$$\begin{aligned} \text{ACMMD}^2(\mathbb{P}_{|}, \mathbb{Q}_{|}) &= \int k_{\mathcal{X}}(p, p') C(p, p') (A_{11} + A_{22} - 2A_{12}) d\mathbb{P}_X(p) d\mathbb{P}_X(p') \\ &\quad + \int k_{\mathcal{X}}(p, p') (T_{11}^0(p, p') + T_{22}^0(p, p') - 2T_{12}^0(p, p')) d\mathbb{P}_X(p) d\mathbb{P}_X(p') \\ &= \int k_{\mathcal{X}}(p, p') C(p, p') (A_{11} + A_{22} - 2A_{12}) d\mathbb{P}_X(p) d\mathbb{P}_X(p') \end{aligned}$$

since T_{ij}^0 does not depend on i, j . We can narrow the variation down even further:

by noting $p_{ij}^A = P(\delta(y_i \neq y'_i) = 0 | F(m))$ (resp $p_{ij}^B = P(\delta(y_i \neq y'_i) = 0 | F(0))$), since $\mathbb{E}(e^{-\lambda \varepsilon}) = p(\varepsilon = 0)(1 - e^{-\lambda}) + e^{-\lambda}$ if ε is a Bernoulli random variable,

$$\begin{aligned} \text{ACMMD}^2(\mathbb{P}_{|}, \mathbb{Q}_{|}) &\int k_{\mathcal{X}}(p, p') C(p, p') (1 - e^{-\lambda})(p_{11}^A + p_{22}^A - 2p_{12}^A) \\ &\quad d\mathbb{P}_X(p) d\mathbb{P}_X(p') \quad (\text{VII.18}) \end{aligned}$$

We now compute the probabilities p_{ij}^A for $i, j \in \{1, 2\}$. In every case, such p_{ij}^A can be written as:

$$p_{ij}^A = \frac{P(y_0 = y'_0 = A) + P(y_0 = y'_0 = B)}{P(\{y_0 \in \{A, B\}\} \cap \{y'_0 \in \{A, B\}\})} = \frac{P(y_0 = y'_0 = A) + P(y_0 = y'_0 = B)}{4pp'}$$

and we have

$$\begin{aligned} p_{11}^A &= \frac{pp' + pp'}{4pp'} = \frac{1}{2} \\ p_{22}^A &= \frac{(p + \Delta p)(p' + \Delta p) + (p - \Delta p)(p' - \Delta p)}{4pp'} = \frac{2pp' + 2\Delta p^2}{4pp'} \\ p_{12}^A &= \frac{(p)(p' + \Delta p) + (p)(p' - \Delta p)}{4pp'} = \frac{1}{2} \\ \implies p_{11}^A + p_{22}^A - 2p_{12}^A &= \frac{2pp' + 2\Delta p^2}{4pp'} - \frac{1}{2} = \frac{2\Delta p^2}{4pp'} \end{aligned}$$

Putting it together We thus have

$$\text{ACMMD}(\mathbb{P}_|, \mathbb{Q}_|) = \int C(p, p') k(p, p') (1 - e^{-\lambda}) \frac{2\Delta p^2}{4pp'} d\mathbb{P}_X(p) d\mathbb{P}_X(p')$$

Recalling that

$$C(p, p') = \frac{(1 - 2p)(1 - 2p')4pp'}{1 - 2pp'(1 + e^{-\lambda})} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right)$$

yields the desired result. \square

VII.D.4.2 Closed-form ACMMD–Rel evaluation

Assuming the same model, it is also possible to evaluate $\text{ACMMD–Rel}(\mathbb{P}_|, \mathbb{Q}_|)$ in closed form. Indeed, ACMMD–Rel becomes a special case of the ACMMD formula given above, with the conditioned variable X set to be the models $\mathbb{Q}_{|X}$. It is thus possible to show:

Lemma VII.D.2. *We have*

$$\text{ACMMD–Rel}^2(\mathbb{P}_|, \mathbb{Q}_|) = C \times \Delta p^2$$

for

$$C = \iint k_{\mathcal{P}(\mathcal{Y})}(q|_p, q|_{p'}) 2(1 - e^{-\lambda}) \frac{(1 - 2p)(1 - 2p')}{1 - 2pp'(1 + e^{-\lambda})} \\ \times \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right) d\mathbb{P}_X(p)d\mathbb{P}_X(p') \quad (\text{VII.19})$$

The above lemma leaves the choice of the kernel $k_{\mathcal{P}(\mathcal{Y})}$ open: the tractability of this expression will follow only if such kernel can be tractably computed. In the next lemma, we derive a closed form solution for $k_{\mathcal{P}(\mathcal{Y})}(q, q')$ when $k_{\mathcal{P}(\mathcal{Y})}(q, q') = e^{-\frac{\text{MMD}^2(q, q')}{2\sigma^2}}$, where the MMD is computed with an Exponentiated Hamming kernel on \mathcal{Y} .

Lemma VII.D.3. *We have*

$$\text{MMD}^2(q|_p, q|_{p'}) = T(p, p) + T(p', p') - 2T(p, p')$$

Where

$$T(p, p') = C(p, p')A(p, p') + T^0(p, p') \\ C(p, p') = \frac{(1 - 2p)(1 - 2p')4pp'}{1 - 2pp'(1 + e^{-\lambda})} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right) \\ A(p, p') = \frac{2pp' + 2\Delta p^2}{4pp'} \times (1 - e^{-\lambda}) + e^{-\lambda} \\ T^0(p, p') = (1 - 2p)(1 - 2p') \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right)$$

Combining the two lemmas allows us to obtain a computable expression for ACMMD–Rel($\mathbb{P}|, Q|$).

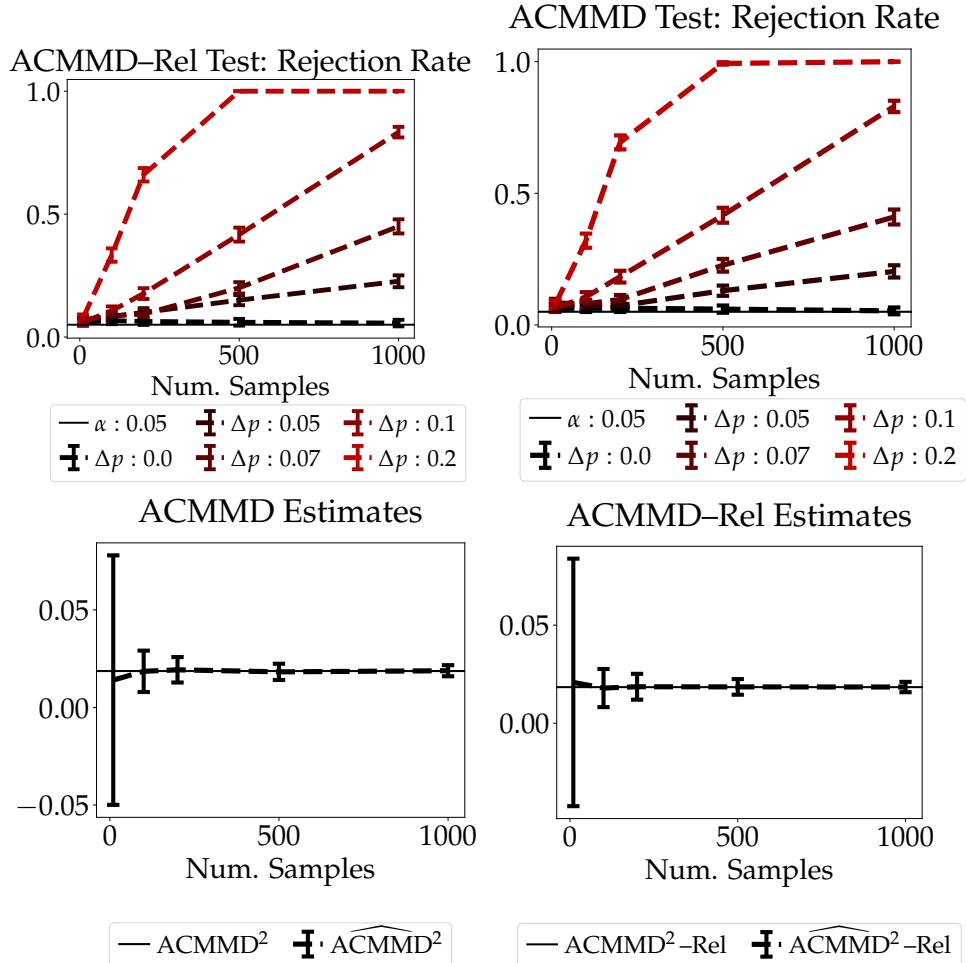


Figure VII.D.1: Top left: Rejection Rate of the ACMMD test as a function of the dataset size, for different values of Δp . Top right: Rejection Rate of the ACMMD test as a function of Δp , for different dataset sizes. Bottom left: Estimated ACMMD as a function of the dataset size. Bottom right: Estimated ACMMD–Rel as a function of Δp . To compute these estimates, we use dataset sizes of $\{10, 100, 200, 500, 1000\}$, used $m = 5$ atoms for the prior on p between $p_1 = 0.3$, $p_2 = 0.45$, used $\lambda = 1$, $\Delta p = 0.25$, and average over 300 runs. In addition, we plot the true value $\text{ACMMD}(\mathbb{P}_1, \mathbb{Q}_1)$ using the closed-form expression derived above.

VII.E Additional Experiments

VII.E.1 Additional Experiments for the semisynthetic ProteinMPNN data

In addition to the figures of Section VII.6.2.1, which use $T = 0.1$ to plot the estimates and rejection rates of ACMMMD and ACMMMD–Rel on the ProteinMPNN synthetic data, we provide here the same plots for $T = 1.0$ the value used to train ProteinMPNN. We notice that detecting a given change in temperature is slightly simpler for $T = 1.0$ than for $T = 0.1$.

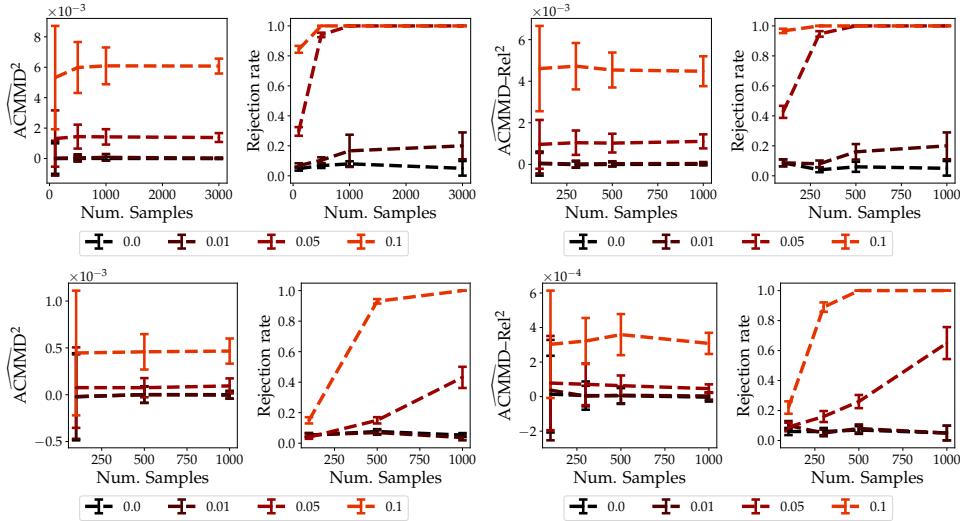


Figure VII.E.1: ACMMMD and ACMMMD–Rel estimates and rejection rate in the semisynthetic setting of Section VII.6.2.1. The different lines indicate a different temperature shift between the two MPNN models. Top panel shows uses a base temperature of $T = 1.0$, while the bottom panel uses $T = 0.1$.

VII.E.2 Additional Experiments for the structural superfamily evaluation

We include in Figure VII.E.2 the values of $\widehat{\text{ACMMD-Rel}}^2$ for different superfamilies (which was not included in Section VII.6.2.2), and compare it with the values of $\widehat{\text{ACMMD}}^2$. In line with the hyperparameter tuning results of Section VII.6.2.2, we notice that high temperature are highly detrimental from a reliability perspective. Intuitively, increasing the temperature of ProteinMPNN makes the model “underconfident”. Since a reliable model is neither over- nor underconfident, this decrease

of confidence is penalized by ACMMMD-Rel. This also shows that increasing the temperature of a model does not make the model fallback to its prior distribution (otherwise the model would be more reliable). Instead, it just increases the uncertainty of the model in a detrimental fashion.

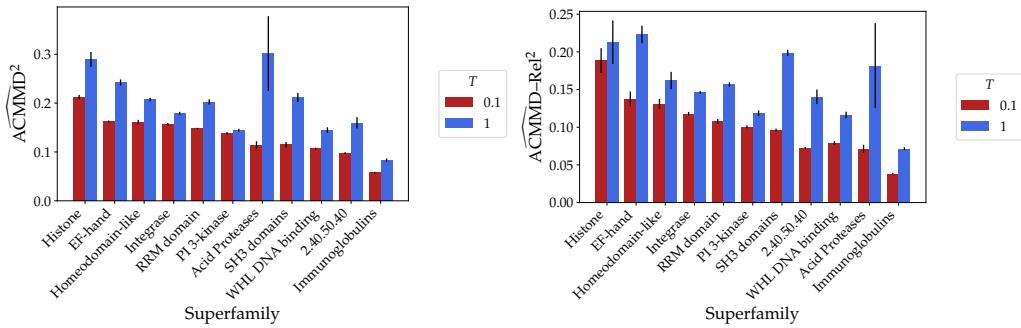


Figure VII.E.2: Value of $\widehat{\text{ACMMD}}^2$ (left) and $\widehat{\text{ACMMD-Rel}}^2$ (right) between Protein-MPNN and the CATH S60 reference dataset on a subset of 10 superfamilies for two different temperatures $T = 1.0$ and $T = 0.1$.

VII.F Known Kernels for protein sequences and structures

In the context of inverse folding, computing the ACMMMD requires a kernel on sequences k_y and a kernel on protein structures. This section contains a brief overview of non neural-network based, known kernels for protein sequences and structures. The main desiderata to achieve when choosing kernels for computing goodness-of-fit criterion is to find kernels that are able to detect (up to statistical noise) any deviation from a perfect fit between the model and the data. Such kernels are referred to as *universal*.

The protein formalism The most general formalism for the space protein structures is the set of equivalence classes of graphs where the equivalence relationship is defined to the existence of a graph isomorphism. The need for equivalence classes stems from the fact that different labelling policies exist for a given protein, meaning that a single protein can be associated to multiple graphs. However, this labelling will in practice not be completely arbitrary: first, the set of candidate labelling can be restricted to the ones consistent with covalent bounds. But in the inverse

folding problem, the setting is even simpler: the protein structure is restricted to its backbone, which is sequential by nature. This limits the set of covalent-bound consistent labelling policies to two (the forward and the backward one), and my vague understanding is that there is a terminal atom in protein, which suggests the existence of a canonical direction: thus, only one labelling policy remain, and protein structures can thus be associated to the set of atom locations $\bigcup_{i=1}^{+\infty} \mathbb{R}^i$. This set differs from the set of protein sequences $\bigcup_{i=1}^{+\infty} \mathcal{A}$ in that the “alphabet” is the real line instead of a finite set of symbols. The restriction from the space of graphs to the space of variable-length sequences since there it is known that no graph kernels commonly in use are even characteristic [138]. The space $\bigcup_{i=1}^{+\infty} \mathcal{X}$ (for arbitrary \mathcal{X} have been investigated by the time series community), which have developed a set of kernels to carry out data analysis on it. We provide some background on such kernels below.

Background: alignment kernels for real-valued sequences of arbitrary length

Alignment kernels [49, 48, 214, 260] refer to a diverse set of variety of kernels constitute a family of kernels on $\bigcup_{i=1}^{+\infty} \mathcal{X}^i$ that are computed based on aggregating the similarities between all possible “alignment candidates” between two inputs x_1 and x_2 . There are two main subfamilies of alignment kernels, which both use slightly different alignment definitions: local alignment kernels, and global alignment kernels.

Local alignment kernels Local alignment kernels [214, 260] are kernels of the form

$$k_{\text{LA}}(x, y) = \sum_{\pi \in \Pi(x, y)} \exp(\beta s(x, y, \pi)) \quad (\text{VII.20})$$

Where

$$s(x, y, \pi) = \sum_{i=1}^{|\pi|} s(x_1^{(\pi_1(i))}, x_2^{(\pi_2(i))}) + \sum_{i=1}^{|\pi|-1} g(\pi_1(i+1) - \pi_1(i)) + g(\pi_2(i+1) - \pi_2(i))$$

and $\Pi(x, y)$ is the set of all possible *alignments* of x and y , e.g. the set of all 2-tuple of p -long sequences

$$\pi := ((\pi_1(1), \dots, \pi_1(p)), (\pi_2(1), \dots, \pi_2(p)))$$

where

$$1 \leq \pi_1(1) < \pi_1(2) \cdots < \pi_1(p) \leq n$$

$$1 \leq \pi_2(1) < \pi_2(2) \cdots < \pi_2(p) \leq m$$

Importantly, local alignment kernels involve a gap function, and thus give a specific status to insertions and deletions, unlike global alignment kernels, as we will see below. The local alignment kernel can be seen as computing the (soft) minimum of a discrepancy within the set of all possible alignments. The use of a soft minimum (and not a hard one) is crucial to ensure positive definiteness. Local alignment kernels seem to have been designed initially for finite alphabets target. When $g = 0$, the necessary and sufficient condition on s to ensure that k_{LA} is a positive definite is for s to be a conditionally positive definite kernel⁴. This is in particular verified if $(s(x^i, y^j))_{1 \leq i, j \leq |\mathcal{A}|}$ is positive definite. I need further reading to investigate whether the case of infinite \mathcal{X} was studied.

Global alignment kernels Global alignment kernels [49, 48] also perform a softmin over alignment, but do not incorporate gaps in their score, and use a slightly different notion of alignment, namely:

$$\boldsymbol{\pi} := ((\pi_1(1), \pi_2(1), \dots, (\pi_1(p), \pi_2(p)))$$

where now, the constraints on π_1 and π_2 are

$$1 = \pi_1(1) < \pi_1(2) \cdots < \pi_1(p) = n$$

$$1 = \pi_2(1) < \pi_2(2) \cdots < \pi_2(p) = m$$

$$\pi_1(i+1) \leq \pi_1(i) + 1 \quad \text{unitary increments}$$

$$(\pi_1(i+1) - \pi_1(i)) + (\pi_2(i+1) - \pi_2(i)) \geq 1 \quad \text{no repetitions}$$

Unlike the previous alignment definition, this one explicitly maps each item in each sequence with another item in the other sequence, and does not try to account for potential gaps. Let us call \mathcal{A} the set of all alignment. The final definition for a global alignment kernel is then:

⁴A kernel is c.p.d if $\sum_{i,j=1}^n c_i c_j s(x^{(i)}, x^{(j)}) \geq 0 \forall c_1, \dots, c_n, c_1 + \dots + c_n = 0$.

$$k_{\text{GA}}(x, y) = \sum_{\pi \in \mathcal{A}(x, y)} \exp\left(\sum_{i=1}^{\pi} s(x_1^{(\pi_1(i))}, x_2^{(\pi_2(i))})\right) \quad (\text{VII.21})$$

As stated in Cuturi et al. [49, Theorem 1], k_{GA} will be positive definite if $k(x, y) := \exp(s(x, y))$ is a positive definite kernel such that $\frac{k}{(1-k)}$ is positive definite.

CHAPTER VIII

Measuring data and model properties with `kdiscs`

This Chapter is based on the following work:

Pierre Glaser, Antonin Schrab, and Arthur Gretton. Measuring data and model properties with `kdiscs`. In preparation, 2025

Abstract

The last two decades have seen the emergence of a versatile class of methods aimed at measuring distributional properties of data and statistical models, relying on kernel methods as their analytical backbone. However, the software ecosystem for these methods is currently fragmented, with no library providing a comprehensive and user-friendly implementation of these methods. To address this issue, we introduce `kdiscs`, a Python package for measuring data and model properties with kernels. `kdiscs` implements estimators (and accompanying tests) of kernel-based measures of most well-known distributional properties, including equality in distribution, independence, (conditional) goodness-of-fit and calibration. `kdiscs` comes with multiple layers, for both (scalable) statistical estimation using (in)complete U-statistics, single hypothesis testing, and test aggregation which are both extensible and interoperable. `kdiscs` is implemented in JAX, making use of its `pytree` model in order to support to a very generic class of data structures.

VIII.1 Introduction

“Are two sets of samples distributed equally?” “Is there a relationship between these two variables?” “Does this model fit the data?”. These questions are routinely asked by scientists and practitioners across a wide range of fields such as medicine to quantify the effect of a treatment, in social sciences to measure the impact of a policy,

or in engineering to assess the performance of a model. They also share the common feature of asking about the properties of the *distribution* of the data, and not about the values of some parameters, another common type of question in statistics. The importance of these such *distributional* questions is well recognized by the statistics and machine learning communities, which have developed, for each question, several methods to answer them. In this paper, we focus on the family of methods that answer these questions using kernels, a major contender in this domain. These methods are scalable, come with strong theoretical guarantees, and come a minimal number of hyperparameters. Their general recipe consists, given a distributional property H_0 , in constructing a measure of deviation D from this property, which can be estimated from the input \mathcal{I} at hand (a combination of random samples – or “data” – and fixed distributions – or “models”). In most cases, it is possible to integrate this estimator in a hypothesis test, in order to determine whether the property holds or not with a controlled number of false positives, and shrinking (as the number of samples grows) number of false negatives. While the minimal variance estimator often has a quadratic sample complexity, linear-time alternative estimators have been proposed to scale such method to large datasets [90, 217]. To increase the test power, multiple methods that can scale to large datasets by trading *aggregating* the results of several such tests have been developed. The theoretical framework behind these methods has grown mature, with Schrab et al. [220] obtaining (minimax) sample complexity rate for aggregated tests, and Domingo-Enrich et al. [60] relating the complexity of the statistical estimation procedure to the desired type-I error. As we show in this paper, all methods in this family share a lot of mathematical structure, calling for a unified implementation. However, to the best of our knowledge, most existing packages only handle a limited subset of these methods [123, 124, 125, 126, 127, 220, 217, 218, 22, 134, 216, 268, 77] and may use different backend engines, limiting their interoperability. To address this issue, we introduce `kdiscs`, a Python library written in JAX [28] for assessing data and model distributional properties using kernel-based measures. `kdiscs` supports measuring most well-known distributional properties, which we detail in the next section. For every property H_0 with associated measure

D , `kdiscs` exposes two main features: a versatile way to estimate the measure D from the input \mathcal{I} , and a hypothesis testing framework in which these estimators can be integrated.

VIII.2 The use-case for `kdiscs`: measuring distributional properties

In this section, we introduce the use-case for `kdiscs`, which consists of assessing distributional properties of data and models.

VIII.2.1 Estimating (Deviations from) Distributional Properties of data and models

When faced with finding out whether a distributional property H_0 of some input data and model \mathcal{I} , a common statistical workflow consists of

- Computing an estimator of a measure quantifying deviations of the data and/or model from H_0
- Running a hypothesis test based on these estimators to detect statistically significant deviations from H_0 .

This general paradigm was used to evaluate a wide range of statistical properties: distributional equality between two samples, (conditional) goodness-of-fit of some model to some data, independence of two variables, and calibration of predictive models [90, 42, 220, 217, 218, 22, 270, 77, 248].

In fact, such distributional properties can all be written as equality between two distributions, whose precise nature depend on the property of interest. As such, many estimators introduced in prior work are estimators of a specific distance between probability distributions, the Maximum Mean Discrepancy (MMD, 90). Given two distributions $\mathbb{U}, \mathbb{V} \in \mathcal{P}(\mathcal{Z})$ and a Reproducing Kernel Hilbert Space \mathcal{H} with associated positive definite kernel $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, we recall that the MMD is defined

as

$$\begin{aligned} \text{MMD}(\mathbb{U}, \mathbb{V}) &:= \sup_{f \in \mathcal{H}: \|f\|_{\mathcal{H}} \leq 1} \left\{ \int f d\mathbb{U} - \int f d\mathbb{V} \right\} \\ &= \left(\iint k d(\mathbb{U} \otimes \mathbb{U}) - 2 \iint k d(\mathbb{U} \otimes \mathbb{V}) + \iint k d(\mathbb{V} \otimes \mathbb{V}) \right)^{1/2} \end{aligned} \quad (\text{VIII.1})$$

The precise nature of \mathbb{U} , \mathbb{V} , k and the estimator of $\text{MMD}(\mathbb{U}, \mathbb{V})$ will depend on the distributional property of interest and the input at hand; it gives rise to metrics D and estimators \hat{D} originally introduced under different names (such as the Kernel Stein Discrepancy, the Kernel Conditional Stein Discrepancy, the Kernel Calibration Error, etc.). Cases falling into this framework, and of interest in this paper include

- equality between two (unknown) distributions \mathbb{P} and \mathbb{Q} given samples $\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$ and $\{Y_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}$ done by estimating $\text{MMD}^2(\mathbb{P}, \mathbb{Q})$ [90]
- equality between an unknown distribution \mathbb{P} and a distribution \mathbb{Q} with known and differentiable density q w.r.t the Lebesgue measure given samples $\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}$, done by estimating $\text{KSD}^2(\mathbb{Q})$ [42, 157, 84]
- almost-everywhere equality between a conditional distribution $\mathbb{P}_{|}(Y \in \bullet | X = \diamond)$ and a “model” conditional distribution $\mathbb{Q}_{|}(Y \in \bullet | X = \diamond)$ with known and differentiable density $q(\bullet | x)$ for almost-all x , given samples $\{(X_i, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY} := \mathbb{P}_X \otimes \mathbb{P}_{|}$, done by estimating $\text{KCSd}^2(\mathbb{Q}_{|})$ [127]
- almost-everywhere equality between two conditional distributions $\mathbb{P}_{|}(Y \in \bullet | X = \diamond)$ and $\mathbb{Q}_{|}(Y \in \bullet | X = \diamond)$ $\{(X_i, Y_i, Y'_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes \mathbb{Q}_{|}$ done by estimating $\text{ACMMD}^2(\mathbb{P}_{|}, \mathbb{Q}_{|})$ [248]
- statistical calibration of a predictive model $Q_{|\bullet} : x \mapsto Q_{|x} \in \mathcal{P}(Y)$ given samples $\{(Q_{|X_i}, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY} := \mathbb{P}_{Q_{|X}} \otimes \mathbb{P}_{|Q_{|X}}$ done by estimating $\text{SKCE}^2(Q_{|\bullet})$ [270]
- statistical calibration of a predictive model $Q_{|\bullet}$ with known and differentiable density $Q_{|x}(\bullet)$ w.r.t the Lebesgue measure almost everywhere, given samples $\{(X_i, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY} := \mathbb{P}_X \otimes \mathbb{P}_{|}$ and done by estimating $\text{KCCSD}^2(Q_{|X})$ [77]

Algorithm 14 General Single Hypothesis Testing Procedure for assessing a property H_0

Input: \mathcal{I}
Parameters: Level α ,
Estimate \widehat{D} with \mathcal{I}
Estimate $\widehat{q}_{1-\alpha}$, the $(1 - \alpha)$ -quantile of \widehat{D} under H_0 ,
if $\widehat{D} \leq \widehat{q}_{1-\alpha}$ **then**
 Fail to reject H_0
else
 Reject H_0
end if

A summary of each setting, and how their associated metrics relate to a given MMD is provided in Table VIII.2.1 As all such metrics are MMDs, defining them and their associated and estimators require choosing specific kernels (and the kernels hyperparameters) on the relevant domains. Moreover, calibration metrics such as the SKCE and the KCCSD are MMDs between distributions *on distributions*: this domain is non-standard, and often requires additional approximations to yield tractable estimators [270, 77, 248]

VIII.2.2 Integrating estimators into hypothesis tests

Given a property H_0 with associated metric D , and input \mathcal{I} , it is common to not only compute an estimator \widehat{D} , but to also run a *hypothesis test* based on it. Given \mathcal{I} , this test returns a binary value r indicating its estimate about whether the property H_0 holds. These tests can typically (approximately) control their false positive rate $\mathbb{P}[v|H_0] = \alpha$ to a desired level $\alpha \in [0, 1]$, while lower bounds on their *power* $\mathbb{P}[v|H_1]$ depending on the number of samples n often exists, sometimes matching minimax upper bounds.

The general pseudocode for such a test is given in Algorithm 14. It proceeds by (i) computing \widehat{D} , (ii), estimating the $(1 - \alpha)$ -quantile of \widehat{D} under H_0 , and (iii) rejecting H_0 if \widehat{D} is larger than this quantile.

Leveraging multiple estimators using composite hypothesis tests As discussed above, the kernels involved in the estimators introduced in the previous section typically come with hyperparameters. The values of these hyperparameters can

Framework	Prop. H_0	Input \mathcal{I}	D	MMD formulation
Two-sample	$X \stackrel{d}{=} X'$	$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X$ $\{X'_i\}_{i=1}^m \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{X'}$	MMD^2	$\text{MMD}^2(\mathbb{P}_X, \mathbb{P}_{X'})$
Goodness-of-fit	$X \sim Q$	$\{X_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_X$ $Q \in \mathcal{P}(\mathcal{X})$	KSD^2	$\text{MMD}^2(\mathbb{P}_X, Q)$
Independence	$X \perp\!\!\!\perp Y$	$\{(X_i, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$	HSIC^2	$\text{MMD}^2(\mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y)$
Conditional Goodness-of-fit	$Y X \sim Q(Y X)$	$\{(X_i, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$ $Q(Y X)$	KCSD^2	$\text{MMD}^2(\mathbb{P}_{XY}, \mathbb{P}_X Q_{Y X})$
Calibration	$Y Q \sim Q$	$\{(Y_i, Q_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XQ}$	KCE^2	$\text{MMD}^2(\mathbb{P}_{YQ}, Q\mathbb{P}_Q)$
Calibration	$Y Q \sim Q$	$\{(Y_i, Q_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}_{XQ}$	KCCSD^2	$\text{MMD}^2(\mathbb{P}_{YQ}, Q\mathbb{P}_Q)$

Table VIII.2.1: Properties measurable using *kdiscs* with associated measures

have a strong impact on the power of the associated hypothesis test; many heuristics exist to select sensible value before running it, perhaps the most simple and popular being the median heuristic [90]. An alternative to heuristically setting the kernel’s hyperparameters to a single value consists in constructing *composite* tests, e.g. tests that leverage estimators of metrics D associated with different hyperparameters to compute its decision. This approach can yield tests with much stronger performance than standard ones: recently, Schrab et al. [220] showed that *aggregated* tests—a specific class of composite tests—achieve minimax optimal power rates (also known as *separation rates*) up to logarithmic factors.

VIII.3 *kdiscs*

VIII.3.1 Overview

Our contribution, *kdiscs*, is a Python package, written in JAX [28], containing implementations of the estimators and algorithms involved in the use-case described above. Namely, *kdiscs* provides:

- an implementation of estimators which can be used to quantify deviations to an array of well-known distributional properties of data and models (given in Table VIII.2.1)

- the implementation of single and composite hypothesis tests based on these estimators to detect statistically significant deviations from these properties, obeying a user-specified level of false positives.

From a technical standpoint, the key aspect making *kdiscs* standing out compared to existing alternative implementations [123, 124, 125, 126, 127, 220, 217, 218, 22, 134, 216, 268] is that it was designed in a modular manner, with most mathematical concepts involved being implemented as their independent type. These concepts can be arbitrarily composed to construct a property-assessemment workflow consisting of any well-defined combination of property H_0 , kernels k , (Composite) Hypothesis test T and estimator \widehat{D} . The main advantages of this design are that

- *kdiscs* exhibits much more feature coverage than any of its alternatives
- *kdiscs* can be easily extended to include a new estimator, kernel, property, or hypothesis testing technique, which will readily be compatible with all other components.

This modularity is illustrated in the code snippet VIII.3.1, where *kdiscs* is used to determine whether two sets of samples come from the same distribution by (1) estimating the MMD between their respective distribution using a Gaussian kernel, and the standard MMD's complete U-statistic estimator (See, e.g. 90, Equation 4, or first line, second column of Table VIII.3.1), and (2) running a hypothesis test using this statistic, with bandwidth selected using the median heuristic. Across lines 14-21, this snippet creates a Gaussian Kernel, an MMD U-Statistic kernel using that kernel, a complete U-Statistics based on that U-statistics kernel, and a hypothesis test using that statistic. These choices can be easily changed to, for instance, use a Laplace kernel instead, test independence instead of equality in distribution, use an incomplete U-statistic with R subdiagonals, or use an aggregated test instead, and this by modifying a single line without changing any other part of the code.



```

● ● ●

1 >>> from jax import random
2 ... from kdiscs.rkhs.kernels import gaussian_kernel
3 ... from kdiscs.statistical_tests.base import SingleTest
4 ... from kdiscs.statistical_tests.mmd import OneSampleMMDUStatFn
5 ... from kdiscs.statistical_tests.u_statistics import CompleteUStat
6 ...
7 ... # Generate synthetic data
8 ... X = random.normal(random.PRNGKey(0), (1000,))
9 ... Y = random.normal(random.PRNGKey(1), (1000,)) + 0.1
10 ...
11 ...
12 ... # Compute the MMD U statistic estimate
13 ... # see Equation 4 of (Gretton et. Al, 2012)
14 ... kernel = gaussian_kernel.create(sigma=1.0)
15 ... stat_fn = OneSampleMMDUStatFn(kernel)
16 ... u_stat = CompleteUStat(stat_fn)
17 ... mmd_est, _ = u_stat.compute((X, Y))
18 ... print(f"MMD estimate: {mmd_est:.3f}")
19 MMD estimate: 0.141
20 >>> # Create and run hypothesis test based on it
21 ... test = SingleTest.create(statistic=u_stat, num_permutations=500, median_heuristic=True)
22 ...
23 ... result = test((X, Y), random.PRNGKey(2))
24 ... print(f"reject: {result.reject}")
25 reject: True

```

Figure VIII.3.1: Running a two-sample test using *kdiscs*

VIII.3.2 Structure and Features

We now briefly introduce the main types and features exposed by *kdiscs*, and the contracts they obey.

VIII.3.2.1 Statistics

Perhaps the most important construct of *kdiscs* is the `Statistic[T]` type, a parameterized type which represents an estimator \hat{D} of a metric D associated with a property H_0 and input \mathcal{I} with type T . Its main method is the `Statistics.compute(I: T)` which returns a (i) the value of the estimator \hat{D} and (ii) a `NullDistribution` object, an approximation of the distribution of \hat{D} under H_0 , which can be sampled from and used to estimate its $(1 - \alpha)$ quantile. We provide its most basic type declaration in Snippet VIII.3.3.

U-statistics subtypes An important and unifying subtype of the `Statistic[T]` type is the `BaseUStat[F, T]` type, which represents the type of (possibly incomplete) one-sample, second-order U-statistics with (U-statistics) kernel¹ of type F

¹In this paper, two notions of “kernel” are used: (i) positive definite kernels k used to define RKHS and MMDs, and (ii) U-statistics kernels h , which are symmetric functions being averaged in U-statistics. These two notions should not be confused; when necessary, we will refer to them as “p.d

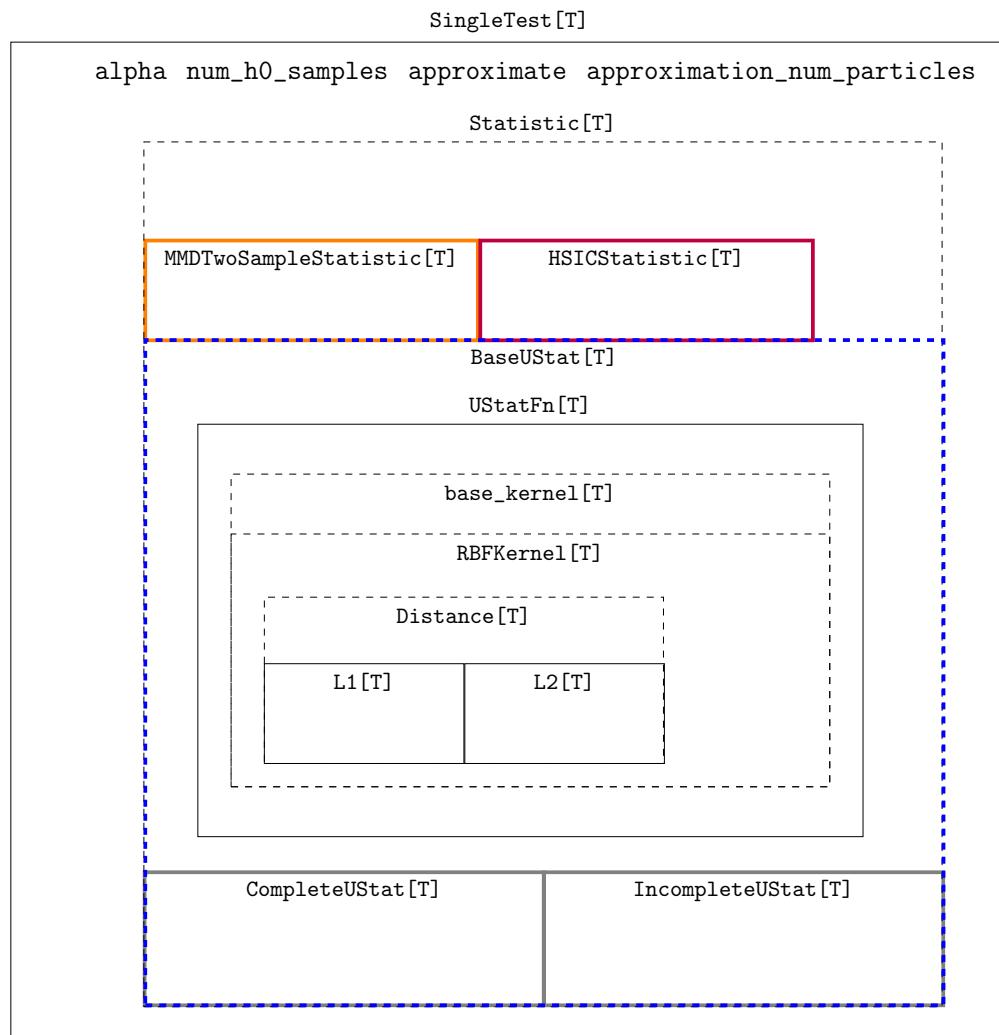


Figure VIII.3.2: Main type structure of kdiscs. Dashed boxes indicate abstract types, while full boxes indicate concrete types. Subtypes of a given type are drawn inside their parent type, with no separation between the boundary of their parent type and their own, or their siblings.

```

1 class Statistic(Generic[T], struct.PyTreeNode, metaclass=abc.ABCMeta):
2     @abstractmethod
3         def compute(
4             self, X: T, /, intermediates: Optional[Intermediate_T] = None
5         ) -> Tuple[Scalar, NullDistribution]:
6             raise NotImplementedError
7
8     @abstractmethod
9         def with_median_heuristic(
10             self: Self, sample: T, key: jax.random.KeyArray, max_num_distances: int
11         ) -> Self:
12             raise NotImplementedError
  
```

Figure VIII.3.3: The `Statistic` type and its main method

(a subtype of `UStatFn` type), and input of type `T`. `kdiscs` defines two subtypes of this type: `CompleteUStat[T]` and `IncompleteUStat[T]`, standing for complete and incomplete U-statistics respectively. We recall that given n i.i.d samples $\{Z_i\}_{i=1}^n \stackrel{\text{i.i.d}}{\sim} \mathbb{U} \in \mathcal{P}(\mathcal{Z})$ a one-sample, second-order *complete* U-statistics [109] with symmetric kernel $h : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$ is given by

$$U := \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} h(Z_i, Z_j) \quad (\text{VIII.2})$$

while an *incomplete* U-statistics with indices subset $\mathcal{D} \subseteq \{(i, j) \in [n] \times [n] : i \neq j\}$ is defined by:

$$U := \frac{1}{|\mathcal{D}|} \sum_{(i,j) \in \mathcal{D}} h(Z_i, Z_j) \quad (\text{VIII.3})$$

with complete U-statistics being a special case of an incomplete one. The main interest of incomplete U-statistics is their computational friendliness: when $|\mathcal{D}| = O(n)$, U will be computable in linear time, while its complete counterpart requires $O(n^2)$ time. The statistical properties of incomplete U-statistics were studied in [23], while the properties of hypothesis tests based on them were studied in [217]. The default indices subset \mathcal{D} used in `IncompleteUStat[T]` is the set of R subdiagonals of the complete U-statistics matrix, with R being a user-specified parameter, yielding an indices set \mathcal{D} of size $|\mathcal{D}| = nR - R(R+1)/2$.

The important role played by second-order U-statistics in `kdiscs` comes from the fact that *all* properties introduced in the previous section admit a U-statistics-based estimator [217, 270, 77]. An overview of all such U-statistics is given in Table VIII.3.1, along with their associated kernels. The `BaseUStat[T]` class provides the general logic, given an arbitrary kernel h , to (i) compute the complete and incomplete U-statistics formulas in Equations VIII.2 and VIII.3, and (ii) provide an approximation of its distribution under H_0 , thereby decoupling the specification of the kernel and its use in U-statistics estimates and further hypothesis tests. We provide the high-level structure of the `BaseUStat[T]` type in Figure VIII.3.4. Subclass of this type should implement (i) a way to generate the indices subset \mathcal{D} , (ii) a way to “kernels” and “U-statistics kernels” respectively.

map evaluation of u-stats kernels over different pairs of inputs Z_i, Z_j , and (iii) a way to sum (“reduce”) all these evaluations to form the final result.



```

1 class BaseUStat(Generic[F, T], Statistic[T]):
2     def compute(
3         self: "BaseUStat[F[T], T]",
4         X: T,
5     ) -> Tuple[Scalar, NullDistribution]:
6         n = infer_num_samples_pytree(X)
7         indices = self._generate_indices(n)
8         func_evals = self._map(self.stat_fn, X, indices)
9         stat_val = self._reduce(func_evals, n)
10        null_dist = self._make_null_distribution(n, func_evals, indices)
11        return stat_val, null_dist
12
13    @abstractmethod
14    def _generate_indices(self, n: int) -> Tuple[Array, Array]:
15        raise NotImplementedError
16
17    @abstractmethod
18    def _map(
19        self,
20        f: Callable[[T1, T1], T2],
21        X: T1,
22        indices: Optional[Tuple[Array, Array]] = None,
23    ) -> T2:
24        raise NotImplementedError
25
26    @abstractmethod
27    def _reduce(self: "BaseUStat[F[T], T]", func_evals: Array, n: int) -> Scalar:
28        raise NotImplementedError
29
30    @abstractmethod
31    def _make_null_distribution(
32        self, n: int, func_evals: Array, indices: Tuple[Array, Array]
33    ) -> NullDistribution:
34        raise NotImplementedError

```

Figure VIII.3.4: Type declaration of the `BaseUStat` class.

Non U-statistics estimators in `kdiscs` Aside from the U-statistics estimators of Table VIII.3.1, `kdiscs` provides `MMDTwoSampleStatistic[T]`, an implementation of the minimum variance MMD estimator for the two sample problem in the case $n \neq m$ [90, Eqaution 3] and `HSICStatistic[T, U]`, the HSIC estimator of [226, Equation 5], which is a fourth-order (and not a second-order) U-statistics.

VIII.3.2.2 Hypothesis Tests

Single Hypothesis Tests With such a U-statistics class, the resulting hypothesis testing logic, implemented in the `SingleTest[T]` type, is just a thin layer of logic (i) computing the U-statistic (optionally before tuning it using the median heuristic),

Framework	U-statistic	Associated kernel
Two-sample ($n = m$)	$\frac{1}{n(n-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{MMD}}^k((X_i, X'_i), (X_j, X'_j))$ k kernel on \mathcal{X}	$\begin{aligned} h_{\text{MMD}}^k((x, y), (x', y')) := \\ k(x, x') + k(y, y') \\ - k(x, y') - k(x', y) \end{aligned}$
Goodness-of-fit	$\frac{1}{n(n-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{KSD}}^{k,q}(X_i, X_j)$ k kernel on \mathcal{X}	$\begin{aligned} h_{\text{KSD}}^{k,q}(x, x') := k(x, x') \\ \times \langle \nabla \log q(x), \nabla \log q(x') \rangle \\ + \langle \nabla_x k(x, x'), \nabla \log q(x') \rangle \\ + \langle \nabla_{x'} k(x, x'), \nabla \log q(x) \rangle \\ + \nabla_{x'} \cdot \nabla_x k(x, x') \end{aligned}$
Independence	$\frac{1}{[n/2]([n/2]-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{HSIC}}^{k,\ell}(\bar{Z}_i, \bar{Z}_j)$ k kernel on \mathcal{X} , ℓ kernel on \mathcal{Y} , $\bar{Z}_i := ((X_i, Y_i), (X_{i+\lfloor n/2 \rfloor}, Y_{i+\lfloor n/2 \rfloor}))$	$\begin{aligned} h_{\text{HSIC}}^{k,\ell}(\bar{z}, \bar{z}') := \\ \frac{1}{4} \times h_{\text{MMD}}^k((x_1, x_2); (x'_1, x'_2)) \\ \times h_{\text{MMD}}^\ell((y_1, y_2), (y'_1, y'_2)) \\ \bar{z} := ((x_1, y_1), (x_2, y_2)) \end{aligned}$
Conditional Goodness-of-fit	$\frac{1}{n(n-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{KCSD}}^{k,\ell,q \bullet}((X_i, Y_i), (X_j, Y_j))$ k kernel on \mathcal{X} , ℓ kernel on \mathcal{Y} ,	$\begin{aligned} h_{\text{KCSD}}^{k,l,q \bullet}((x, y), (x', y')) := \\ k(x, x') \times \left(\ell(y, y') \times \right. \\ \langle \nabla \log q(y x), \nabla \log q(y' x') \rangle \\ + \langle \nabla_y \ell(y, y'), \nabla_{y'} \log q(y' x') \rangle \\ + \langle \nabla_{y'} \ell(y, y'), \nabla_y \log q(y x) \rangle \\ \left. + \nabla_{y'} \cdot \nabla_y \ell(y, y') \right) \end{aligned}$
Calibration	$\frac{1}{n(n-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{SKCE}}^{k,\ell}((q _{X_i}, Y_i), (q _{X_j}, Y_j))$ k kernel on $\mathcal{P}(\mathcal{Y})$, ℓ kernel on \mathcal{Y} ,	$\begin{aligned} h_{\text{SKCE}}^{k,\ell}((q, y), (q', y')) := \\ k(q, q') \times \left(\ell(y, y') \right. \\ - \mathbb{E}_{Z \sim q} \ell(Z, y') \\ - \mathbb{E}_{Z' \sim q'} \ell(y, Z') \\ \left. + \mathbb{E}_{Z \sim q, Z' \sim q'} k(Z, Z') \right) \end{aligned}$
Calibration	$\frac{1}{n(n-1)} \sum_{\substack{1 \leq i < j \leq n \\ i \neq j}} h_{\text{KCCSD}}^{k,\ell,q \bullet}((q _{X_i}, Y_i), (q _{X_j}, Y_j))$ k kernel on $\mathcal{P}(\mathcal{Y})$, ℓ kernel on \mathcal{Y} ,	$\begin{aligned} h_{\text{KCCSD}}^{k,\ell,q \bullet}((y, q), (y', q')) = \\ h_{\text{KCSD}}^{k,\ell,q':q \rightarrow q}((y, q), (y', q')) \end{aligned}$

Table VIII.3.1: U-statistics estimators for the metrics supported in *kdiscs*

(ii) estimating the $(1 - \alpha)$ quantile of its null distribution (whose construction is deferred to Section VIII.3.2.3), and rejecting (or not) H_0 by comparing the two resulting values. As shown in Snippet VIII.3.1, `SingleTest` instances contain the statistic as an attribute, and can be thus used to create a test using any statistic following the `Statistic[T]` contract.

Composite Hypothesis Tests Additionally, `kdiscs` implements three types of composite hypothesis tests: one based on Bonferroni corrections, one based on the aggregation method of Albert et al. [3], Schrab et al. [220, 218, 217], and one based on the Fuse method of [22]. These composite tests take as input a certain `Statistic` subtype (such as a MMD one sample U-statistic), and a list of kernels (e.g. Gaussian and Laplace kernels with different bandwidths) used to create multiple `Statistic` instances used by the composite test to make its decision. Code snippet VIII.3.5 provides an example of how to construct and run a composite test using the aggregation method using `kdiscs`. This snippet makes use of the `OneSampleAggregatedTest` type, which slightly simplifies creation of the aggregated test in the U-statistics case.

```

● ● ●

1 >>> from jax import random
2 ... from kdiscs.rkhs.kernels import gaussian_kernel
3 ... from kdiscs.statistical_tests.base import SingleTest
4 ... from kdiscs.statistical_tests.mmd import OneSampleMMDUStatFn
5 ... from kdiscs.statistical_tests.u_statistics import CompleteUStat
6 ... from kdiscs.statistical_tests.adaptive.aggregation import AggregationMultipleOneSampleTest
7 ...
8 ... # Generate synthetic data
9 ... X = random.normal(random.PRNGKey(0), (100,))
10 ... Y = random.normal(random.PRNGKey(1), (100,)) + 1
11 ...
12 ... test = AggregationMultipleOneSampleTest.create(
13 ...     stat_fn_cls=OneSampleMMDUStatFn,
14 ...     kernels=[{"kernel": gaussian_kernel.create(sigma=2**i)} for i in range(-5, 5)],
15 ... )
16 ...
17 ... result = test((X, Y), random.PRNGKey(2))
18 >>> print(f"reject: {result.reject}")
19 reject: True

```

Figure VIII.3.5: Constructing and running an aggregated test using `kdiscs`

This subsection concludes the introduction of the two main features of `kdiscs` (Statistics computation and hypothesis testing). We now discuss two important additional features of `kdiscs`, which are crucial to its practical use.

VIII.3.2.3 Null Distributions

We now briefly discuss how *kdiscs* constructs the approximate distributions of statistics under H_0 , henceforth called “null distributions”.

The U-Statistic case U-statistics supported by *kdiscs* come with null distributions \mathbb{P}_0 that use the wild bootstrap strategy to draw their samples. Wild bootstrap-based distributions produce sample $\tilde{Z} \sim \mathbb{P}_0$ using the following procedure

$$\tilde{Z} := \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \varepsilon_i \varepsilon_j h(Z_i, Z_j), \quad \varepsilon_1, \dots, \varepsilon_n \stackrel{\text{i.i.d}}{\sim} \text{Rad} \quad (\text{VIII.4})$$

where Rad is the Rademacher distribution (returning ± 1 w.p $1/2$). The approximate quantiles obtained by drawing B samples from that distribution converge to the true quantiles as $n, B \rightarrow \infty$, and come with (1) non-asymptotic guarantees in the two-sample and SKCE-based calibration settings [60], and asymptotic guarantees in the (conditional) goodness-of-fit scenarios (including the KCCSD-based calibration settings). The number of samples B to draw is left as a user-defined hyperparameter of the resulting hypothesis tests, with a default value of 500.

The non U-statistics case The two remaining non-U-statistic estimators covered by *kdiscs* come with associated null-distribution obtained using a permutation approach [135]. In the two-sample case, a sample from the null distribution is obtained by shuffling the samples $\{X_i\}_{i=1}^n, \{X'_i\}_{i=1}^m$ uniformly at random to construct two datasets, and computing the statistic on the shuffled two samples. In the independence testing case, only the samples $\{X_i\}_{i=1}^n$ are shuffled (thus breaking any dependence between Y_i and the resulting X_i) before recomputing the statistic. While permutation-based null distributions also exist for certain U-statistics, wild-bootstrap based null distributions admit implementations that are more computationally efficient, as they do not require shuffling arrays, an expensive operation.

VIII.3.3 Additional features

VIII.3.3.1 Efficient Composite tests by caching intermediate computations

Composite tests often compute multiple statistics only differing by the parameter value of a given p.d. kernel (such as the bandwidth of a Gaussian kernel). In particular, when the kernel is of the form

$$k(z, z') = \phi\left(\frac{\|z - z'\|}{\sigma}\right) \quad (\text{VIII.5})$$

for a given norm $\|\bullet\|$, a function ϕ and a “bandwidth” parameter σ , the distances $\|Z_i - Z_j\|$ for $(i, j) \in \mathcal{D}$ —which form the main bottleneck, scaling in $\mathcal{O}(d|\mathcal{D}|)$, for computing the statistic—only need to be computed once for each different choice of norm $\|\cdot\|$, regardless of the number of different σ values used in the composite test.

kdiscs leverages this observation to speed up composite tests when computing its sequence of statistics by maintaining a cache containing such shared intermediate quantities, progressively populated as statistics are computed, and which is looked-up when computing a new statistics in order to avoid recomputing them. This reduces the overall computational cost of a composite statistics using a kernel with p different bandwidth values from $\mathcal{O}(pd|\mathcal{D}|)$ to $\mathcal{O}(d|\mathcal{D}|)$.

The rules for specifying which are the intermediate quantities involved are typically specified for a single input pair (z, z') at both the U-Statistics kernel *and* at the p.d kernel level, with the mapping of the computations of such intermediate quantities across all possible indices in \mathcal{D} being handled at the U-statistics (e.g. the `BaseUStat[T]`) level.

For instance, the U-statistics kernel for the MMD statistic, given by:

$$h_{\text{MMD}}^k((x_1, x'_1), (x_2, x'_2)) := k(x_1, x_2) + k(x'_1, x'_2) - k(x_1, x'_2) - k(x_2, x'_1)$$

will ask its p.d kernel (when of the form of Equation VIII.5) to compute the relevant

intermediates for each of the four kernel evaluations appearing in its definition: $\|x_1 - x_2\|$, $\|x'_1 - x'_2\|$, $\|x_1 - x'_2\|$ and $\|x_2 - x'_1\|$. The U-statistic instance will then map this request across pairs $(X_i, X'_i), (X_j, X'_j), (i, j) \in \mathcal{D}$.

From a typing standpoint, p.d kernels and U-statistic kernels implement two methods, `make_intermediate` and `from_intermediate`, responsible respectively for (i) computing the intermediate quantities for a single pair of inputs, and (ii) computing the kernel evaluation from these intermediate quantities. Second, a `StatCollection[T]` type is used to compute multiple statistics by computing intermediate quantities only when needed, store them, and compute the kernel evaluations from them.

Currently, all statistics, U-statistics kernels, and p.d kernels provided by `kdiscs` provide non-trivial caching rules, described in Appendix [VIII.A](#). Again, if one wishes to construct a composite testing pipeline for a new U-statistics kernel not present in `kdiscs`, the only task required to enable caching of intermediate computations is to specify the caching rules (`make_intermediate` and `from_intermediate` methods) at the U-statistics kernel level, as the remaining logic at the other layers (p.d kernels, U-statistics, statistics collection) will be reused.

VIII.3.3.2 Built-in approximation of intractable quantities in calibration settings

Calibration measures such as the SKCE and the KCCSD can contain intractable quantities within their statistics:

- in the case of the SKCE for instance, expectations under the models $q_{|X^i}$ of k are required, but yet these expectations are rarely available in closed form. While [270] suggests that when these quantities are intractable, approximations methods such as quadrature rules should be used instead, it does not provide a final algorithm to compute an approximate statistics, leaving its construction to the user.
- Moreover, in the case of both the SKCE and the KCCSD, it is required to

compute evaluations of p.d kernels on distributions $k(q|_{X_i}, q|_{X_j})$. Powerful instances of such kernels are exponentiated MMD kernels, of the form:

$$k(q, q') := \exp\left(-\frac{\text{MMD}_m^2(q, q')}{2\sigma^2}\right), \quad q, q' \in \mathcal{P}(\mathcal{Y}) \quad (\text{VIII.6})$$

where the MMD is computed using a kernel m on \mathcal{Y} , as well as the exponentiated Generalized Fisher Divergence (GFD) kernels [77]

$$\begin{aligned} k(q, q') &:= \exp\left(-\frac{\text{GFD}_v^2(q, q')}{2\sigma^2}\right), \\ \text{GFD}_v(q, q') &:= \left(\int \|\nabla \log q - \nabla \log q'\|^2 dv\right)^{1/2} \end{aligned} \quad (\text{VIII.7})$$

for a measure $v \in \mathcal{P}(\mathcal{Y})$ with full support on \mathcal{Y} , and the one using the kernelized variant of the GFD, the exponentiated Kernelized Generalized Fisher Divergence (KGFD) kernel [77]

$$\begin{aligned} k(q, q') &:= \exp\left(-\frac{\text{KGFD}_{m,v}^2(q, q')}{2\sigma^2}\right), \\ \text{KGFD}_v(q, q') &:= \left\| \int m(x, \cdot) (\nabla \log q(x) - \nabla \log q'(x)) dv(x) \right\|_{\mathcal{H}_v} \end{aligned} \quad (\text{VIII.8})$$

where m is a p.d kernel on \mathcal{X} , and \mathcal{H}_v is the vector valued RKHS associated with the operator-valued kernel $m(\bullet, \diamond)I_{\mathbb{R}^d}$. All three kernels involve possibly intractable expectations, which need to be approximated. Yet, these kernels possess valuable properties such as universality [77, 39].

Importantly, in both the case of the SKCE and of the KCCSD, *kdiscs* will automatically approximate these quantities for the user, in a way that does not compromise the type-I error of the resulting tests [248]. In particular, *kdiscs* uses the following approximations of the SKCE and KCCSD U-statistics kernels:

$$\begin{aligned} h_{\text{SKCE}}^{k,\ell}((Q|_{X_i}, y), (Q|_{X_j}, y')) &\approx \widehat{k}((\widehat{Q}|_{X_i,1}, y), (\widehat{Q}|_{X_j,1}, y')) \times \left(k(y, y') \right. \\ &\quad \left. - \mathbb{E}_{Z \sim \widehat{Q}|_{X_i,2}} k(Z, y') - \mathbb{E}_{Z' \sim \widehat{Q}|_{X_j,2}} k(y, Z') + \mathbb{E}_{Z \sim \widehat{Q}|_{X_i,2}, Z' \sim \widehat{Q}|_{X_j,2}} k(Z, Z') \right) \end{aligned} \quad (\text{VIII.9})$$

$$\begin{aligned}
h_{\text{KCCSD}}^{k,\ell}((Q|_{X_i}, Y_i), (Q|_{X_j}, Y_j)) \approx & \widehat{k}(\widehat{Q}|_{X_i,1}, \widehat{Q}|_{X_j,1}) \times \left(k(y, y') \right. \\
& \times \left\langle \nabla \log Q|_{X_i}(y), \nabla \log Q|_{X_j}(Y_j) \right\rangle + \left\langle \nabla_y k(y, y'), \nabla_{y'} \log q(y'|x') \right\rangle \\
& \left. + \left\langle \nabla_{y'} k(y, y'), \nabla_y \log q(y|x) \right\rangle + \nabla_{y'} \cdot \nabla_y k(y, y') \right) \quad (\text{VIII.10})
\end{aligned}$$

Here, the value of \widehat{k} , $\widehat{Q}|_{X,1}$ and $\widehat{Q}|_{X,2}$ depends on the kernel k and the type of model used: For an exponentiated (K) GFD kernel $k_{\text{GFD},v}$ of the form of Equations [VIII.7](#) or [VIII.8](#), $\widehat{Q}|_{X,1} = Q|_X$, and $\widehat{k}_{\text{GFD},v} = k_{\text{GFD},\widehat{v}}$, where \widehat{v} is an empirical approximation of v , e.g.

$$\begin{aligned}
\widehat{k}_{\text{GFD},v}(\widehat{Q}|_{X_i,1}, \widehat{Q}|_{X_j,1}) := & \exp \left(-\frac{\text{GFD}_{\widehat{v}}(Q|_{X_i}, Q|_{X_j})^2}{2\sigma^2} \right), \quad \widehat{v} := \frac{1}{p} \sum_{i=1}^p \delta_{V_i}, \\
V_1, \dots, V_p \stackrel{\text{i.i.d.}}{\sim} v \quad (\text{VIII.11})
\end{aligned}$$

For an exponentiated MMD_{*m*} kernel k_{MMD_m} of the form of Equation [VIII.6](#), we have $\widehat{k}_{\text{MMD}_m} = k_{\text{MMD}}$, and $\widehat{Q}|_X$ is an empirical version of $Q|_X$, e.g.

$$\begin{aligned}
\widehat{k}_{\text{MMD}_m}(\widehat{Q}|_{X_i,1}, \widehat{Q}|_{X_j,1}) := & k_{\text{MMD}_m} \left(\frac{1}{p} \sum_{i=1}^p \delta_{S_i}, \frac{1}{p} \sum_{i=1}^p \delta_{S'_i} \right) \\
S_1, \dots, S_p \stackrel{\text{i.i.d.}}{\sim} Q|_{X_i}, \quad S'_1, \dots, S'_p \stackrel{\text{i.i.d.}}{\sim} Q|_{X_j} \quad (\text{VIII.12})
\end{aligned}$$

The logic defining \widehat{k} , and the $Q|_{X,1}$ is specified first at the p.d kernel level. For the SKCE U-statistics kernel, additional empirical approximations $\widehat{Q}|_{X,2}$ are needed to approximate the expectations contained in its formula. Again, these empirical approximation are simply set to be:

$$Q|_{X_i,2} := \frac{1}{p} \sum_{j=1}^p \delta_{T_{i,j}}, \quad T_{i,1}, \dots, T_{i,p} \stackrel{\text{i.i.d.}}{\sim} q, \quad i = 1, \dots, n \quad (\text{VIII.13})$$

To toggle the approximations of $h_{\text{SKCE}}^{k,\ell}$ and $h_{\text{KCCSD}}^{k,\ell}$ given in Equations [VIII.9](#) and [VIII.10](#), *kdiscs* exposes an approximate flag during the test construction, which a user must set to true when known analytic intractability that can be al-

leviated through sampling appear when computing the statistic. The value of p in Equations [VIII.11](#), [VIII.12](#) and [VIII.13](#) can be changed by specifying the `approximation_num_particles` argument (defaulting to 10). Currently, the input models given to the tests must subclass the `SampleableModel` type present in `kdiscs.conditional_models.base` module. An example of running a test requiring numerical approximation is given in snippet [VIII.3.6](#).

The special case of Gaussian models In the case where all components involved take the form of Gaussian objects (Gaussian kernels or distributions), the SKCE and KCCSD U-statistics kernels may be computed in closed form, without resorting to numerical approximations. Indeed, assuming $q = \mathcal{N}(\mu_1, \sigma_1^2 I)$, $q' = \mathcal{N}(\mu_2, \sigma_2^2 I)$, and one can show, using Gaussian integral identities, that

$$\begin{aligned} \text{GFD}_v^2(q, q') &= (d\sigma_b^2 + \|v_b\|^2) \times (1/\sigma_1^2 - 1/\sigma_2^2)^2 \\ &\quad - 2 \times (1/\sigma_1^2 - 1/\sigma_2^2) \times \langle v_b, \mu_1/\sigma_1^2 - \mu_2/\sigma_2^2 \rangle + \|\mu_1/\sigma_1^2 - \mu_2/\sigma_2^2\|^2 \\ \text{KGFD}^2(q, q') &= \frac{\sigma_k(1 + 2\sigma_b^2/\sigma_k^2)^{-(d-1)/2}}{\sigma_b \sqrt{2 + \sigma_k^2/\sigma_b^2}} (f(q, q) + f(q', q') - 2 \times f(q, q')) \\ \text{MMD}^2(q, q') &= g(q, q) + g(q', q') - 2g(q, q') \end{aligned} \tag{VIII.14}$$

assuming $v = \mathcal{N}(v_b, \sigma_b^2 I)$ for the GFD and Kernelized GFD (KGFD) [[77](#)], a Gaussian kernel k with bandwidth σ_k for the MMD and KGFD, and we defined

$$\begin{aligned} f(\mathcal{N}(\mu_e, \sigma_e^2 I), \mathcal{N}(\mu_f, \sigma_f^2 I)) &:= \frac{-\langle \mu_b, (\mu_e + \mu_f) \rangle + \langle \mu_e, \mu_f \rangle}{\sigma_e^2 \sigma_f^2} \\ g(\mathcal{N}(\mu_e, \sigma_e^2 I), \mathcal{N}(\mu_f, \sigma_f^2 I)) &:= \frac{\exp(-\|\mu_e - \mu_f\|^2 / (2(\sigma_e^2 + \sigma_f^2 + \sigma_k^2)))}{(1 + (\sigma_e/\sigma_k)^2 + (\sigma_f/\sigma_k)^2)^{d/2}}. \end{aligned} \tag{VIII.15}$$

Thanks to these formulas, the exponentiated MMD and (K)GFD kernels of Equation [VIII.6](#) and [VIII.7](#), can be evaluated in closed form, and thus, so can the KCCSD U-statistics kernel. Moreover, since, for $X \sim \mathcal{N}(\mu, \sigma^2 I)$, $X' \sim \mathcal{N}(\mu', \sigma'^2 I)$, for the

same Gaussian kernel,

$$\begin{aligned}\mathbb{E}[k(X, y)] &= \frac{\exp(-\|\mu - y\|^2/(2(\sigma^2 + \sigma_k^2)))}{(1 + (\sigma/\sigma_k)^2)^{d/2}} \\ \mathbb{E}[k(X, X')] &= \frac{\exp(-\|\mu - \mu'\|^2/(2(\sigma^2 + \sigma'^2 + \sigma_k^2)))}{(1 + (\sigma/\sigma_k)^2 + (\sigma'/\sigma_k)^2)^{d/2}}\end{aligned}\tag{VIII.16}$$

the SKCE U-statistics kernel can also be computed in closed form in this case. `kdiscs` implements all such closed form formulas in dedicated kernels, (the `ExpMMDGaussianKernel`, the `GaussianExpFisherKernel`, as well as `GaussianExpKernelizedFisherKernel`), which take instances of the class `GaussianConditionalModel` as inputs. These analytical formulas enable useful sensitivity studies: indeed, by comparing, for the same Gaussian input, the results of the hypothesis tests using these analytical formulas with the ones using the approximation formulas of the previous paragraph, one can then assess empirically the impact of the number of approximation particles p on the test power.

VIII.4 Experiments

Finally, we perform and comment on a series of experiments using `kdiscs` to showcase the computational and statistical performance of the algorithms it implements.

`kdiscs` is implemented in JAX: as such, it can run its algorithm on both CPUs and GPUs. Moreover, all `kdiscs` routines are fully jit-compilable using JAX.

VIII.4.1 Synthetic Data Generation Mechanism

The experiments were run on synthetic data whose structure is described below. Overall, all datasets are determined by three parameters (which we vary across experiments): (1) their number of samples n , (2) their dimension d and (3) a “shift” parameter $\delta > 0$ which is a natural quantification of how much the null hypothesis H_0 is violated for the dataset in question. For $\delta = 0$, H_0 holds, while as δ grows, H_0 becomes progressively more violated.

```

1 >>> import jax.numpy as jnp
2 ... from jax import random, vmap
3 ... from kdiscs.conditional_models.gaussian_models import GaussianConditionalModel
4 ... from kdiscs.kernels import ExpMMDKernel
5 ... from kdiscs.rkhs.kernels import gaussian_kernel
6 ... from kdiscs.statistical_tests.base import GenericOneSampleTest
7 ... from kdiscs.statistical_tests.skce import OneSampleSKCEUStatFn
8 ...
9 ... # generate data
10 ... Ps = GaussianConditionalModel(
11 ...     mu=random.normal(random.PRNGKey(0), (100, 2)),
12 ...     sigma=jnp.ones((100,)))
13 ...
14 ... Ys = (
15 ...     vmap(type(Ps).sample_from_conditional, in_axes=(0, 0, None))(
16 ...         Ps, random.split(random.PRNGKey(1), 100), 1
17 ...     )
18 ...     + 1.0
19 ... )
20 ...
21 ... test = GenericOneSampleTest.create(
22 ...     stat_fn_cls=OneSampleSKCEUStatFn,
23 ...     p_kernel=ExpMMDKernel.create(),
24 ...     y_kernel=gaussian_kernel.create(sigma=1.0),
25 ...     # approximate=False # raises NotImplementedError
26 ...     approximate=True
27 ... )
28 ...
29 ... result = test((Ps, Ys), random.PRNGKey(2))
30 ... print(f"reject: {result.result}")
31 reject: True

```

Figure VIII.3.6: Running a SKCE test requiring numerical approximations.

Two Sample In the two sample scenario, we set \mathbb{P} to a d -dimensional standard Normal distribution, $\mathbb{P} = \mathcal{N}(0, I)$, and $\mathbb{Q} = \mathcal{N}(\delta e_1, I)$, where $\delta > 0$ is a shift parameter, and $e_1 = (1, 0, \dots, 0)$ is the first vector of \mathbb{R}^d 's canonical basis. We then generate $\{X_i\}_{i=1}^n, \{X'_i\}_{i=1}^n$ according to Table VIII.2.1.

Goodness-of-fit In the goodness-of-fit case, we set, similarly, $\mathbb{P} = \mathcal{N}(0, I)$, and $\mathbb{Q} = \mathcal{N}(0, \delta e_i)$, and generate $\{X_i\}_{i=1}^n$ as in Table VIII.2.1.

Independence For independence testing, the data $\{X_i, Y_i\}_{i=1}^n$ is generated using

$$X \sim \mathcal{N}(0, I),$$

$$U \sim \mathcal{N}(0, I),$$

$$Y = \delta X + \sqrt{1 - \delta^2} U.$$

Conditional Goodness-of-fit In the conditional goodness-of-fit case, We set $\mathbb{P}_X = \mathcal{N}(0, I)$, $\mathbb{P}(Y|X) = \mathcal{N}(x, I)$, and $Q_{|\bullet} : x \mapsto Q_{|x} = \mathcal{N}(x + \delta e_1, \mathcal{I})$

Calibration Finally, in the calibration setup, we set $\mathbb{P}_X = \mathcal{N}(0, I)$, $\mathbb{P}(Y|X) = \mathcal{N}(x, I)$, and $Q_{|\bullet} : x \mapsto Q_{|x} = \mathcal{N}(x + \delta e_1, \mathcal{I})$. Note that for $\delta > 0$, the model $Q_{|\bullet}$ is not only an inaccurate model of $\mathbb{P}(Y|X)$ it is also not calibrated, as

$$\mathbb{P}(Y|Q_{|X} = \mathcal{N}(\mu, I)) = \mathbb{P}(Y|X = \mu - \delta e_1) = \mathcal{N}(\mu - \delta e_1, I) \neq \mathcal{N}(\mu, I).$$

VIII.4.2 Experiments and Results

We record two key metrics: estimated test power (e.g. the mean rejection rate under when H_0 does not hold) and estimated mean total runtime. To estimate the test power, we run each test on s different datasets using the same data generation mechanism, only differing by the pseudo-randomness with which they were generated, and set our power estimate to be the empirical rejection rate. We also report the total runtime by averaging individual runtimes across these s runs. We report frequentist 95% confidence intervals for these values by empirically estimating the standard deviation of the values. Unless specified, we use Gaussian kernels, and we use the analytical formulas for $h_{\text{SKCE}}^{k,\ell}$ and $h_{\text{KCCSD}}^{k,\ell,q}$ available for Gaussian models provided in the previous section. The tests are run by default using a prescribed type-I error rate of 0.05 and 500 approximate null-distribution samples to estimate $q_{1-\alpha}$. All of our experiments were run sequentially on a machine with 20 CPU cores, and a single Nvidia A100 GPU with 40 GB of vram.

VIII.4.2.1 Benchmarking all scenarios across dimensions

First, we benchmark standard single hypothesis tests in all scenarios covered by `kdiscs` (two sample, goodness-of-fit, independence, conditional goodness-of-fit, calibration), we plot the average rejection rate as a function of the sample size n for different values of the dimension d . We do not notice significant performance differences across scenarios, suggesting that for this data, all tasks are approximately equally hard.

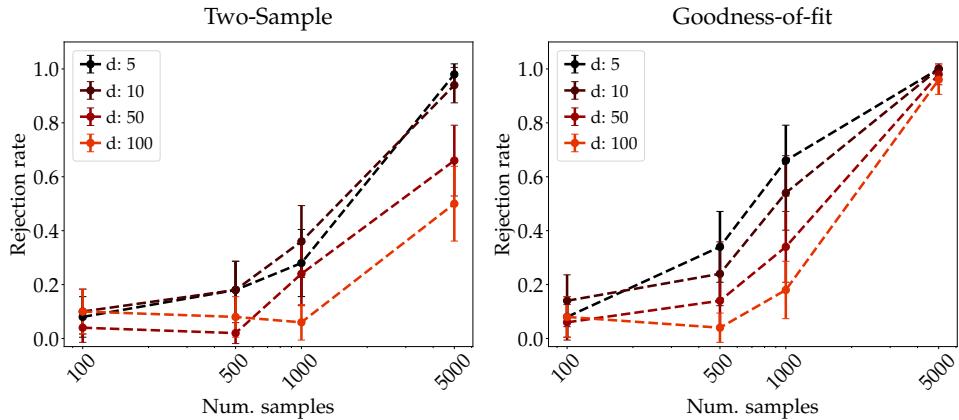


Figure VIII.4.1: Rejection rate across scenarios and data dimensions, using 50 seeds, with the median heuristic, and $\delta = 0.25$

VIII.4.2.2 Performance on CPU vs. GPU

Next, we compare the runtimes obtained by running hypothesis tests with `kdiscs` using either CPU or GPUs. We notice that, for $d = 100$, and n approaches 10^4 , the total runtime using the JAX CPU backend becomes high (around 100 seconds), while with the JAX GPU backend, the runtime remains under 1 second. These runtimes do not account for the JIT compilation times. We perform a further breakdown of the total runtime for the case $d = 100$, by reporting separately the total time required for (1) computing the test statistic, (2) estimating $q_{1-\alpha}$ by sampling from the null distribution, and (3) tuning the kernel’s bandwidth using the median heuristic. In the non-calibration scenarios (Two sample, goodness-of-fit, conditional goodness-of-fit), we see that for $n = 100$, all times are comparable, while as $n = 10000$, computing the U-statistic becomes the main bottleneck, in particular compared to estimating the quantiles. While both tasks in theory should have comparable runtimes for such hyperparameters ($\mathcal{O}(n^2d)$ for computing the u-statistic, $\mathcal{O}(n^2B + B \log B)$ for estimating $q_{1-\alpha}$ with B null distribution samples). In the calibration settings, we see that for low sample sizes ($n = 100$ to 1000), performing the median heuristic takes up the majority of the time, due to the need to estimate the bandwidths kernels used in the MMD of the exponentiated MMD kernels. On GPU, the situation is even worse, with the median heuristic taking the majority of the time for in all sample sizes. To alleviate this cost, the user can set the number of indices pairs used to

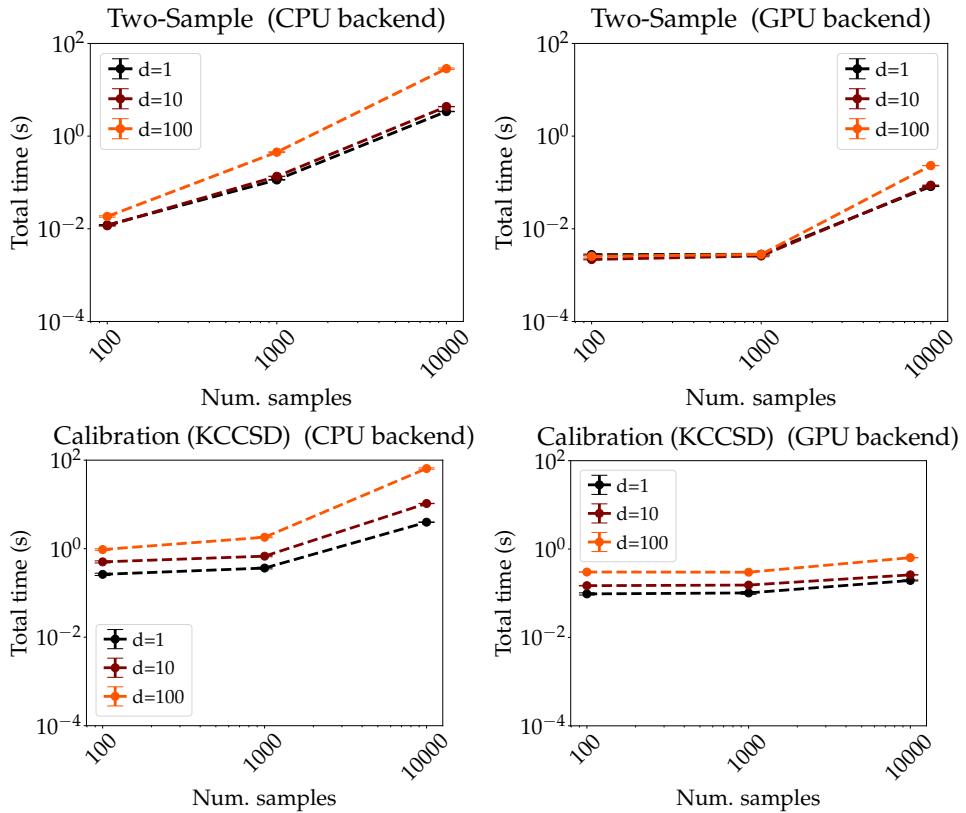


Figure VIII.4.2: Total runtime of hypothesis test in a non-calibration and a calibration scenario as a function of the number of samples, on CPU and GPUs, averaged over 50 seeds, with the median heuristic, and 5000 permutations

estimate the median (5000 by default) by setting the `max_num_distances` argument of the test constructor.

VIII.4.2.3 On using Complete vs. Incomplete U-Statistics

Next, we run an experiment to compare the power of the hypothesis tests obtained by using a complete U statistics or an incomplete one with R subdiagonals. The average rejection rate and runtime for the case of a two-sample test are provided in Figure VIII.4.5. We see three main takeaways from this experiment. For a fixed number of samples, the quadratic test perform much better than its linear counterparts, as expected. Second, the compute-cost per kernel evaluation of the incomplete U-statistic is higher than the one of the complete U-statistic, due to the fact that complete U-statistics and their null distributions can be computed using highly matrix-vector and matrix-matrix products, while incomplete U-statistics require gathering individual kernel evaluations. However, interestingly, we see that a

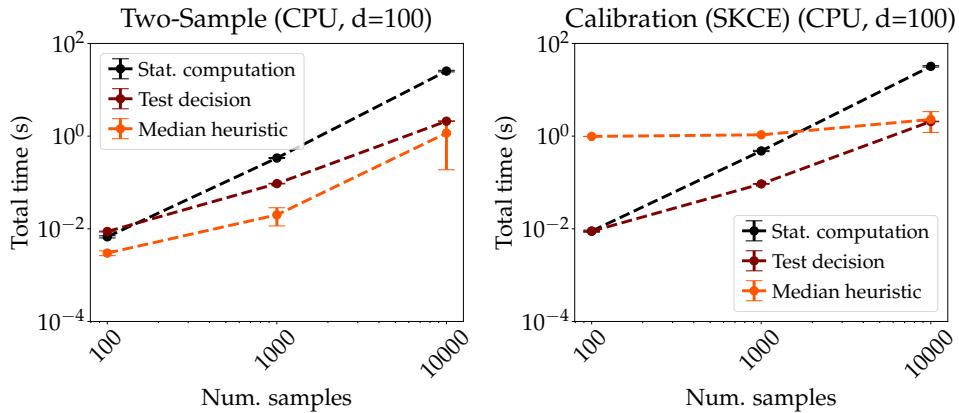


Figure VIII.4.3: Cost Breakdown in the two-sample and calibration (SKCE) scenarios using the CPU backend

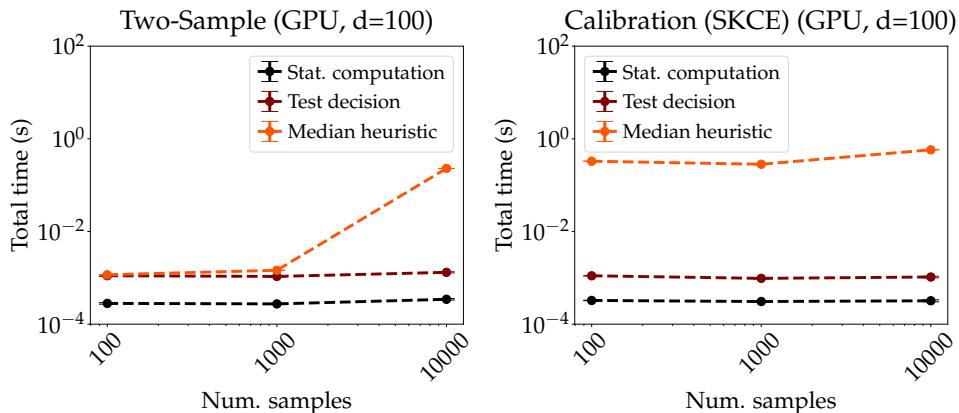


Figure VIII.4.4: Cost Breakdown in the Two sample and Calibration (SKCE) scenarios using the GPU backend

linear-time test using $n = 15000$ samples and $R = 50$ subdiagonals, which amounts to $750000 \approx 866^2$ kernel evaluations, has a much better test power than a quadratic test using $n = 866$ with the same number of kernel evaluations (0.4 vs. 0.7). The sample diversity conferred by the larger sample size of the linear test thus gives the latter clear edge over the quadratic test.

VIII.4.2.4 Impact of caching intermediate quantities in multiple tests

Next, we run an experiment to assess the impact of caching intermediate quantities when running composite tests. We use composite tests using complete U-statistics, with 11 Gaussian kernels with log-2 bandwidths evenly spread between -5 and 5, and 11 Laplacian kernels with the same bandwidth. We record the total runtime of running the test with and without caching intermediate quantities. The results are

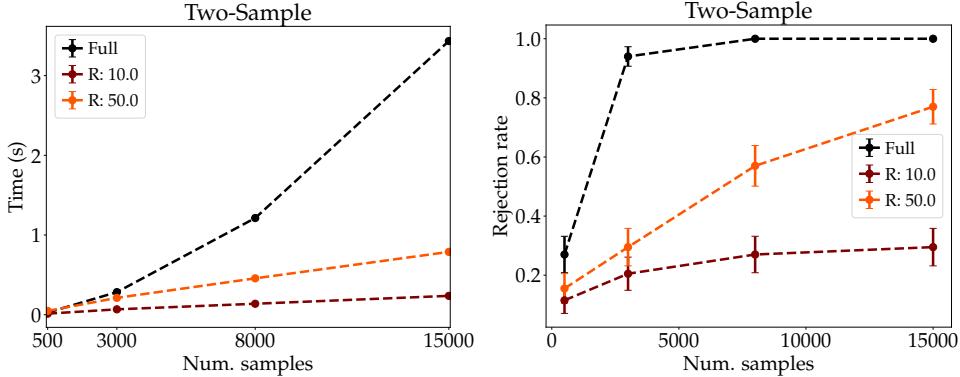


Figure VIII.4.5: Runtime and average rejection rate of an MMD test, using 100 seeds, $\delta = 0.1$, and $d = 2$.

reported in Figure in the case of the two-sample test in VIII.4.6, where we plot the cumulative runtime of the test as a function of the number of the individual statistics computed. For $B=500$ null distribution samples, the total runtime without caching is around 60 seconds, while being around 10 seconds with caching enabled. The runtime bump at the 12th statistics appearing for caching enabled illustrate the need to compute new intermediates when switching kernel types, as the Laplace kernel requires computing ℓ_1 distances $\|z - z'\|_1 := \sum_{i=1}^n |z_i - z'_i|$, while the Gaussian kernel requires (squared) ℓ_2 distances $\|z - z'\|_2^2 = \sum_{i=1}^n (z_i - z'_i)^2$. Interestingly, the cost of computing $q_{1-\alpha}$, which in theory scales in $\mathcal{O}(n^2B)$, is almost negligible compared to the cost of computing the U-statistic, which scales in $\mathcal{O}(n^2d)$, even for $B = 5000$ and $d = 50$. This likely comes from the fact that the former may have a larger space complexity, as it may require the computation of intermediate quantities $z - z'$, amounting to a total space complexity of $\mathcal{O}(n^2d)$, compared to the space complexity of the latter, which is of $\mathcal{O}(n^2 + nB)$. This further justifies the interest of caching intermediate quantities when running multiple tests.

VIII.4.2.5 Impact of the numerical approximation used in calibration settings

Next, we assess the impact of the approximations performed when computing the KCCSD and the SKCE statistics. To do so, we compare the average rejection rate of two SKCE experiments using the same Gaussian models and Gaussian kernel

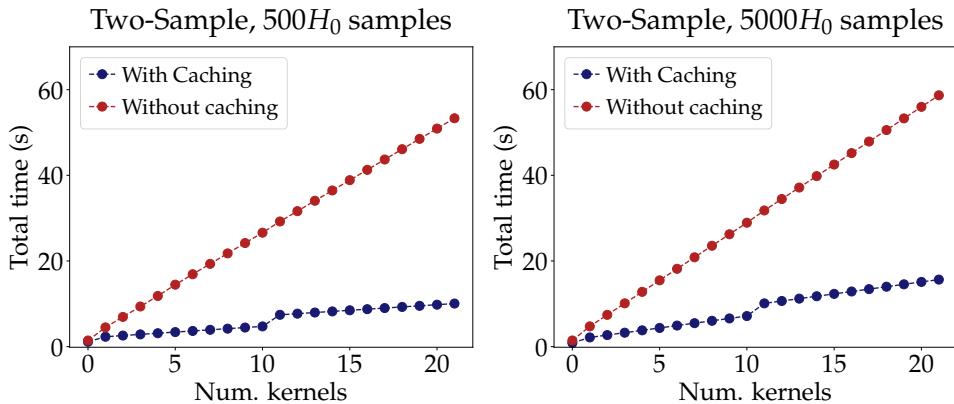


Figure VIII.4.6: Runtime of running a MMD aggregated two-sample test, using a Complete U-statistics with $n = 500$ samples of dimension $d = 50$, samples, enabling or disabling the caching of the intermediate quantities. Results are averaged over 500 seeds.

on \mathcal{Y} , and an exponentiated MMD kernel on $P(\mathcal{Y})$. In the first experiment the SKCE statistic is computed using the closed form formula for the SKCE U-statistics available for Gaussian models and kernels, as well as for exponentiated MMD kernels. In the second experiment, we use the approximation formulas described in Section 3.3. The results are given in Figure VIII.4.7. We see that, in particular for $d = 2$, using a large p does result in an increase in test power, with the test using the exact formula having the best power. However, using even very small p (e.g. $p = 3$) compared to larger $p = 20$ already provides a significant fraction of the test power obtained using the exact formula. Given that the computational cost of computing the statistics scales as $\mathcal{O}(n^2 p^2 d)$, using smaller values of p may still constitute a reasonable compromise between power and computational cost.

Analyzing the impact of approximate p.d kernel on distributions only Finally, we assess the impact of using exact vs. approximate p.d kernels evaluations, while using the approximate formula for the SKCE U-statistics kernel. This allows to isolate the impact of using approximate kernels on distributions only. The results are given in Figure VIII.4.8.

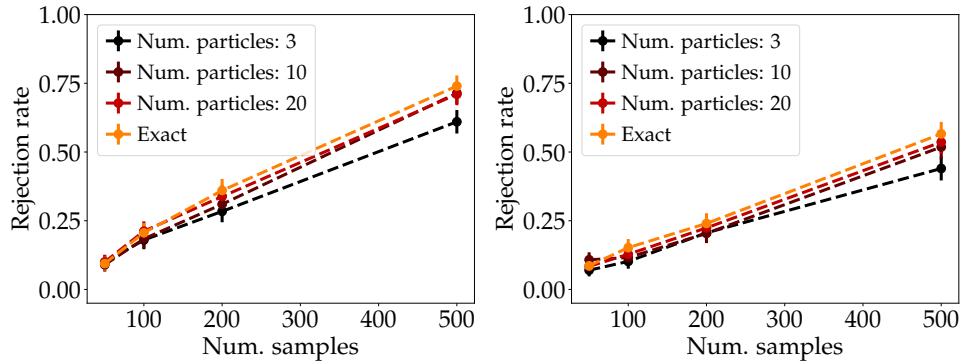


Figure VIII.4.7: Rejection rate of SKCE tests using the exact and approximate formulas for the SKCE U-statistics kernel, in dimension $d = 2$ (left) and $d = 10$ (right), using $\delta = 0.15$, and a varying number of approximation particles p , as well as using the exact formula. Results are averaged over 500 seeds.

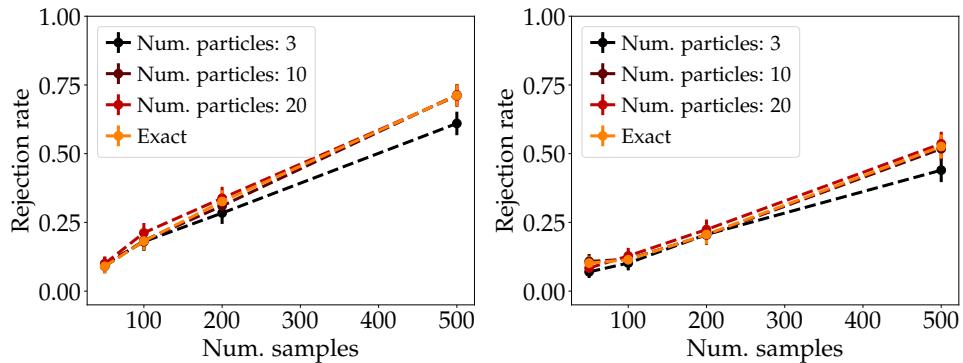


Figure VIII.4.8: Rejection rate of SKCE tests using the approximate formulas for the SKCE U-statistics kernel, in dimension $d = 2$ (left) and $d = 10$ (right), using $\delta = 0.15$ and with either analytical or approximate formulas for kernels on distributions. Results are averaged over 500 seeds.

VIII.5 Conclusion

In this paper, we introduced `kdiscs`, a modern, extensible JAX-based Python package for performing kernel-based hypothesis testing and discrepancy estimation in a variety of settings. With this library we hope to provide the new standard implementation of these methods, and that it will be progressively extended by the community to include new methods and scenarios.

VIII.6 Acknowledgements

P.G and A.G acknowledge support from the Gatsby Charitable Foundation. A.S acknowledges support from the U.K. Research and Innovation (EP/S021566/1).

Appendix

VIII.A List of intermediate quantities by Statistic

Caching of intermediate quantities is only done when the kernel involved are RBF kernel, e.g. kernels for the form

$$k(z, z') = \phi(d(z, z')/\sigma)$$

for a distance function $d : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}_+$, an RBF function $\phi : \mathbb{R}_+ \rightarrow \mathbb{R}$, and a bandwidth parameter σ . Evaluating the distance function is considered to be the main computational bottleneck for computing $k(z, z')$.

VIII.A.1 For U-statistics

For the MMD U-statistic The MMD U-statistic kernel is given by

$$h_{\text{MMD}}^k((x_1, x_2), (x'_1, x'_2)) = k(x_1, x'_1) + k(x_2, x'_2) - k(x_1, x'_2) - k(x_2, x'_1)$$

If one seeks to evaluate this function for multiple bandwidths, it suffices to compute once $a = d(x_1, x'_1)$, $b = d(x_2, x'_2)$, $c = d(x_1, x'_2)$, and $d = d(x_2, x'_1)$, from which the U-statistic kernel can be computed for any bandwidth σ as

$$h_{\text{MMD}}^k((x_1, x_2), (x'_1, x'_2)) = \phi(a/\sigma) + \phi(b/\sigma) - \phi(c/\sigma) - \phi(d/\sigma)$$

For the HSIC paired U-statistic Recall that the HSIC U-statistic kernel is given by

$$\begin{aligned} h_{\text{HSIC}}^{k,\ell}(((x_1, y_1), (x_2, y_2), ((x'_1, y'_1), (x'_2, y'_2)))) := \\ \frac{1}{4} \times h_{\text{MMD}}^k((x_1, x_2); (x'_1, x'_2)) \times h_{\text{MMD}}^\ell((y_1, y_2), (y'_1, y'_2)), \end{aligned}$$

e.g. the product of two MMD U-statistic. If one seeks to evaluate this function for multiple bandwidths for both kernels k and ℓ , it suffices to apply the caching strategy described above for the MMD statistic to both h_{MMD}^k and h_{MMD}^ℓ .

For the KSD U-statistic For RBF kernels, the KSD U-statistic kernel is given by

$$h_{\text{KSD}}^k(x, x') = (1) + (2) + (3) + (4)$$

$$\begin{aligned}
(1) &= \langle \nabla \log q(x), \nabla \log q(x') \rangle \phi(d(x, x')/\sigma) \\
(2) &= \langle \nabla \log q(x), \nabla'_x d(x, x')/\sigma \rangle \phi'(d(x, x')/\sigma) \\
(3) &= \langle \nabla_x d(x, x')/\sigma, \nabla \log q(x') \rangle \phi'(d(x, x')/\sigma) \\
(4) &= (\text{Tr}(\nabla_x \nabla'_x d(x, x'))/\sigma \phi'(d(x, x')/\sigma) + \langle \nabla'_x d(x, x'), \nabla_x d(x, x') \rangle/\sigma^2 \\
&\quad \times \phi''(d(x, x')/\sigma)).
\end{aligned}$$

If one seeks to evaluate this function for multiple bandwidths, it suffices to compute the following terms once and for all:

$a = d(x, x')$	used for (1), (2), (3), (4)
$b = \langle \nabla \log q(x), \nabla \log q(x') \rangle$	for (1)
$c = \langle \nabla \log q(x), \nabla'_x d(x, x') \rangle$	for (2)
$d = \langle \nabla_x d(x, x'), \nabla \log q(x') \rangle$	for (3)
$e = \text{Tr}(\nabla'_x \nabla_x d(x, x'))$	for (4)
$f = \langle \nabla_x d(x, x'), \nabla'_x d(x, x') \rangle$	for (4)

and given these for a specific σ , the KSD U-statistic kernel can be written as

$$\begin{aligned}
h_{\text{KSD}}^k(x, x') &= (\phi(a/\sigma) \times b + \phi'(a/\sigma) \times c + \phi'(a/\sigma) \times d \\
&\quad + (\phi'(a/\sigma) \times e/\sigma + \phi''(a/\sigma) \times f/\sigma^2)).
\end{aligned}$$

For the KCSD U-statistic Assume that $k(y, y') = \phi_1(d_1(y, y')/\sigma_1)$ and $\ell(x, x') = \phi_2(d_2(x, x')/\sigma_2)$ are RBF kernels. The associated KCSD U-statistic kernel is given by

$$h_{\text{KCSD}}^{k, \ell}((x, y), (x', y')) = ((1) + (2) + (3) + (4)) \times (5)$$

where

$$(1) = \langle \nabla \log q(y|x), \nabla \log q(y'|x') \rangle \phi_1(d_1(y, y')/\sigma_1)$$

$$\begin{aligned}
(2) &= \langle \nabla_y \log q(y|x), \nabla_2 d_1(y, y')/\sigma_1 \rangle \phi'_1(d_1(y, y')/\sigma_1) \\
(3) &= \langle \nabla_1 d_1(y, y')/\sigma_1, \nabla_y \log q(y'|x') \rangle \phi'_1(d_1(y, y')/\sigma_1) \\
(4) &= (\text{tr}(\nabla_2 \nabla_1 d_1(y, y'))/\sigma_1 \phi'_1(d_1(y, y')/\sigma_1) + \langle \nabla_1 d_1(y, y'), \nabla_2 d_1(y, y') \rangle)/\sigma_1^2 \\
&\quad \times \phi''_1(d_1(y, y')/\sigma_1)) \\
(5) &= \phi_2(d_2(x, x')/\sigma_2).
\end{aligned}$$

If one seeks to evaluate this function for multiple bandwidths, it suffices to compute the following terms once and for all:

$a = d_1(x, y)$	used for (1), (2), (3), (4)
$b = \langle \nabla \log q(y x), \nabla \log q(y' x') \rangle$	for (1)
$c = \langle \nabla \log q(y x), \nabla_2 d_1(y, y') \rangle$	for (2)
$d = \langle \nabla_1 d(y, y'), \nabla \log q(y' x') \rangle$	for (3)
$e = \text{tr}(\nabla_2 \nabla_1 d_1(y, y'))$	for (4)
$f = \langle \nabla_1 d(y, y'), \nabla_2 d_1(y, y') \rangle$	for (4)
$g = d_2(x, x')$	for (5)

and given these for a specific σ , the Stein kernel can be written as

$$\begin{aligned}
k_s(x, y) &= (\phi_1(a/\sigma_1) \times b + \phi'_1(a/\sigma_1) \times c + \phi'_1(a/\sigma_1) \times d \\
&\quad + (\phi'_1(a/\sigma_1) \times e/\sigma + \phi''_1(a/\sigma_1) \times f/\sigma_1^2)) \times \phi_2(g/\sigma_2).
\end{aligned}$$

For the KCCSD U-Statistic The KCCSD being a special case of the KCSD, it inherits the caching strategy of the latter. Moreover, it uses, for computing approximations of $k_2(q, q')$, the p.d kernel on distributions evaluations, the caching strategy for approximate kernels on distributions described in section VIII.A.2.

For the SKCE Statistic Recall that the SKCE statistic is given by

$$h_{\text{SKCE}}^{k,\ell}((q, y), (q', y')) = k(q, q') \times h_1((q, y), (q', y'))$$

where

$$h_1((q, y), (q', y')) := \ell(y, y') - \mathbb{E}_{Z \sim q} \ell(Z, y') - \mathbb{E}_{Z' \sim q'} \ell(y, Z') + \mathbb{E}_{Z \sim q, Z' \sim q'} \ell(Z, Z')$$

The goal here is to compute the approximate SKCE statistics given in Equation VIII.9. Assuming discretized version of q, q', \hat{q}, \hat{q}' , given by;

$$\hat{q} = \frac{1}{p} \sum_{i=1}^p \delta_{S_i}, \quad \hat{q}' = \frac{1}{p} \sum_{i=1}^p \delta_{S'_i}, \quad S_1, \dots, S_p \stackrel{\text{i.i.d}}{\sim} q, \quad S'_1, \dots, S'_p \stackrel{\text{i.i.d}}{\sim} q' \quad (\text{VIII.17})$$

and that k is an RBF kernel with distance d , intermediate quantities to cache are

$$\begin{aligned} b &:= d(y', y) \\ c &:= (d(S_1, y'), \dots, d(S_n, y')) \\ d &:= (d(S'_1, y), \dots, d(S'_m, y)) \\ e &:= (d(S_i, S'_j))_{i,j} \end{aligned}$$

From which the approximate SKCE U-statistic kernel can be computed for any bandwidth σ as

$$\begin{aligned} h_{\text{SKCE}}^{k,\ell}((q, y), (q', y')) &\approx \phi(b/\sigma) - \frac{1}{p} \sum_{i=1}^p \phi(c_i/\sigma) - \frac{1}{p} \sum_{j=1}^p \phi(d_j/\sigma) \\ &\quad + \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p \phi(e_{i,j}/\sigma). \quad (\text{VIII.18}) \end{aligned}$$

VIII.A.2 For kernels on distributions

RBF kernels on distributions, for the form

$$k(q, q') = \phi(d(q, q')/\sigma), \quad q, q' \in \mathcal{P}(\mathcal{Z})$$

usually cannot be evaluated exactly. Yet, their approximate variant, given by, $\hat{k}(\hat{q}, \hat{q}')$, contains intermediate quantities which can be reused across multiple bandwidths σ . Moreover, the distance d can itself depend on a kernel k_z , and when k_z is an RBF

kernel evaluations of d with bandwidth σ_z also can contain intermediate quantities which can be reused to compute d for different values of σ_z .

For the MMD When $d(q, q') = \text{MMD}(q, q')$, where the MMD is computed using an RBF kernel k_z with bandwidth σ_z , and distance d_1 , and q and q' are approximated by discrete models of the form of Equation VIII.17, the intermediate quantities to cache are the individual distances evaluations of d_1 across all pairs of samples, e.g.

$$A = (d_1^2(S_i, S_j))_{i,j \in [p]^2}, \quad B = (d_1^2(S'_i, S'_j))_{i,j \in [p]^2}, \quad C = (d_1^2(S_i, S'_j))_{i,j \in [p]^2}$$

from which the squared MMD (and consequently the exponentiated MMD kernel) can be computed for any bandwidth σ_z as

$$\text{MMD}^2(q, q') = \frac{1}{p^2} \times (\langle \mathbf{1}, \phi(A/\sigma_z)\mathbf{1} \rangle + \langle \mathbf{1}, \phi(B/\sigma_z)\mathbf{1} \rangle - 2 \langle \mathbf{1}, \phi(C/\sigma_z)\mathbf{1} \rangle)$$

where ϕ is applied element-wise to the matrices A, B, C , and $\mathbf{1}$ is the vector of ones of size p . The Exponentiated GFD kernel can then be computed for any pair bandwidth σ_z, σ as

$$k(q, q') \approx \exp \left(-\frac{1}{\sigma^2} \left(\frac{1}{p^2} \times (\langle \mathbf{1}, \phi(A/\sigma_z)\mathbf{1} \rangle + \langle \mathbf{1}, \phi(B/\sigma_z)\mathbf{1} \rangle - 2 \langle \mathbf{1}, \phi(C/\sigma_z)\mathbf{1} \rangle) \right) \right)$$

For the GFD When $d(q, q') = \text{GFD}_v(q, q')$, assuming v has been discretized into \hat{v} of the form

$$\hat{v} := \frac{1}{p} \sum_{i=1}^p \delta_{V_i}, \quad V_1, \dots, V_p \stackrel{\text{i.i.d}}{\sim} v \quad (\text{VIII.19})$$

The intermediate quantities to cache are

$$a_i = \|\nabla \log q(V_i) - \nabla \log q'(V_i)\|^2, \quad i \in [p]$$

from which the $\text{GFD}_{\hat{v}}$ and the Exponentiated GFD kernel can be computed as

$$\text{GFD}_{\hat{v}}(q, q') = \frac{1}{p} \sum_{i=1}^p a_i, \quad k(q, q') \approx \exp \left(-\frac{1}{\sigma^2} \times \frac{1}{p} \sum_{i=1}^p a_i \right)$$

and the Exponentiated GFD kernel can be computed for any bandwidth σ_z as

For the KGFD When $d(q, q') = \text{KGF}(q, q')$, where the KGFD is computed using an RBF kernel k_z with bandwidth σ_z , and distance d_1 , and q and q' are approximated by discrete models of the form of Equation VIII.17, and assuming v has been discretized and is of the form given by Equation VIII.19, the intermediate quantities to cache are

$$a_i = \|\nabla \log q(V_i) - \nabla \log q'(V_i)\|^2, \quad i \in [p]$$

$$B_{ij} = d_1^2(V_i, V_j), \quad i, j \in [p]^2$$

from which the squared KGFD (and consequently the exponentiated KGFD kernel) can be computed for any bandwidth σ_z as

$$\begin{aligned} \text{KGFD}(q, q') &\approx \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p a_i \times \phi(B_{ij}/\sigma_z) \\ k(q, q') &\approx \exp \left(-\frac{1}{\sigma^2} \times \frac{1}{p^2} \sum_{i=1}^p \sum_{j=1}^p (a_i \times \phi(B_{ij}/\sigma_z)) \right) \end{aligned}$$

VIII.A.3 Kernels and U-statistics Kernels on Gaussian distributions

When q and q' are Gaussian distributions, the closed-form expressions for the MMD, GFD, and KGFD between Gaussian distributions, given in Equations VIII.14, as well as the SKCE U-statistic kernel contain intermediate quantities which can be reused across multiple bandwidths σ , and possibly σ_z , or even the standard deviations σ_b of the base (isotropic) distribution v_b .

VIII.A.3.1 Kernels on Gaussian distributions

For the GFD, given distributions $q = \mathcal{N}(\mu_1, \sigma_1^2 I)$, $q' = \mathcal{N}(\mu_2, \sigma_2^2 I)$, and $v_b = \mathcal{N}(\mu_b, \sigma_b^2 I)$,

$$a = \left\langle \mu_b, \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right) \right\rangle, \quad b = \left\| \frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2} \right\|^2, \quad s = \|\mu_b\|^2$$

from which the GFD can be computed as:

$$\text{GFD}^2(q, q') = (d \times \sigma_b^2 + s) \times \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right)^2 - 2 \times \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) \times a + b$$

For the KGFD, with kernel bandwidth σ_k , one needs to cache

$$a = f(q, q), \quad b = f(q, q'), \quad c = f(q', q')$$

where f is given in Equation VIII.15 from which the KGFD can be computed for any bandwidth σ_k and base distribution standard deviation σ_b as:

$$\text{KGFD}^2(q, q') = \frac{\sigma_k (1 + 2\sigma_b^2/\sigma_k^2)^{-(d-1)/2}}{\sigma_b \sqrt{2 + \sigma_k^2/\sigma_b^2}} (a + c - 2b).$$

Finally, for the MMD, with kernel bandwidth σ_k , one needs to cache

$$a = \|\mu_1 - \mu_2\|^2$$

where g is given in Equation VIII.15, from which the squared MMD can be computed for any bandwidth σ_k as:

$$\begin{aligned} \text{MMD}^2(q, q') = & \frac{1}{(1 + 2(\sigma_e/\sigma_k)^2)^{d/2}} + \frac{1}{(1 + 2(\sigma_f/\sigma_k)^2)^{d/2}} \\ & - 2 \times \frac{\exp(-a^2/2(\sigma_e^2 + \sigma_f^2 + \sigma_k^2))}{(1 + (\sigma_e/\sigma_k)^2 + (\sigma_f/\sigma_k)^2)^{d/2}}. \end{aligned}$$

VIII.A.3.2 SKCE U-statistics kernel between Gaussian Distributions

Note that for the SKCE, the closed-form expression of the expectations under the Gaussian kernel given in Equation VIII.16 given inputs $(q, y), (q', y')$, one can cache

$$A = \|\mu_1 - y\|^2, \quad B = \|\mu_2 - y'\|^2, \quad C = \|\mu_1 - \mu_2'\|^2,$$

from which the SKCE U-statistic kernel can be computed for any bandwidth σ_k as

$$k(q, q') \times \left(\ell(y, y') - \frac{\exp\left(-\frac{A}{2(\sigma^2 + \sigma_k^2)}\right)}{(1 + (\sigma/\sigma_k)^2)^{d/2}} - \frac{\exp\left(-\frac{B}{2(\sigma^2 + \sigma_k^2)}\right)}{(1 + (\sigma/\sigma_k)^2)^{d/2}} \right. \\ \left. + 2 \times \frac{\exp\left(-\frac{C}{2(\sigma^2 + \sigma'^2 + \sigma_k^2)}\right)}{(1 + (\sigma/\sigma_k)^2 + (\sigma'/\sigma_k)^2)^{d/2}} \right)$$

VIII.B Additional Plots for the experiments

VIII.B.1 Benchmarking All Statistics

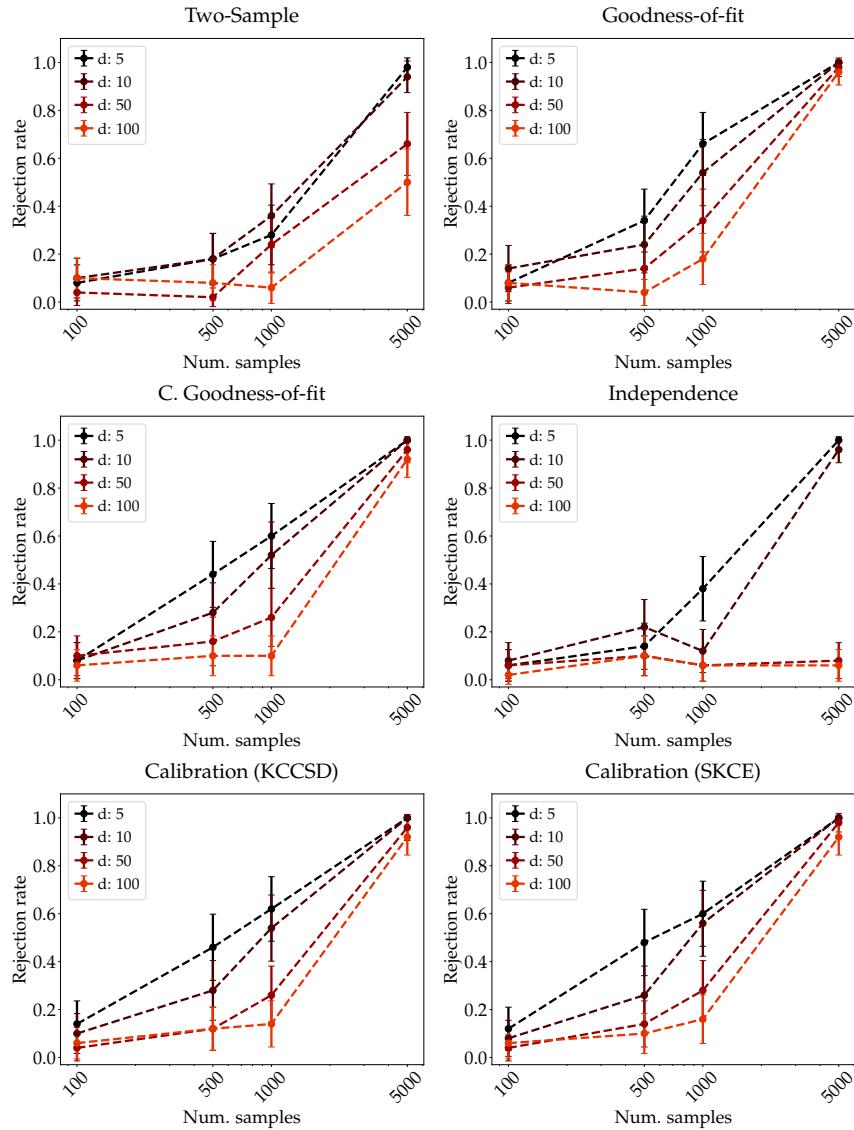


Figure VIII.B.1: Rejection rate across scenarios and data dimensions, using 50 seeds, with the median heuristic, and $\delta = 0.25$

VIII.B.2 CPU vs. GPU runtime

VIII.B.2.1 Total CPU vs. GPU runtime

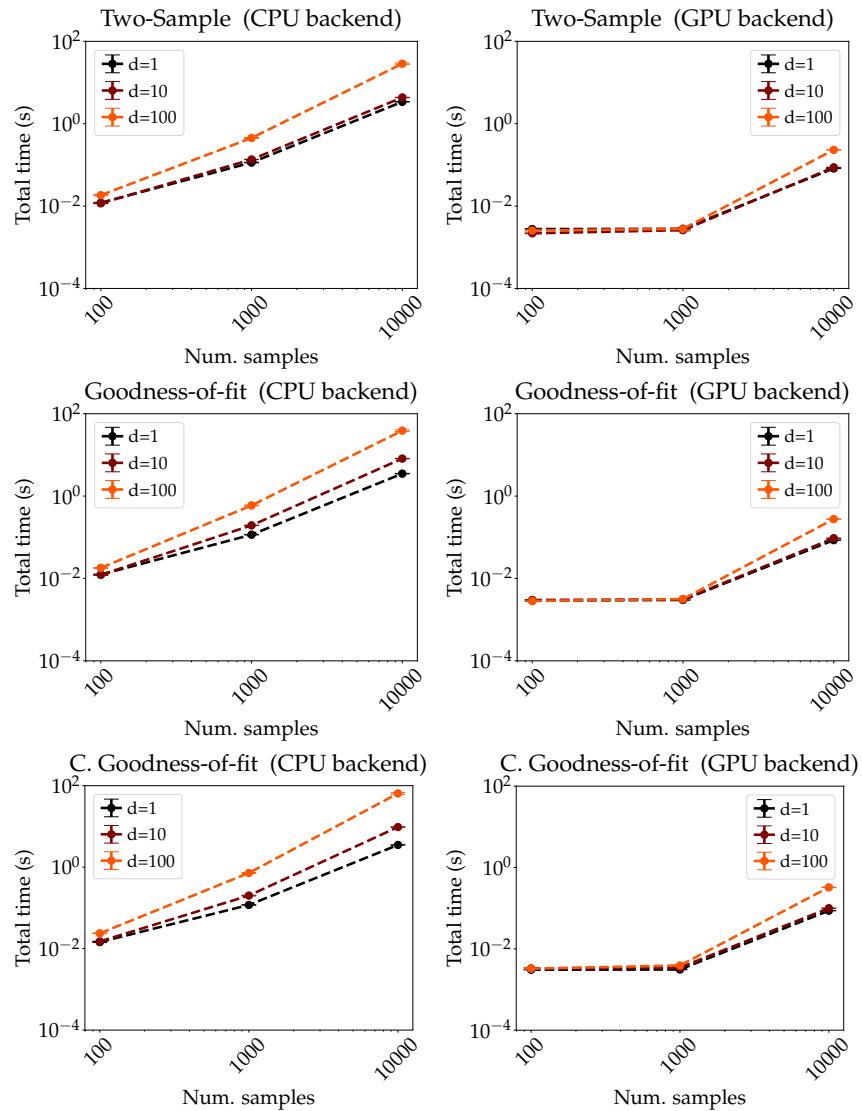


Figure VIII.B.2: Total runtime of hypothesis test in different scenarios as a function of the number of samples, on CPU and GPUs, using the median heuristic, and 50 seeds.

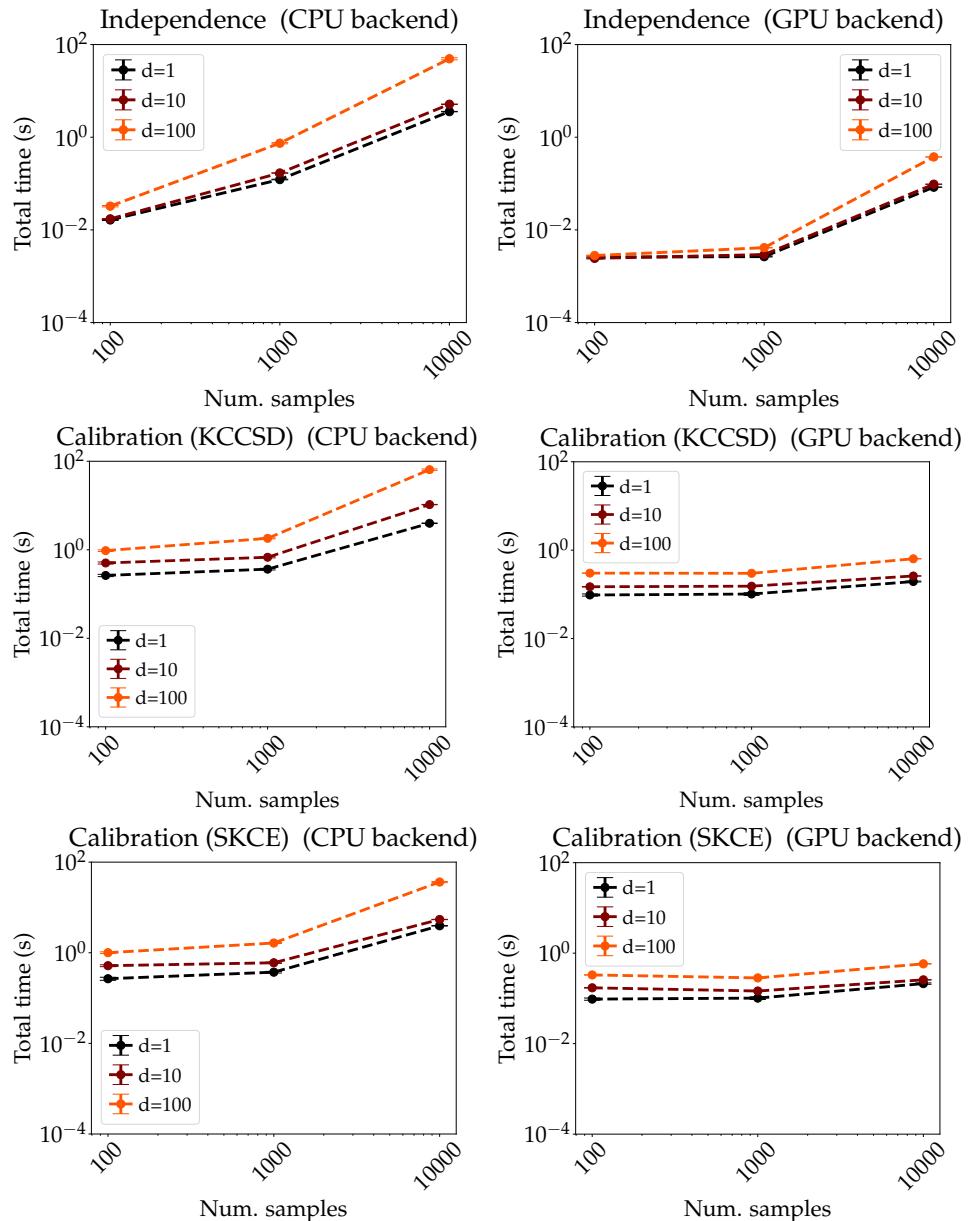


Figure VIII.B.3: Total runtime of hypothesis test in different scenarios as a function of the number of samples, on CPU and GPUs, using the median heuristic, and 50 seeds.

VIII.B.3 Breakdown of total CPU vs. GPU runtime

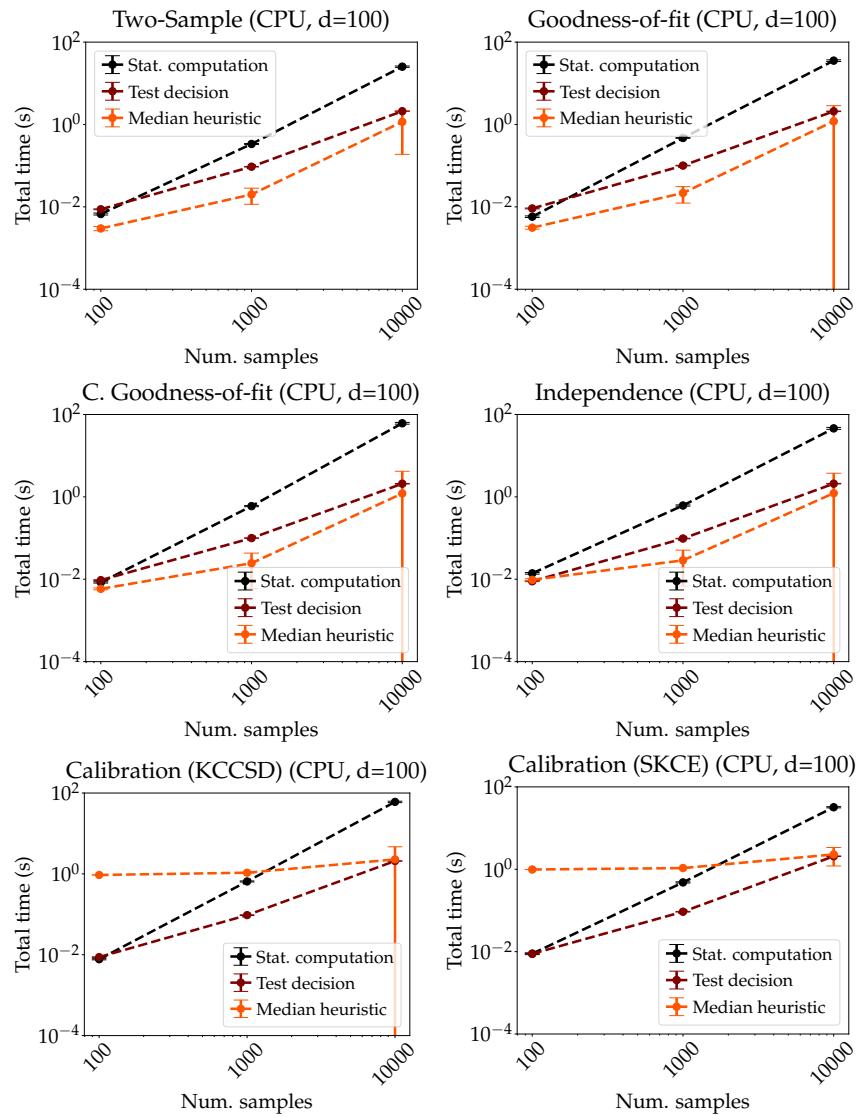


Figure VIII.B.4: Runtime breakdown of hypothesis tests in different scenarios as a function of the number of samples, on CPUs, using the median heuristic, and 50 seeds.

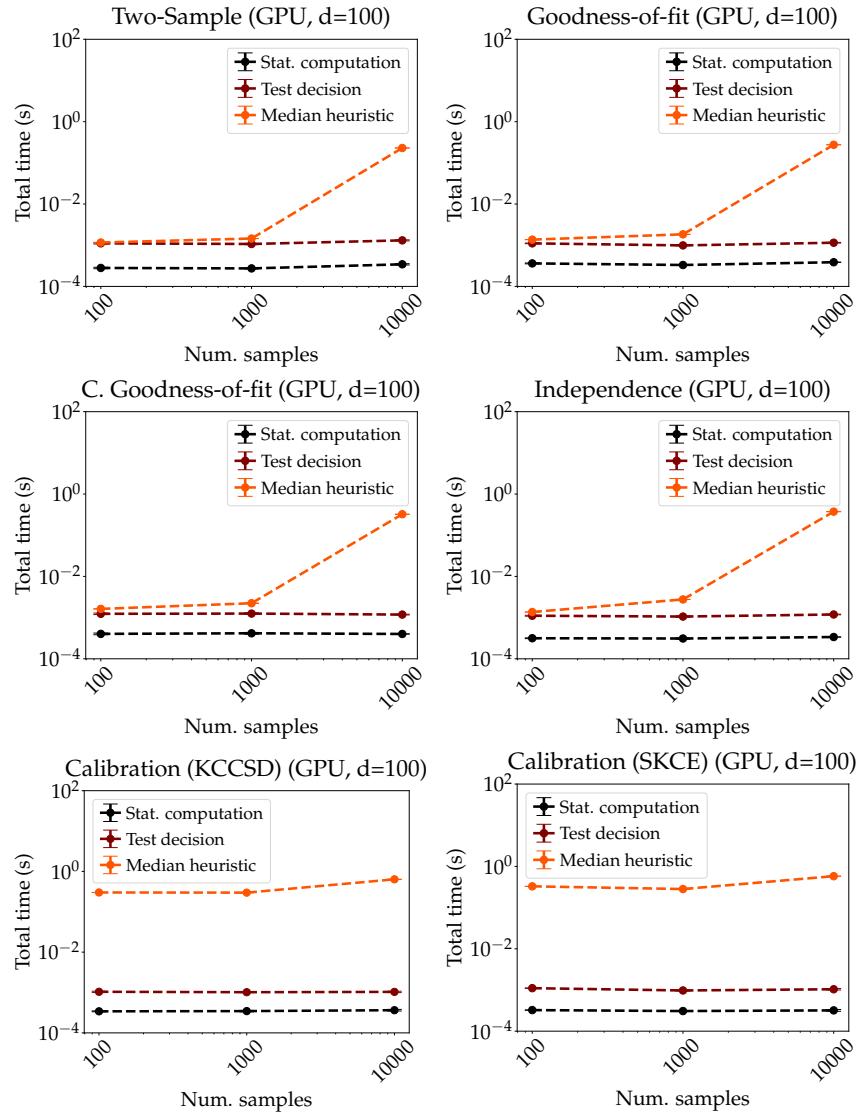


Figure VIII.B.5: Runtime breakdown of hypothesis tests in different scenarios as a function of the number of samples, on GPUs, using the median heuristic, and 50 seeds.

VIII.B.4 Using Quadratic vs. Linear U-statistic

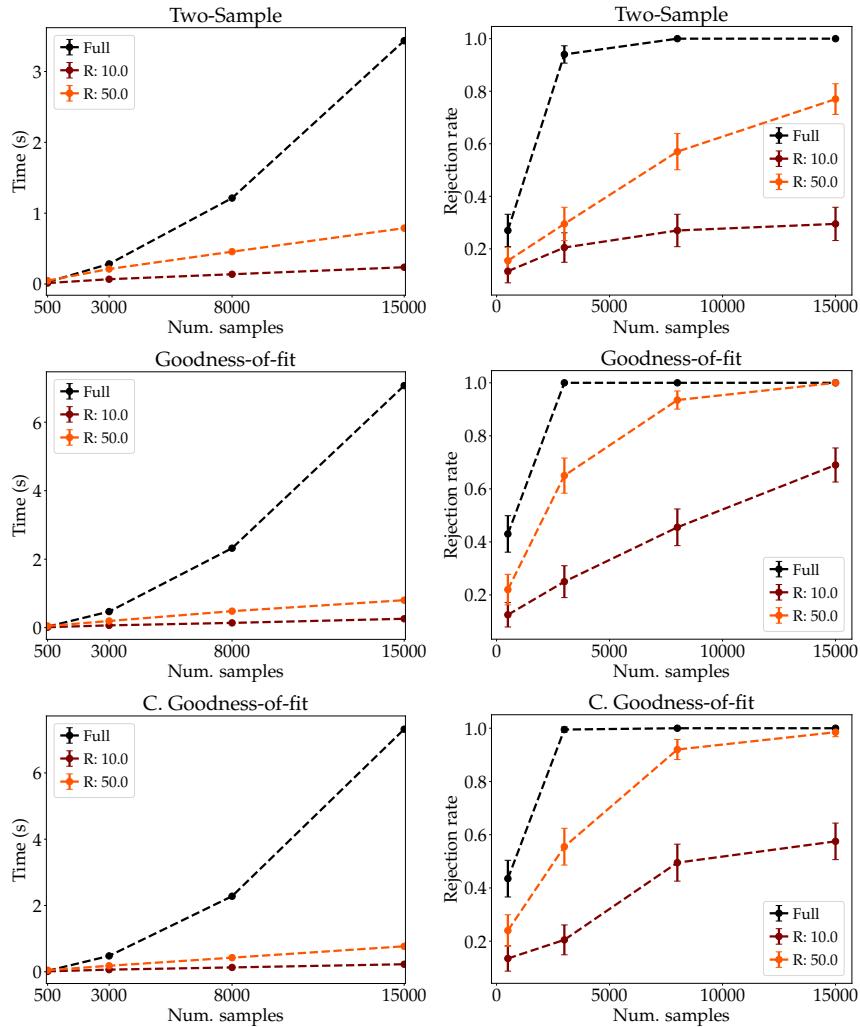


Figure VIII.B.6: Runtime and rejection rate of hypothesis tests using the quadratic and linear time U-statistic variants of the MMD (top), KSD (middle) and KCSD (bottom) statistics, 200 seeds, and $\delta = 0.1$.

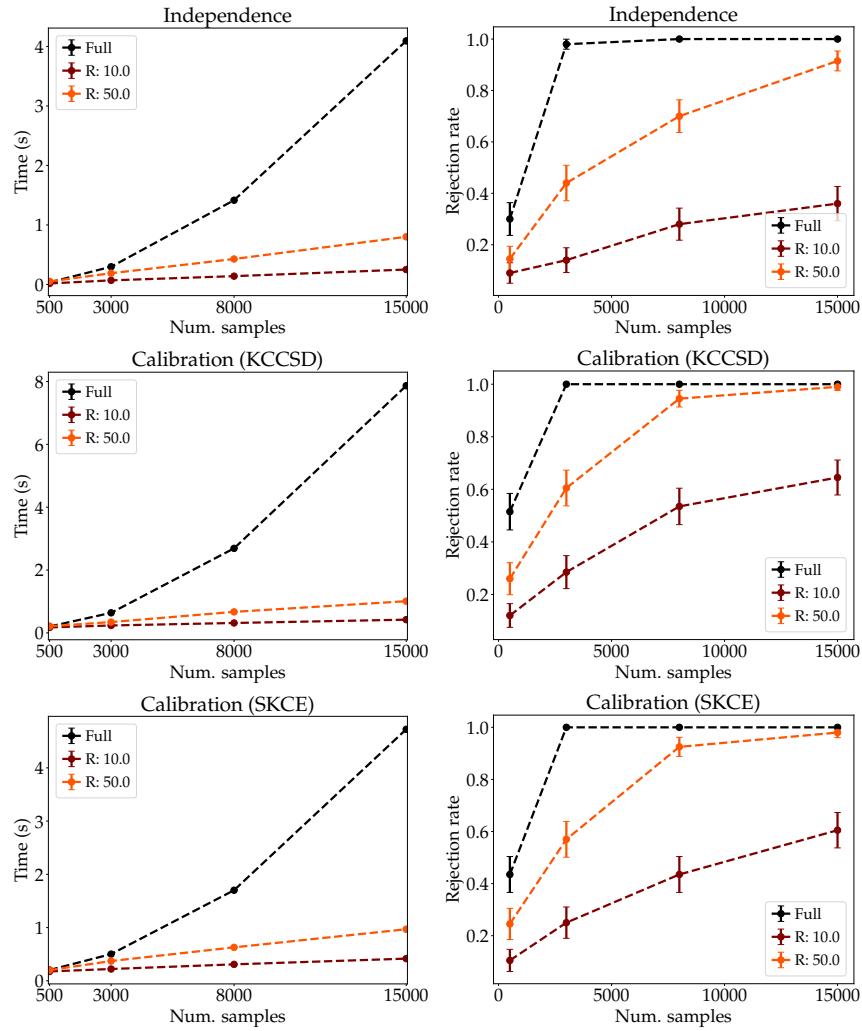


Figure VIII.B.7: Runtime and rejection rate of hypothesis tests using the quadratic and linear time U-statistic variants of the KCCSD (top), and SKCE statistics (bottom), 200 seeds, and $\delta = 0.1$.

VIII.B.5 Caching Intermediate Quantities

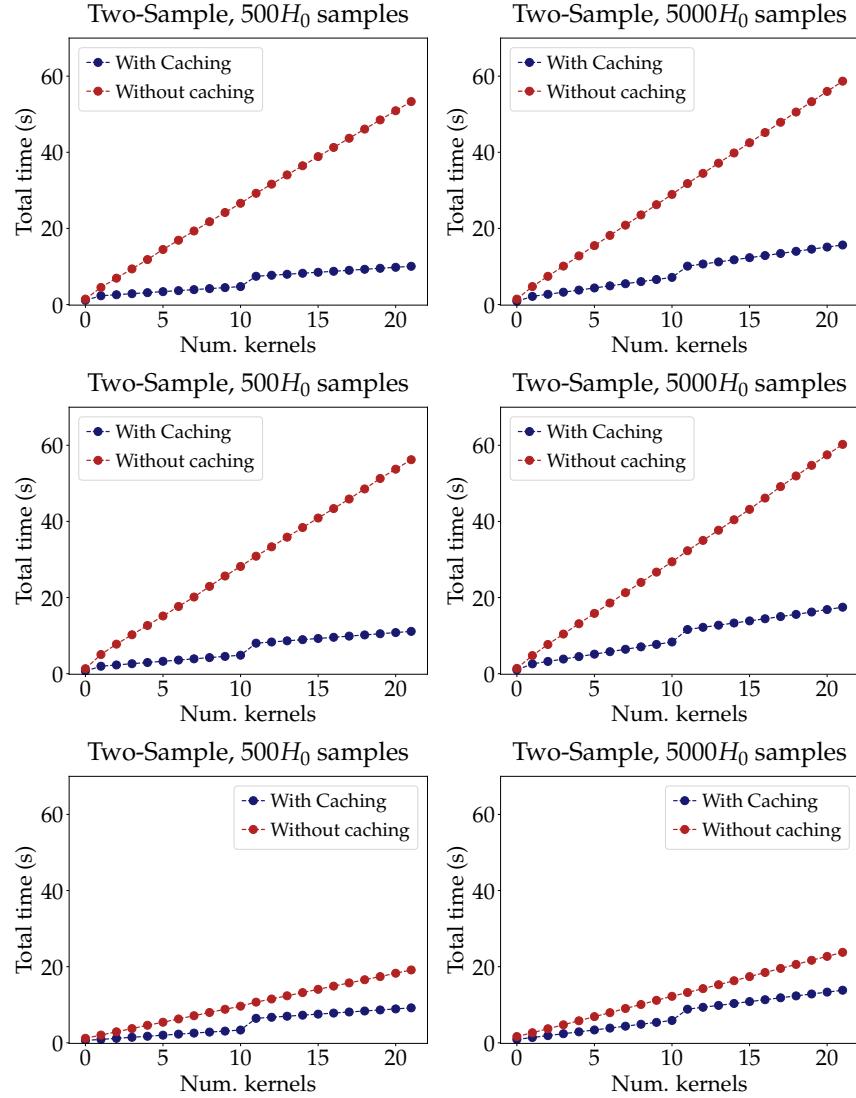


Figure VIII.B.8: Cumulative runtime of a composite MMD hypothesis test with the Bonferroni (top), Aggregation [220] (middle), and FUSE [22] aggregation method as a function of the number of statistics computed, with and without caching of intermediate quantities (single seed).

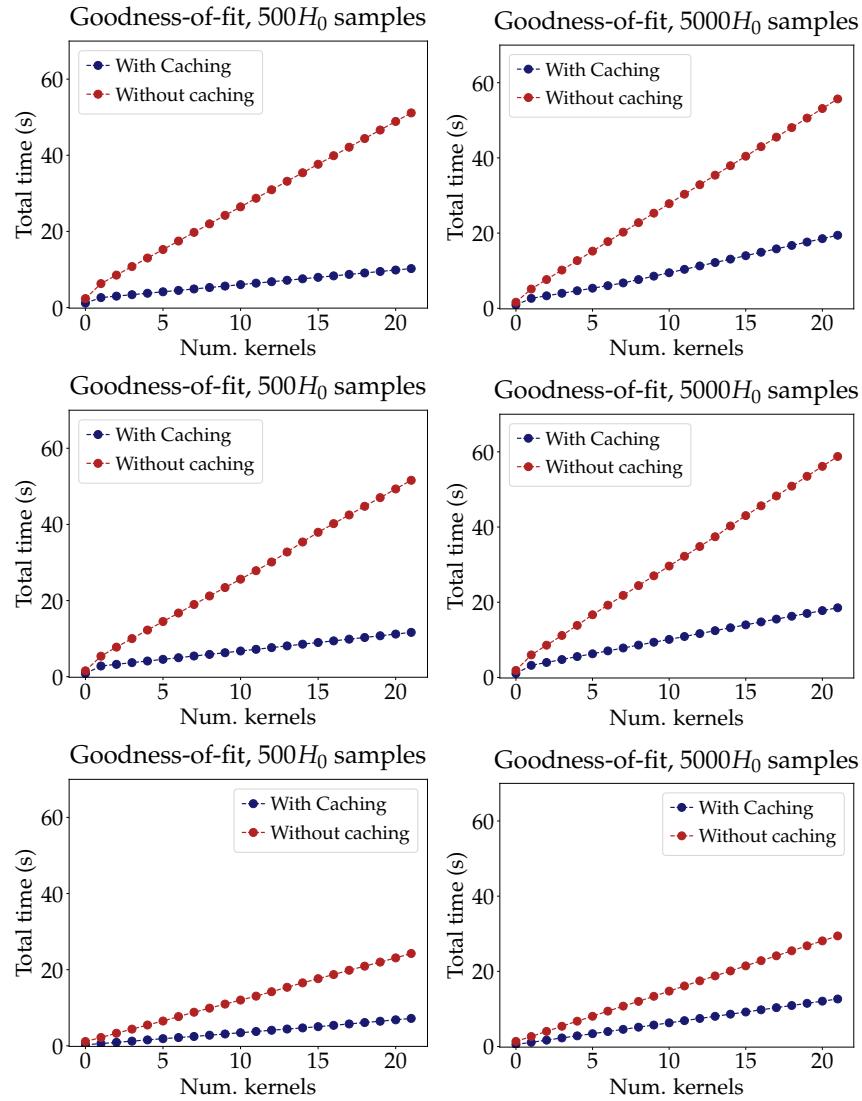


Figure VIII.B.9: Cumulative runtime of a composite KSD hypothesis test with the Bonferroni (top), Aggregation [218] (middle), and FUSE [22] (bottom) aggregation method as a function of the number of statistics computed, with and without caching of intermediate quantities (single seed).

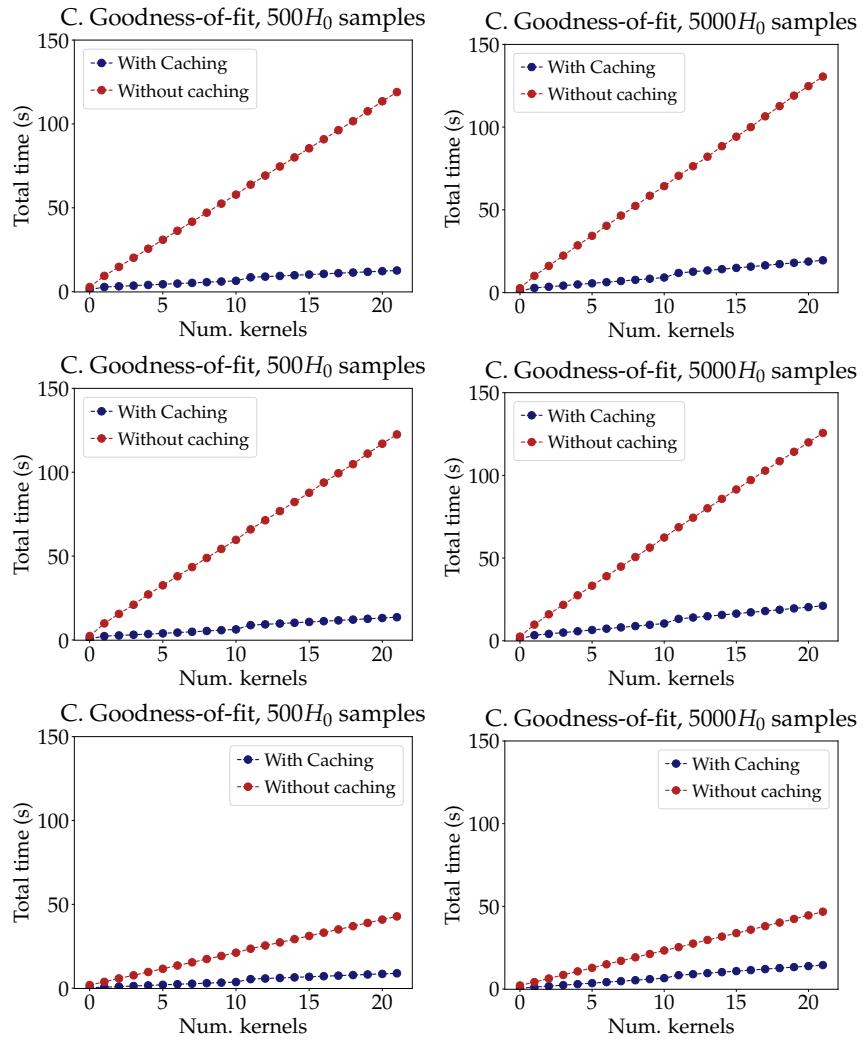


Figure VIII.B.10: Cumulative runtime of a composite KCSd hypothesis test with the Bonferroni (top), Aggregation (middle), and FUSE [22] aggregation method as a function of the number of statistics computed, with and without caching of intermediate quantities (single seed).

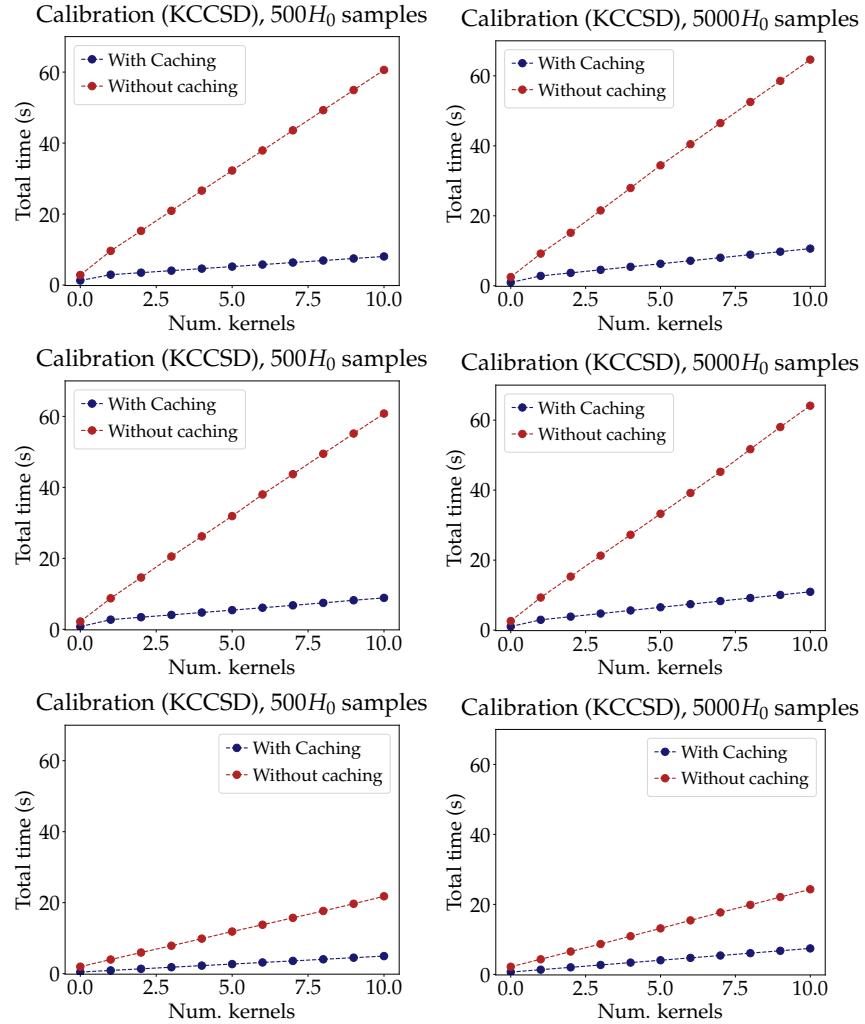


Figure VIII.B.11: Cumulative runtime of a composite KCCSD hypothesis test with the Bonferroni (top), Aggregation (middle), and FUSE [22] aggregation method as a function of the number of statistics computed, with and without caching of intermediate quantities (single seed).

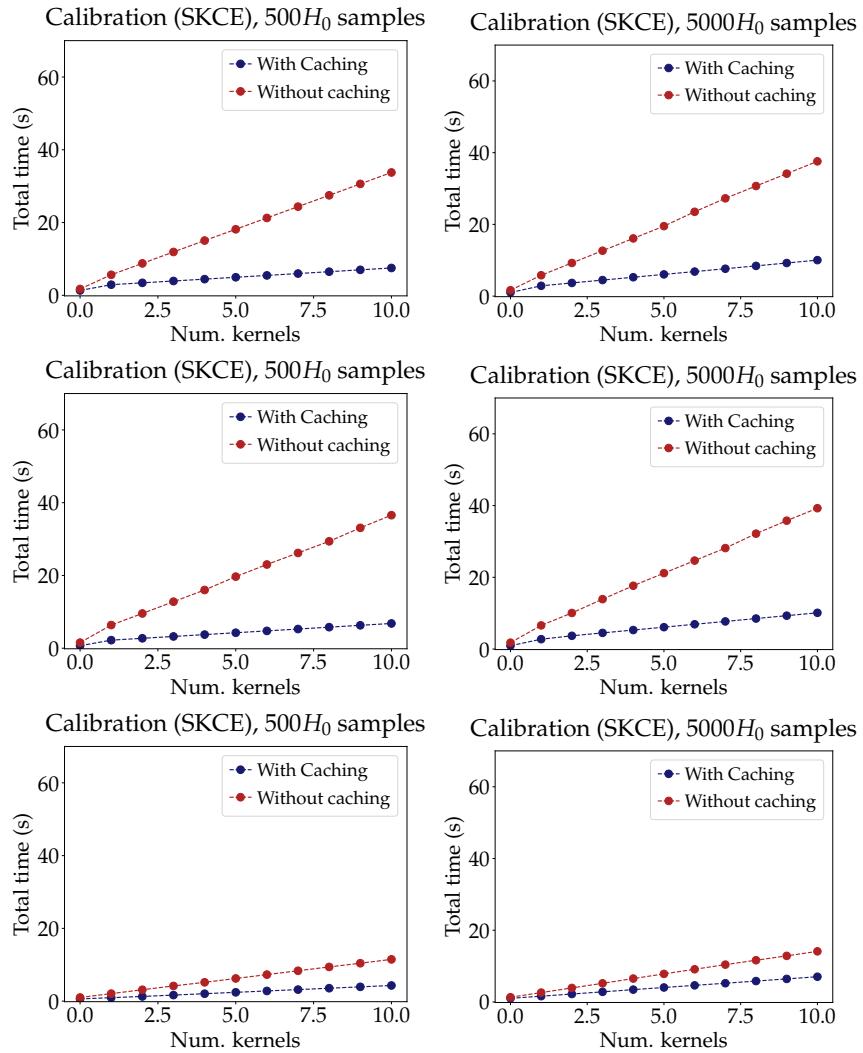


Figure VIII.B.12: Cumulative runtime of a composite SKCE hypothesis test with the Bonferroni (top), Aggregation (middle), and FUSE [22] (bottom) aggregation method as a function of the number of statistics computed, with and without caching of intermediate quantities (single seed).

Conclusion and Future Directions

Conditional Density Models are key components of AI for science. Despite the field’s high activity , I believe that a lot remains to be done to train and evaluate them efficiently.

Finite-sample bounds for noise-contrastive estimators in SBI contexts The statistical analysis of NRE presented in this thesis is asymptotic. Asymptotic analyses are common in the analysis of machine learning techniques [92, 37, 151, 136, 247] as they are relatively simple to obtain, often allow for both lower and upper bounds, and often align well with finite-sample behavior, they target a quantity (the limit of the rescaled mean-squared error) that is a priori meaningless in practice. Such analyses should thus be taken with caution, as they do not account for certain finite-sample phenomena; More satisfying results would consist in obtaining finite-sample bounds, e.g. bounds that hold for all settings of d, n . To achieve this, one could either rely on general finite sample analyses of M-estimators [185, 229], or on more specific analyses of classification [36, 139, 100]. Moreover, a very interesting research avenue would be to extend the analysis to Neural Posterior Estimation (NPE), which learns its posterior using a contrastive multi-class loss, and is widely used in practice.

Designing evaluation methods with provable power guarantees In my PhD, I developed metrics which are, in theory, able to detect any inaccuracy on unreliability

in probabilistic predictive models. In practice, these metrics must be estimated from a finite amount of data, and the resulting estimator will be noisy. To be able to make high-confidence statements regarding the accuracy of the model in the finite sample regime, I developed hypothesis tests, which are able to factor out the estimation noise by comparing the estimator values to the ones obtained by comparing the model and *its own samples*. Prior work suggests that this approach may yield tests able to detect with high probability any discrepancy between the model of the data. In the future, I plan to formally establish this by providing power guarantees for these tests. Of particular interest will be the case of calibration, in which the conditioned variable is a probability distribution, and not a simple finite-dimensional vector. This kind of input is non-standard in the theory of hypothesis testing. Different approaches may be considered to establish these guarantees, depending on whether these probability distributions are parametric or not.

Low-variance evaluation metrics for biological sequence models In [248], I developed accuracy and calibration metrics for predictive models of biological sequences. Currently, estimating these metrics requires sampling from the predictive models; however, the variance of the estimator depends on how many samples are drawn from the model; thus, obtaining low-variance estimators can be computationally prohibitive, which hinders the practical value of these metrics. In the future, I plan to develop sample-free estimators for these metrics, which will be able to provide low-variance estimates of the calibration and accuracy of the models without the need to sample from them. To do this, I plan to use Stein operators [11], a celebrated tool to construct sample-free estimators. Traditionally employed to model continuous data, these operators were adapted to the case of discrete data in [8]. Extending such operators to the conditional setting will allow the creation of sample-free estimators for the calibration and accuracy of models in computational biology.

Tradeoffs in conditional goodness-of-fit Accuracy and calibration can be shown to be two instances of a more general property, called conditional goondess-of-fit [128]. Currently, conditional goodness-of-fit metrics (including the ones I developed) proceed by augmenting conditional distributions with a marginal distribution, and

comparing the resulting joint distributions. This method has clear upsides: first, it is a valid approach to comparing two conditional distributions, in the sense that the resulting metrics will be able to detect any discrepancy between them. Second, the resulting estimators are low-variance, and can be integrated within hypothesis tests. Despite these benefits, this augmentation step may not be necessary: instead, it may be possible to design metrics which directly compare conditional distributions at each input [192]. These metrics are likely to be (1) stronger, in the sense that they will be able to detect more subtle discrepancies between the models, and (2) more interpretable, as they will be able to pinpoint input regions of the input space where the models disagree. However, current methods which directly compare conditional distributions must estimate *conditional mean embeddings* [170], which are high variance and can thus be less informative when only finitely many data points are available to evaluate the models. Understanding the tradeoff between discriminative power and variance will require a fine analysis, but has the potential to yield metrics for the evaluation of probabilistic models which will be provably more powerful.

Towards problem-specific methods for uncertainty quantification The mainstream definitions of accuracy and calibration used in my PhD work, are very general. This generality allows designing evaluation methods applicable to a wide range of problems. However, in practice, the definition of accuracy and calibration may vary depending on the problem at hand. For instance, in the case of protein design, specific parts of a protein (such as complementarity determining regions for antibodies) will be more important than others to predict their function, and even a small mismatch in such regions may lead to a complete loss of function [213]. In SBI on the other hand, overconfidence may be more problematic, and thus more important to detect, than underconfidence [104]. However, metrics relying on a general definition of calibration or accuracy may not be able to capture these specificities. In the future, I believe that designing more fine-grained measures of accuracy and calibration—which could be used both for training and evaluation—will constitute a key step to maximize their usefulness and adoption in practice.

Bibliography

- [1] Sur les contraintes imposées par les passages la limite usuels en statistique.
In *Bulletin of the International Statistical Institute*.
- [2] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar.
Information-theoretic lower bounds on the oracle complexity of convex
optimization. *Advances in Neural Information Processing Systems*, 22, 2009.
- [3] Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar
Meynaoui. Adaptive test of independence based on hsic measures. *The
Annals of Statistics*, 50(2):858–879, 2022.
- [4] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of
divergence of one distribution from another. *Journal of the Royal Statistical
Society: Series B (Methodological)*, 28(1):131–142, 1966.
- [5] Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy monte
carlo: Convergence of markov chains with approximate transition kernels.
Statistics and Computing, 2016.
- [6] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in
metric spaces and in the space of probability measures*. Springer Science &
Business Media, 2005.

- [7] Alan Amin, Eli N Weinstein, and Debora Marks. A generative nonparametric bayesian model for whole genomes. *NeurIPS*, 34:27798–27812, 2021.
- [8] Alan Nawzad Amin, Eli N Weinstein, and Debora Susan Marks. A kernelized stein discrepancy for biological sequences. In *International Conference on Machine Learning*, pages 718–767. PMLR, 2023.
- [9] Alan Nawzad Amin, Eli Nathan Weinstein, and Debora Susan Marks. Biological sequence kernels with guaranteed flexibility. *arXiv preprint arXiv:2304.03775*, 2023.
- [10] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein’s method meets computational statistics: a review of some recent developments. *Statistical Science*, 2022.
- [11] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein’s method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.
- [12] Ingo Steinwart Andreas Christmann. *Support Vector Machines*. Springer New York, 2008.
- [13] Miguel A Arcones and Evarist Giné. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992.
- [14] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- [15] Yves F Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

- [16] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.
- [17] Jerome Baum, Heishiro Kanagawa, and Arthur Gretton. A kernel stein test of goodness of fit for sequential models. In *ICML*, 2023.
- [18] Mark A Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 2010.
- [19] Nathaniel R Bennett, Brian Coventry, Inna Goreshnik, Buwei Huang, Aza Allen, Dionne Vafeados, Ying Po Peng, Justas Dauparas, Minkyung Baek, Lance Stewart, Frank DiMaio, Steven De Munck, Savvas N Savvides, and David Baker. Improving de novo protein binder design with deep learning. *Nat. Commun.*, 2023.
- [20] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- [21] Dimitri Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.
- [22] Felix Biggs, Antonin Schrab, and Arthur Gretton. MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, 2023.
Python package mmdfuse:
<https://github.com/antoninschrab/mmdfuse>.
- [23] Gunnar Blom. Some properties of incomplete u-statistics. *Biometrika*, pages 573–580, 1976.
- [24] Michael GB Blum and Olivier François. Non-linear regression models for approximate bayesian computation. *Statistics and computing*, 20:63–73, 2010.

- [25] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.
- [26] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vis.*, 51(1): 22–45, 2015.
- [27] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.
- [28] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL
<http://github.com/jax-ml/jax>.
- [29] J. Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 2009.
- [30] J. Bröcker and L. A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 2007.
- [31] Jochen Bröcker. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 2008.
- [32] Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.
- [33] Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

- [34] José A Carrillo, Robert J McCann, and Cédric Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Revista Matematica Iberoamericana*, 2003.
- [35] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.
- [36] Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression. *arXiv preprint arXiv:2411.02137*, 2024.
- [37] Omar Chehab, Alexandre Gramfort, and Aapo Hyvärinen. The optimal noise in noise-contrastive learning is not what you think. In *Uncertainty in Artificial Intelligence*, pages 307–316. PMLR, 2022.
- [38] Zonghao Chen, Aratrika Mustafi, **Pierre Glaser**, Anna Korba, Arthur Gretton, and Bharath K Sriperumbudur. (De)-regularized maximum mean discrepancy gradient flow. Accepted (with minor revisions) at JMLR. Accessible at <https://arxiv.org/pdf/2409.14980.pdf>, 2024.
- [39] Andreas Christmann and Ingo Steinwart. Universal kernels on Non-Standard input spaces. In *NeurIPS*, 2010.
- [40] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. *Advances in neural information processing systems*, 23, 2010.
- [41] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, 2016.
- [42] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016.
- [43] Erhan Çinlar. *Probability and stochastics*. Springer, 2011.

- [44] D. R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4):562, December 1958.
- [45] Harald Cramér. *Mathematical methods of statistics*. Princeton university press, 1946.
- [46] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 2020.
- [47] 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.
- [48] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017.
- [49] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *ICASSP*, 2007.
- [50] Hugh Dance, **Pierre Glaser**, Peter Orbanz, and Ryan Adams. Efficiently vectorized MCMC on modern accelerators. In *Forty-second International Conference on Machine Learning*, 2025.
- [51] J Dauparas, I Anishchenko, N Bennett, H Bai, R J Ragotte, L F Milles, B I M Wicky, A Courbet, R J de Haas, N Bethel, P J Y Leung, T F Huddy, S Pellock, D Tischer, F Chan, B Koepnick, H Nguyen, A Kang, B Sankaran, A K Bera, N P King, and D Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 2022.
- [52] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 1983.
- [53] Michael Deistler, Pedro J Goncalves, and Jakob H Macke. Truncated proposals for scalable and hassle-free simulation-based inference. *Advances in Neural Information Processing Systems*, 35:23135–23149, 2022.

- [54] 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.
- [55] Giulia Denevi, Massimiliano Pontil, and Carlo Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. In *NeurIPS*, 2020.
- [56] S. W. Dharmadhikari, V. Fabian, and K. Jogdeo. Bounds on the moments of martingales. *Ann. Math. Statist.*, pages 1719–1723, 1968.
- [57] Nicolae Dinculeanu. *Vector integration and stochastic integration in Banach spaces*. John Wiley & Sons, 2000.
- [58] Thinh T Doan. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2022.
- [59] Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress then test: Powerful kernel testing in near-linear time. *AISTATS*, 2023.
- [60] Carles Domingo-Enrich, Raaz Dwivedi, and Lester Mackey. Compress Then Test: Powerful Kernel Testing in Near-linear Time. *International Conference on Artificial Intelligence and Statistics, AISTATS, PMLR*, pages 1174–1218, 2023. Python package goodpoints:
<https://github.com/microsoft/goodpoints>.
- [61] Joseph L Doob. *Measure theory*, volume 143. Springer Science & Business Media, 2012.
- [62] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32, 2019.
- [63] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy- based models. *arXiv preprint arXiv:2021.01316*, 2021.

- [64] Ayoub El Hanchi, Chris J Maddison, and Murat A Erdogdu. On the efficiency of erm in feature learning. *Advances in Neural Information Processing Systems*, 37:98596–98624, 2024.
- [65] Andre Elisseeff, Theodoros Evgeniou, Massimiliano Pontil, and Leslie Pack Kaelbling. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(1), 2005.
- [66] Richard G Everitt. Bayesian parameter estimation for latent markov random fields and social networks. *Journal of Computational and graphical Statistics*, 21(4):940–960, 2012.
- [67] M. Fasiolo, S. N. Wood, M. Zaffran, R. Nedellec, and Y. Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 2020.
- [68] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.
- [69] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press, 1925.
- [70] Roman Frigg and S Hartmann. Models in science. *Stanford Encyclopedia of Philosophy*, 2006.
- [71] Roy Frostig, Matthew James Johnson, and Chris Leary. Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, 2018.
- [72] Zhangyang Gao, Cheng Tan, Pablo Chacón, and Stan Z Li. PiFold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

- [73] Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- [74] Tom George, **Pierre Glaser**, Kim Stachenfeld, Caswell Barry, and Claudia Clopath. Simpl: Scalable and hassle-free optimisation of neural representations from behaviour. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [75] Charles J Geyer. Introduction to markov chain monte carlo. *Handbook of markov chain monte carlo*, 20116022:45, 2011.
- [76] Daniel Gilman, Simon Birrer, Tommaso Treu, Charles R Keeton, and Anna Nierenberg. Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses. *Monthly Notices of the Royal Astronomical Society*, 2018.
- [77] Pierre Glaser, David Widmann, Fredrik Lindsten, and Arthur Gretton. Fast and Scalable Score-based Kernel Calibration Tests. *Uncertainty in Artificial Intelligence, UAI, PMLR*, 216:691–700, 2023. Python package kccsd: <https://github.com/pierreglaser/kccsd>.
- [78] Manuel Glöckler, Michael Deistler, and Jakob H Macke. Variational methods for simulation-based inference. In *International Conference on Learning Representations*, 2021.
- [79] Manuel Glöckler, Michael Deistler, and Jakob H Macke. Variational methods for simulation-based inference. *arXiv preprint arXiv:2203.04176*, 2022.
- [80] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 2007.
- [81] Antoine Godichon-Baggioni, Nicklas Werge, and Olivier Wintenberger. Non-asymptotic analysis of stochastic approximation algorithms for streaming data. *ESAIM: Probability and Statistics*, 27:482–514, 2023.

- [82] Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, et al. Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife*, 9:e56261, 2020.
- [83] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *ICML*, 2017.
- [84] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, 2017.
- [85] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, and Lester Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 2019.
- [86] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *ICML*, 2020.
- [87] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2019.
- [88] David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *International Conference on Machine Learning*, 2019.
- [89] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.
- [90] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 2012.

- [91] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.
- [92] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.
- [93] Hyukjun Gweon. A power-controlled reliability assessment for multi-class probabilistic classifiers. *Advances in Data Analysis and Classification*, 2022.
- [94] Sara Ann Haddad and Eve Marder. Recordings from the *c. borealis* stomatogastric nervous system at different temperatures in the decentralized condition. URL <https://doi.org/10.5281/zenodo>, July 2021.
- [95] Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.
- [96] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.
- [97] Gary Harris and Clyde Martin. Shorter notes: The roots of a polynomial vary continuously as a function of the coefficients. *Proceedings of the American Mathematical Society*, pages 390–392, 1987.
- [98] Ian R Harris. Predictive fit for natural exponential families. *Biometrika*, pages 675–684, 1989.
- [99] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. *2.4 Statistical Decision Theory*, page 18. Springer, 2 edition, 2009.
- [100] Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209. PMLR, 2014.

- [101] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free MCMC with amortized approximate ratio estimators. In *International Conference on Machine Learning*, 2020.
- [102] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting a crisis in simulation-based inference, 2021.
- [103] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, and Gilles Louppe. Averting A Crisis In Simulation-Based Inference. *arXiv:2110.06581 [cs, stat]*, October 2021.
- [104] Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=LHAbHkt6Aq>.
- [105] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.
- [106] Y. H. S. Ho and S. M. S. Lee. Calibrated interpolated confidence intervals for population quantiles. *Biometrika*, 2005.
- [107] Joseph Lawson Hodges Jr.
- [108] Wassily Hoeffding. The strong law of large numbers for u-statistics. *Institute of Statistics mimeo series*, 1961.
- [109] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Breakthroughs in statistics: Foundations and basic theory*, pages 308–334, 1992.
- [110] Wassily Hoeffding. *On sequences of sums of independent random vectors*. Springer, 1994.

- [111] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022.
- [112] Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2418–2437. PMLR, 2024.
- [113] Tim (<https://stats.stackexchange.com/users/35989/tim>). Can prediction and inference be used interchangeably? Cross Validated. URL <https://stats.stackexchange.com/q/558834>. URL: <https://stats.stackexchange.com/q/558834> (version: 2021-12-31).
- [114] Peter Flom (<https://stats.stackexchange.com/users/686/peter.flom>). Inference vs prediction terminology. Cross Validated. URL <https://stats.stackexchange.com/q/632385>. URL: <https://stats.stackexchange.com/q/632385> (version: 2023-11-27).
- [115] Peter J Huber. Robust statistics. Wiley, 1981.
- [116] Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 1996.
- [117] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005.
- [118] Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.
- [119] John Ingraham, Adam Riesselman, Chris Sander, and Debora Marks. Learning protein structure with a differentiable simulator. In *International Conference on Learning Representations*, 2018.
- [120] John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *NeurIPS*, 2019.

- [121] Bai Jiang, Tung-Yu Wu, Yifan Jin, and Wing H Wong. Convergence of contrastive divergence algorithm in exponential family. *The Annals of Statistics*, 46(6A):3067–3098, 2018.
- [122] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-Antigen docking and design via hierarchical equivariant refinement. *ICML*, 2022.
- [123] Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. Interpretable Distribution Features with Maximum Testing Power. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, pages 181–189, 2016. Python package `interpretable-test`:
<https://github.com/wittawatj/interpretable-test>.
- [124] Wittawat Jitkrittum, Zoltán Szabó, and Arthur Gretton. An Adaptive Test of Independence with Analytic Kernel Embeddings. *International Conference on Machine Learning, ICLR, PMLR*, 2017. Python package `fsic-test`:
<https://github.com/wittawatj/fsic-test>.
- [125] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A Linear-Time Kernel Goodness-of-Fit Test. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, pages 262–271, 2017. Python package `kernel-gof`:
<https://github.com/wittawatj/kernel-gof>.
- [126] Wittawat Jitkrittum, Heishiro Kanagawa, Patsorn Sangkloy, James Hays, Bernhard Schölkopf, and Arthur Gretton. Informative Features for Model Comparison. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, pages 816–827, 2018. Python package `kernel-mod`:
<https://github.com/wittawatj/kernel-mod>.
- [127] Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing Goodness of Fit of Conditional Density Models with Kernels. *Uncertainty in Artificial Intelligence, UAI, PMLR*, 124, 2020. Python package `kernel-cgof`: <https://github.com/wittawatj/kernel-cgof>.

- [128] Wittawat Jitkrittum, Heishiro Kanagawa, and Bernhard Schölkopf. Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, 2020.
- [129] Oliver Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.
- [130] Sean R Johnson, Xiaozhi Fu, Sandra Viknander, Clara Goldin, Sarah Monaco, Aleksej Zelezniak, and Kevin K Yang. Computational scoring and experimental evaluation of enzymes generated by neural networks. 2023.
- [131] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Learning in graphical models*, pages 105–161, 1998.
- [132] Olav Kallenberg and Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.
- [133] Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme. In *Conference on Learning Theory*, pages 1944–1974. PMLR, 2019.
- [134] Ilmun Kim and Antonin Schrab. Differentially Private Permutation Tests: Applications to Kernel Methods. *arXiv preprint 2310.19043*, 2023. Python package dpkernel: <https://github.com/antoninschrab/dpkernel>.
- [135] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022.
- [136] Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2022.

- [137] Deqian Kong, Bo Pang, Tian Han, and Ying Nian Wu. Molecule design by latent space energy-based modeling and gradual distribution shifting. In *Uncertainty in Artificial Intelligence*, pages 1109–1120. PMLR, 2023.
- [138] Nils M Kriege, Fredrik D Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 2020.
- [139] Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *Sankhya A*, pages 1–70, 2024.
- [140] Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, 2017.
- [141] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In *NeurIPS*, pages 12316–12326, 2019.
- [142] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *ICML*, 2018.
- [143] Vered Kunik and Yanay Ofran. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering, Design & Selection*, 2013.
- [144] Kit Fun Lau and Ken A Dill. Theory for protein mutability and biogenesis. *Proceedings of the National Academy of Sciences*, 87(2):638–642, 1990.
- [145] Lucien Le Cam et al. Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 245–261. University of California Press, 1972.
- [146] Nicolas Le Roux and Yoshua Bengio. Representational power of restricted boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

- [147] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006.
- [148] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- [149] A J Lee. *U-statistics: Theory and Practice*. Routledge, 2019.
- [150] Donghwan Lee, Xinxin Huang, Hamed Hassani, and Edgar Dobriban. T-Cal: An optimal test for the calibration of predictive models, 2022.
- [151] Holden Lee, Chirag Pabbaraju, Anish Prasad Sevekari, and Andrej Risteski. Pitfalls of gaussians as a noise distribution in nce. In *The Eleventh International Conference on Learning Representations*, 2023.
- [152] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- [153] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.
- [154] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.
- [155] Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 2020.
- [156] Qiang Liu. A short introduction to kernelized Stein discrepancy, 2016.
- [157] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

- [158] James Robert Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. *Adv. Neural Inf. Process. Syst.*, 2015-Janua: 829–837, 2015.
- [159] Alfred J Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 1920.
- [160] J.-M. Lueckmann, J. Boelts, D. S. Greenberg, P. J. Gonçalves, and J. H. Macke. Benchmarking simulation-based inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.
- [161] 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 *Advances in Neural Information Processing Systems*, 2017.
- [162] Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.
- [163] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv preprint arXiv:1809.01812*, 2018.
- [164] Peter McCullagh. *Tensor methods in statistics: Monographs on statistics and applied probability*. Chapman and Hall/CRC, 2018.
- [165] Dimitri Meunier, Massimiliano Pontil, and Carlo Ciliberto. Distribution regression with sliced Wasserstein kernels. In *ICML*, 2022.
- [166] Dimitri Meunier, Massimiliano Pontil, and Carlo Ciliberto. Distribution regression with sliced Wasserstein kernels. In *ICML*, 2022.
- [167] Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 2005.

- [168] Jesper Möller, Anthony N Pettitt, Robert Reeves, and Kasper K Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 2006.
- [169] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.
- [170] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- [171] Nicole Mücke, Gergely Neu, and Lorenzo Rosasco. Beating sgd saturation with tail-averaging and minibatching. *Advances in Neural Information Processing Systems*, 32, 2019.
- [172] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.
- [173] A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 1977.
- [174] Iain Murray and Zoubin Ghahramani. Bayesian learning in undirected graphical models: approximate mcmc algorithms. *arXiv preprint arXiv:1207.4134*, 2012.
- [175] Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. Mcmc for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006.
- [176] M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

- [177] Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.
- [178] Radford M Neal. *Probabilistic inference using Markov chain Monte Carlo methods*. Department of Computer Science, University of Toronto Toronto, ON, Canada, 1993.
- [179] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
- [180] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.
- [181] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.
- [182] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. *Advances in Neural Information Processing Systems*, 2019.
- [183] Erik Nijkamp, Mitch Hill, Tian Han, Song-Chun Zhu, and Ying Nian Wu. On the anatomy of mcmc-based maximum likelihood learning of energy-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [184] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 2019.
- [185] Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance. 2021.

- [186] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 2000.
- [187] Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36, 2024.
- [188] Lorenzo Pacchiardi and Ritabrata Dutta. Score matched neural exponential families for likelihood-free inference. *Journal of Machine Learning Research*, 2022.
- [189] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.
- [190] 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*, 2019.
- [191] Jaewoo Park and Murali Haran. Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390, 2018.
- [192] Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. *Advances in neural information processing systems*, 33:21247–21259, 2020.
- [193] Cynthia Passmore, Julia Svoboda Gouvea, and Ronald Giere. Models in science and in learning science: Focusing scientific practice on sense-making. In *International handbook of research in history, philosophy and science teaching*, pages 1171–1202. Springer, 2013.
- [194] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through

- multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.
- [195] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.*, 10, 06 2000.
- [196] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- [197] Astrid A Prinz, Cyrus P Billimoria, and Eve Marder. Alternative to hand-tuning conductance-based models: construction and analysis of databases of model neurons. *Journal of neurophysiology*, 2003.
- [198] Astrid A Prinz, Dirk Bucher, and Eve Marder. Similar network activity from disparate circuit parameters. *Nature neuroscience*, 2004.
- [199] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems*, 35:9538–9551, 2022.
- [200] Maxim Rabinovich, Aaditya Ramdas, Michael I Jordan, and Martin J Wainwright. Function-specific mixing times and concentration away from equilibrium. 2020.
- [201] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [202] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

- [203] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a non-asymptotic analysis. In *Conference on Learning Theory*, 2017.
- [204] Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.
- [205] C Radhakrishna Rao et al. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc*, 37(3):81–91, 1945.
- [206] Franz Rellich. *Perturbation theory of eigenvalue problems*. CRC Press, 1969.
- [207] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [208] Leonard CG Rogers and David Williams. *Diffusions, Markov processes, and martingales: Volume 1, foundations*. Cambridge University Press, 2000.
- [209] Arturo Rosenblueth and Norbert Wiener. The role of models in science. *Philosophy of science*, 12(4):316–321, 1945.
- [210] Walter Rudin. *Real and complex analysis, 3rd ed.* McGraw-Hill, Inc., USA, 1987. ISBN 0070542341.
- [211] Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.
- [212] M. Rueda, S. Martinez-Puertas, H. Martinez-Puertas, and A. Arcos. Calibration methods for estimating quantiles. *Metrika*, 2006.
- [213] Jeffrey A Ruffolo, Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J Gray. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature communications*, 14(1):2389, 2023.

- [214] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. Protein homology detection using string alignment kernels. *Bioinformatics*, 2004.
- [215] Antoine Salmona, Valentin De Bortoli, Julie Delon, and Agnès Desolneux. Can push-forward generative models fit multimodal distributions? *Advances in Neural Information Processing Systems*, 35:10766–10779, 2022.
- [216] Antonin Schrab and Ilmun Kim. Robust Kernel Hypothesis Testing under Data Corruption. *arXiv preprint 2405.19912*, 2024. Python package `dckernel`: <https://github.com/antoninschrab/dckernel>.
- [217] Antonin Schrab, Benjamin Guedj, and Arthur Gretton. Efficient Aggregated Kernel Tests using Incomplete U-statistics. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, 2022. Python package `agginc`: <https://github.com/antoninschrab/agginc>.
- [218] Antonin Schrab, Benjamin Guedj, and Arthur Gretton. KSD Aggregated Goodness-of-fit Test. *Advances in Neural Information Processing Systems, NeurIPS, PMLR*, 2022. Python package `ksdagg`: <https://github.com/antoninschrab/ksdagg>.
- [219] Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete U -statistics. *arXiv preprint arXiv:2206.09194*, 2022.
- [220] Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research*, 24(194):1–81, 2023. Python package `mmdagg`: <https://github.com/antoninschrab/mmdagg>.
- [221] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

- [222] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [223] David Shortle, Hue Sun Chan, and Ken A Dill. Modeling the effects of mutations on the denatured states of proteins. *Protein Science*, 1(2):201–215, 1992.
- [224] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla S M Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutařová Varekova, Radka Svobodova, Jon Lees, and Christine A Orengo. CATH: increased structural coverage of functional space. *Nucleic acids research*, 2021.
- [225] Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *ICML*, 2019.
- [226] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.
- [227] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.
- [228] Yang Song and Diederik P Kingma. How to train your energy-based models. *arXiv preprint arXiv:2101.03288*, 2021.
- [229] Vladimir Spokoiny. Parametric estimation. finite sample theory. 2012.
- [230] Bharath Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. On the relation between universality, characteristic kernels and rkhs embedding of measures. In *AISTATS*, 2010.

- [231] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.*, 2017.
- [232] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.
- [233] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.
- [234] Kiera H Sumida, Reyes Núñez-Franco, Indrek Kalvet, Samuel J Pellock, Basile I M Wicky, Lukas F Milles, Justas Dauparas, Jue Wang, Yakov Kipnis, Noel Jameson, Alex Kang, Joshmyn De La Cruz, Banumathi Sankaran, Asim K Bera, Gonzalo Jiménez-Osés, and David Baker. Improving protein expression, stability, and function with ProteinMPNN. *Journal of the American Chemical Society*, 2024.
- [235] Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.
- [236] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath K. Sriperumbudur. Two-stage sampled learning theory on distributions. In *AISTATS*, 2015.
- [237] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath K. Sriperumbudur. Two-stage sampled learning theory on distributions. In *AISTATS*, 2015.
- [238] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *J. Mach. Learn. Res.*, 2016.
- [239] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *JMLR*, 2016.

- [240] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 2013.
- [241] M. Taillardat, O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 2016.
- [242] Terence Tao. *An introduction to measure theory*, volume 126. American Mathematical Soc., 2011.
- [243] Michael Eugene Taylor. *Partial differential equations. 1, Basic theory*. Springer, 1996.
- [244] **Pierre Glaser** and Arthur Gretton. Statistical Analysis of Neural Ratio Estimation. In preparation, 2025.
- [245] **Pierre Glaser**, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed KL gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021.
- [246] **Pierre Glaser**, David Widmann, Fredrik Lindsten, and Arthur Gretton. Fast and scalable score-based kernel calibration tests. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2023. URL <https://proceedings.mlr.press/v216/glaser23a.html>.
- [247] **Pierre Glaser**, Kevin Han Huang, and Arthur Gretton. Near-optimality of contrastive divergence algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=Q74JVgKCP6>.
- [248] **Pierre Glaser**, Steffanie Paul, Alissa M Hummer, Charlotte Deane, Debora Susan Marks, and Alan Nawzad Amin. Kernel-based evaluation of conditional biological sequence models. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=2dlmcTXfcY>.

- [249] **Pierre Glaser**, Michael Arbel, Arnaud Doucet, and Arthur Gretton. Maximum likelihood learning of energy-based models for simulation-based inference. A previous version of this work is available at
<https://arxiv.org/abs/2210.147562025>, 2025.
- [250] **Pierre Glaser**, Antonin Schrab, and Arthur Gretton. Measuring data and model properties with `kdiscs`. In preparation, 2025.
- [251] Tijmen Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, 2008.
- [252] Graham Upton and Ian Cook. *A dictionary of statistics 3e*. Oxford university press, 2014.
- [253] user1592380 (<https://stats.stackexchange.com/users/137298/user1592380>). What is the difference between prediction and inference? Cross Validated.
URL <https://stats.stackexchange.com/q/244017>.
URL: <https://stats.stackexchange.com/q/244017> (version: 2020-06-11).
- [254] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *AISTATS*, 2019.
- [255] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- [256] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- [257] Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [258] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

- [259] Csilla Varnai, Nikolas S Burkoff, and David L Wild. Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: a maximum likelihood approach. *Journal of chemical theory and computation*, 9(12):5718–5733, 2013.
- [260] Jean-Philippe Vert, Hiroto Saigo, and Tatsuya Akutsu. Local alignment kernels for biological sequences. *Kernel methods in computational biology*, 2004.
- [261] Vladimir Vovk. Superefficiency from the vantage point of computability. 2009.
- [262] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- [263] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.
- [264] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.
- [265] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, Basile I M Wicky, Nikita Hanikel, Samuel J Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 2023.
- [266] Li K Wenliang and Heishiro Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

- [267] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *NeurIPS*, 2019.
- [268] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration Tests Beyond Classification. *International Conference on Learning Representations, ICLR, PMLR*, 2021. Python package pycalibration: <https://github.com/devmotion/pycalibration>.
- [269] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification. In *ICLR*, 2021.
- [270] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2022.
- [271] Jacob Wolfowitz. The efficiency of sequential estimates and wald's equation for sequential processes. *The Annals of Mathematical Statistics*, 18(2): 215–230, 1947.
- [272] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 2010.
- [273] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. In *International Conference on Learning Representations*, 2021.
- [274] Kaizhi Yue and Ken A Dill. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences*, 89(9): 4163–4167, 1992.
- [275] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *ICML*, 2001.

- [276] Mingtian Zhang, Oscar Key, Peter Hayes, David Barber, Brooks Paige, and François-Xavier Briol. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.
- [277] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.
- [278] Zuobai Zhang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. Enhancing protein language models with structure-based encoder and pre-training. *arXiv preprint arXiv:2303.06275*, 2023.
- [279] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 2008.
- [280] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. Risk bounds of multi-pass sgd for least squares in the interpolation regime. *Advances in Neural Information Processing Systems*, 35: 12909–12920, 2022.