Scaling Bayesian Inference: Theoretical Foundations and Practical Methods

by

Jonathan Hunter Huggins

B.A., Columbia University (2012)
S.M., Massachusetts Institute of Technology (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author .....
Department of Electrical Engineering and Computer Science
February 28, 2018

Certified by .....
Tamara Broderick
ITT Career Development Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Scaling Bayesian Inference:
Theoretical Foundations and Practical Methods

by

Jonathan Hunter Huggins

Submitted to the Department of Electrical Engineering and Computer Science
on February 28, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Bayesian statistical modeling and inference allow scientists, engineers, and companies to learn from data while incorporating prior knowledge, sharing power across experiments via hierarchical models, quantifying their uncertainty about what they have learned, and making predictions about an uncertain future. While Bayesian inference is conceptually straightforward, in practice calculating expectations with respect to the posterior can rarely be done in closed form. Hence, users of Bayesian models must turn to approximate inference methods. But modern statistical applications create many challenges: the latent parameter is often high-dimensional, the models can be complex, and there are large amounts of data that may only be available as a stream or distributed across many computers. Existing algorithm have so far remained unsatisfactory because they either (1) fail to scale to large data sets, (2) provide limited approximation quality, or (3) fail to provide guarantees on the quality of inference.

To simultaneously overcome these three possible limitations, I leverage the critical insight that in the large-scale setting, much of the data is *redundant*. Therefore, it is possible to *compress* data into a form that admits more efficient inference. I develop two approaches to compressing data for improved scalability. The first is to construct a *coreset*: a small, weighted subset of our data that is representative of the complete dataset. The second, which I call PASS-GLM, is to construct an exponential family model that approximates the original model. The data is compressed by calculating the finite-dimensional sufficient statistics of the data under the exponential family.

An advantage of the compression approach to approximate inference is that an approximate likelihood substitutes for the original likelihood. I show how such approximate likelihoods lend them themselves to *a priori* analysis and develop general tools for proving when an approximate likelihood will lead to a high-quality approximate posterior. I apply these tools to obtain *a priori* guarantees on the approximate posteriors produced by PASS-GLM. Finally, for cases when users must rely on algorithms that do not have *a priori* accuracy guarantees, I develop a method for comparing the quality of the inferences produced by competing algorithms. The method comes equipped with provable guarantees while also being computationally efficient.

Thesis Supervisor: Tamara Broderick
Title: ITT Career Development Assistant Professor of Electrical Engineering and

Computer Science

Dedicated to Oliver Marland Huggins

...and this is the wonder that's keeping the stars apart
i carry your heart(i carry it in my heart)
– E. E. Cummings

Acknowledgments

My graduate career at MIT has been a time of immense personal and intellectual growth. I am indebted to so many friends, colleagues, teachers, mentors, administrators, and classmates for making my experience at MIT fulfilling and memorable.

I must start by thanking my advisor, Tamara Broderick, who has been a superb mentor to me these past three years. Tamara's enthusiasm is boundless, as was her willingness to let me work on the problems that interested me most. It was her guidance that led to me to pursue the bulk of the work that appears in this thesis. Tamara is also a tremendous scientific communicator – I can only hope to approach her clarity of thought and ability to tell a compelling story. She has taught me so much about the under-appreciated part of being a scientist: how to convey the ideas and results I have worked so hard to produce.

I also have to extend a special thanks to Josh Tenenbaum, who served as my advisor during the first few years of my PhD. He gave me incredible flexibility to explore my interests, many of them quite theoretical. Josh always provided a stimulating perspective, forcing me to carefully evaluate my premises and consider the implications of my proposed lines of research. He guided me onto a successful research path, introducing me to Dan Roy early on in my PhD and then encouraging me to work with Tamara after she arrived at MIT.

Next, I am most grateful to my other committee members: Ryan Adams and Piotr Indyk. During his time at Harvard, Ryan welcomed me into his group, where I spent many enjoyable and illuminating afternoons. Ryan's expansive knowledge of statistics, machine learning, and beyond is an inspiration and a delight. I always learn something unexpected and new when I speak with him. I am thankful to Piotr for agreeing to serve so ably on my committee.

Research is not an individual effort, and I have been lucky enough to have a slew of fantastic collaborators. At a formative point in my graduate career, I worked with Dan Roy on the project that would become my Master's thesis. Dan's methodical approach taught me so much about how to do good theory research. James Zou's quiet brilliance continuously amazes me. Trevor Campbell has been the best lab mate and collaborator I could have asked for. I will miss our weekly research chats at Area 4. My summer at Microsoft Research working with Lester Mackey was a highlight of my PhD.

During my graduate career, I've also been lucky enough to collaborate with Ardavan Saeedi, Cynthia Rudin, Matt Johnson, Lorenzo Moesero, Will Stephenson, Miriam Shiffman, Geoffrey Schiebinger, Aviv Regev, Karthik Narasimhan, and Vikash Mansinghka. My research has also benefited from numerous conversations with, and the insights of, Jacob Andreas, Elaine Angelino, David Duvenaud, Hilary Finucane, Nicolo Fusi, Ryan Giordano, Peter Krafft, Tejas Kulkarni, Jennifer Listgarten, Rachael Meager, David Reshef, Yakir Reshef, Jacob Steinhardt, and many other members of CoCoSci, HIPS, and the Broderick lab.

I owe a huge debt of gratitude to my wonderful parents for all they have done for me over nearly three decades. Their active support of my interests in math, science, and research never wavered throughout my life. Thanks also to my fantastic

mother-in-law, Irma, for her love and support, and raising such an amazing daughter.

Finally, thanks to my incredible wife, Diana. The ways in which she has supported me and made it possible for me to complete my PhD are too numerous to list. I was unbelievably lucky to find such an amazing partner so early in life.

Contents

List of Figures

Chapter 1

Introduction

Bayesian statistical modeling and inference allow scientists, engineers, and companies to learn from data while incorporating prior knowledge, quantifying their uncertainty about what they have learned, sharing of power across experiments via hierarchical models, and making predictions about an uncertain future. Bayesian methods in particular make all these tasks conceptually straightforward, via the use of prior distributions and Bayes' theorem. The use of prior knowledge obtained from domains experts or previous inferences can improve data-efficiency, particularly when the amount of data is small relative to the complexity of the model. The uncertainty quantification provided by Bayesian inference is invaluable because, particularly in the context of decision-making, accounting for uncertainty is critical. Instead of just obtaining a point estimate for the latent parameter, one can also calculate covariances, tail probabilities, and other functionals of the posterior distribution over the latent parameter. Consider the following examples:

- A self-driving car estimates that the object moving in front of it is a small rodent scampering across the street. However, due to fog, the estimate contains substantial uncertainty. Therefore the car slows down, avoiding a collision with a creature that turns out to be a much-loved family dog.

- Using observational data, a scientist estimates the effect sizes of genes that may cause a particular type of cancer. Genes that appear to have a substantial effect will be investigated further in the lab, but investigating a gene is costly both in terms of time and money. The scientist notices that a number of the genes that have large estimated effects also have large uncertainties about the effect sizes. Instead of chasing down these noisy leads the scientist spends her time investigating the genes that have much greater certainty of affecting the cancer's growth.

- An ambulance is rushing a patient who just had a heart attack to the hospital. The medics determine they must arrive within 15 minutes for the patient to have a positive outcome. Route A is estimated to take 10 minutes while route B is estimated to take 12 minutes. But route A has a much higher probability than route B of taking more than 15 minutes due to uncertainty about the

traffic conditions. Hence the ambulance driver takes route B and arrives at the hospital in 13 minutes.

In each case, accounting for uncertainty when the agent (a car, scientist, or medic) was making a decision led to a superior outcome. While non-Bayesian methods can incorporate prior knowledge and uncertainty estimates, doing so tends to require ad hoc constructions. Furthermore, when using Bayesian inference methods, there is no need to know the exact question ahead of time. Once we have (an approximation to) the posterior, we can interrogate it repeatedly, after which a colleague might come along and ask her own questions.

A core challenge of Bayesian inference is applying Bayes' theorem – that is, calculating the posterior distribution. In particular, we typically want to be able to calculate posterior expectations. However, for all but the simplest models and most basic questions, such expectations cannot be calculated in closed form. Hence, we must turn to approximate inference methods. Modern applications of Bayesian methods present many challenges for classical approximate inference algorithms because any or all of the following apply: the latent parameter of interest is high-dimensional; the models are complex and even simulating from the model may be computationally expensive; and there are large amounts of data, which often are only available as a stream or are distributed across many computers. To overcome these challenges we require:

1. **Scalability.** Approximate inference algorithms must be applicable to datasets with large numbers of observations. Settings with streaming or distributed data are particularly important. Inference algorithms should also be able to operate on high-dimensional data and high-dimensional latent parameters.

2. **Arbitrary accuracy.** Approximate inference algorithms should also be able to provide arbitrary levels of accuracy, given a sufficiently large computational budget.

3. **Validation of the approximation quality.** We must be able to validate the quality of approximate posteriors output by inference algorithms. Validation can take the form of

   (a) *a priori* finite-time, finite-data guarantees; or

   (b) *post hoc* quality measures that are scalable and theoretically sound.

Requirement 1 ensures that an inference algorithm is suitable for modern applications. Requirement 2 is important because applications have varying precision requirements. For example, many scientific applications require very accurate posteriors while for industry prediction tasks a lower level of precision could be acceptable. There is an inherent tension between these two requirements: greater accuracy tends to require greater computational resources, which makes an algorithm less scalable. Hence, the computational–statistical tradeoffs of an inference algorithm must be considered and deemed favorable in at least some regime to be worth using. Finally,

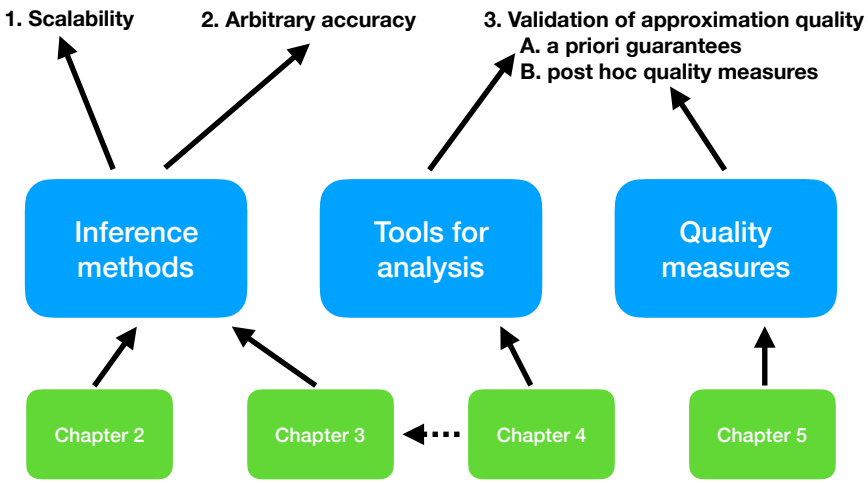


Figure 1-1: Overview of objectives and thesis structure.

Requirement 3 is necessary for practitioners to have faith in the reliability of the decisions and conclusions they will make based on an inference algorithm's output. *A priori* guarantees (Requirement 3a) are ideal since they provide confidence even before any inferential work is done. However, *post hoc* guarantees (Requirement 3b) are also useful because *a priori* guarantees may be unavailable or we may wish to compare methods which have incomparable of guarantees.

How should we approach the development of approximate inference algorithms that will simultaneously scale to large datasets, provide attractive computational–statistical tradeoffs, and come equipped with *a priori* guarantees? An important insight, which I will repeatedly leverage in this thesis, is that in the large-scale setting, much of the data can be *redundant* (though there may also be a small set of data points that are distinctive). Therefore, it is possible that we can *compress* our data into a form that admits more efficient inference. Approaching approximate inference from this perspective makes computational–statistical tradeoffs conceptually straightforward: less compression translates to greater accuracy but less computational savings. Chapters 2 and 3 describe two approaches to compressing data for improved scalability. In Chapter 2, my approach is to construct a *coreset*: a small, weighted subset of our data that is representative of the complete dataset. In Chapter 3, I show how to construct an exponential family model that is close to the original model. The data is then compressed by calculating the finite-dimensional sufficient statistics of the data under the exponential family.

Another advantage of the compression approach is that it involves replacing the original likelihood with a (deterministic) approximate likelihood. In Chapter 4, I show how such approximate likelihoods lend them themselves to *a priori* analysis. I am thus able to develop general tools for achieving Requirement 3a when using approximate likelihoods. I apply the tools developed in Chapter 4 to obtain *a priori* guarantees on the approximate posteriors produced by the methodology from Chapter 3. Finally, in Chapter 5, I develop computationally efficient quality measures

with provable guarantees, toward the goal of satisfying Requirement 3b. Fig. 1-1 provides an outline of the thesis and indicates how each chapter relates to the three requirements.

1.1 Related Work

1.1.1 Approximate Bayesian Inference

It is difficult to find approximate inference methods that meet Requirements 1, 2, and 3a. Most existing methodologies fall into the trilemma of offering at most two out of the three. Markov chain Monte Carlo (MCMC) methods provide an approximate posterior, and the approximation typically becomes arbitrarily good as the amount of computation time grows asymptotically; thereby MCMC satisfies Requirements 2 and 3a. But scalability of MCMC can be an issue. Conversely, variational Bayes (VB) and expectation propagation (EP) [95, 139] have grown in popularity due to their scalability to large data and models—though they typically lack guarantees on quality (failing Requirement 3a). Furthermore, traditional variational methods produce approximations of limited fidelity, hence failing Requirement 2. However, some recent work has sought to overcome the latter limitation [65, 94, 115, 121, 122].

Subsampling methods have been proposed to speed up MCMC [3, 14, 15, 40, 81, 89, 91, 111, 142] and VB [71]. Only a few of these algorithms preserve accuracy guarantees asymptotic in time (Requirement 2), and they often require restrictive assumptions. On the scalability front (Requirement 1), many though not all subsampling MCMC methods have been found to require examining a constant fraction of the data at each iteration [5, 15, 17, 108, 109, 135], so the computational gains are limited. Moreover, the random data access required by these methods may be infeasible for very large datasets that do not fit into memory. Finally, they do not apply to streaming and distributed data, and thus fail to fully satisfy Requirement 1.

Recently, authors have proposed subsampling methods based on piecewise deterministic Markov processes (PDMPs) [20, 22, 49, 106]. These methods are promising since subsampling data here does not change the invariant distribution of the continuous-time Markov process. But these methods have not yet been validated on large datasets nor is it understood how subsampling affects the mixing rates of the Markov processes.

Authors have also proposed methods for coalescing information across distributed computation (Requirement 1) in MCMC [47, 96, 112, 123, 129], VB [26, 29], and EP [58, 69]—and in the case of VB, across epochs as streaming data is collected [26, 29]. While these methods lead to gains in computational efficiency, most lack rigorous justification and provide no guarantees on the quality of inference (failing Requirement 3a). Some approaches, such as the consensus method of Minsker et al. [96], do have supporting theory, but cannot be made arbitrarily accurate (failing Requirement 2). See Angelino et al. [6] for a broader discussion of issues surrounding scalable Bayesian inference.

1.1.2 A priori guarantees

A priori accuracy guarantees for non-trivial models are most widely available for Markov chain Monte Carlo methods. Quantitative convergence of MCMC is a well-studied, though challenging, topic. Jones and Hobert [77] and Roberts and Rosenthal [117] provide excellent, though now dated, overviews of the techniques used to analyze MCMC convergence. The primary objective in the literature is to prove specific Markov chains are geometrically ergodic, which means that the Markov chain will converge exponentially quickly to producing samples from to the correct distribution. Example applications include convergence rates of the (block) Gibbs sampler for a Bayesian hierarchical version of the one-way random effects model [78] and conditions for fast convergence of parallel and simulated tempering algorithms for certain Gaussian mixture models and mean field Ising models [145]. Recently Khare and Hobert [80] and Choi and Hobert [32] proved, respectively, the geometric ergodicity of the Bayesian Lasso algorithm [107] and the uniform geometric ergodicity of the Polya-gamma Gibbs sampler for Bayesian logistic regression. Further examples can be found in Ge et al. [54], Qin and Hobert [110] and Zanella and Roberts [147]. More generally, many fast convergence results are known for sampling from strongly log-concave distributions [31, 42, 43, 45, 92, 118]

Recently, some posterior quality results have become available for variational Bayes [4, 141, 148]. However, these are asymptotic in the number of observation, whereas we seek non-asymptotic guarantees since uncertainty is typically relevant precisely in the non-asymptotic regime.

1.1.3 Quality measures

Approximation quality measures have a long history, particularly in the MCMC literature where they are often called "convergence diagnostics". A popular measure is the Gelman-Rubin statistic [55, 56]. It requires running multiple Markov chains and operates by comparing the inter-chain variance to the intra-chain variance. If the inter-chain variance is much larger, this suggests the chains have not mixed well. While simple to use, there are a number of weaknesses to the method. It may be insufficiently expressive to detect convergence failure. Also, it only works under the assumption that the chains are converging to the correct distribution, which is not the case for many of the scalable MCMC algorithms discussed in Section 1.1.1 (e.g. Ahn et al. [3], Bardenet et al. [14, 15], Korattikara et al. [81], Maire et al. [91], Welling and Teh [142]).

Grosse et al. [64] and Cusumano-Towner and Mansinghka [37] independently introduced an algorithm for stochastically estimating the symmetrized KL divergence (known as the Jeffreys divergence) between approximate posterior samples and the true posterior. The algorithm, which is based on bidirectional annealed importance sampling, is exact in the case of data simulated from the underlying generative model. To estimate performance on real data, Grosse et al. [64] propose a less direct method: hyperparameters for the generative model are learned based on the real data, then the Jeffreys divergence is estimated for synthetic data simulated from the model con-

ditional on those data-dependent hyperparameters. Thus, while this approach can provide a general indicator of MCMC performance, its results cannot be considered a decisive means of evaluation. Also, like the Gelman-Rubin statistic, it is only applicable to asymptotically exact samplers.

A final approach is based on (kernelized) Stein discrepancies [60, 61]. Stein discrepancies guarantee error bounds when calculating expectations for a large class of functions. This class can be chosen such that the Stein discrepancy dominates weak convergence [61], meaning it can distinguish between arbitrary distributions (unlike the Gelman-Rubin statistic). Furthermore, it is applicable to biased samplers. A challenge with using Stein discrepancies is computational efficiency. The original approach of Gorham and Mackey [60] requires solving an expensive linear program. The kernel-based approach of Gorham and Mackey [61] scales quadratically in the number of samples for evaluation.

1.2 Bayesian inference in generalized linear models

In this thesis, and particularly for the scalable inference methods developed in Chapters 2 and 3, I will focus on Bayesian *generalized linear models* (GLMs). GLMs combine the interpretability of linear models with the flexibility of more general outcome distributions—including binary, ordinal, and heavy-tailed observations. For this reason, they are some of the most widely used models by practitioners, both in isolation and as a building block for more complex hierarchical models. The inference algorithms we develop are applicable to such hierarchical models as long as the data groups are fixed. Also GLMs are considered to be "simple" models, scaling them to large datasets in a computationally efficient manner remains difficult.

Formally, the basic Bayesian GLM setting is as follows. Let $\mathcal{Y} \subseteq \mathbb{R}$ be the observation space, $\mathcal{X} \subseteq \mathbb{R}^d$ be the covariate space, and $\Theta \subseteq \mathbb{R}^d$ be the parameter space, where $d \geq 1$ denotes both the covariate and parameter dimensionality. Let $\mathcal{D} := \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ be the observed data, where $N \geq 1$ denotes the number of observations. Write $\mathbf{X} \in \mathbb{R}^{N \times d}$ for the matrix of all covariates and $\mathbf{y} \in \mathbb{R}^N$ for the vector of all observations. For a single data point $(\mathbf{x}_n, y_n)$, a GLM likelihood can be written as

$$p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) = p(y_n \,|\, g^{-1}(\mathbf{x}_n \cdot \boldsymbol{\theta})),$$

where $\mu := g^{-1}(\mathbf{x}_n \cdot \boldsymbol{\theta})$ is the expected value of y_n and $g^{-1} : \mathbb{R} \to \mathbb{R}$ is the *inverse link function*. I will assume that the observations are independent conditional on the covariates and the parameter:

$$\log p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta}) = \textstyle\sum_{n=1}^N \log p(y_n \,|\, g^{-1}(\mathbf{x}_n \cdot \boldsymbol{\theta})).$$

I will also sometimes use the shorthand notation

$$\phi(y_n, \mathbf{x}_n \cdot \boldsymbol{\theta}) := \log p(y_n \,|\, g^{-1}(\mathbf{x}_n \cdot \boldsymbol{\theta}))$$

for the log-likelihood. I will call $\phi(y, s) := \log p(y \,|\, g^{-1}(s))$ the GLM *mapping function*.

Examples of GLMs include some of the most widely used models in the statistical toolbox.

Example 1.2.1 (Binary regression)**.** For binary observations $y \in \{\pm 1\}$, the likelihood model is Bernoulli, $p(y = 1 \,|\, \mu) = \mu$, and the link function is often either the logit $g(\mu) = \log \frac{\mu}{1-\mu}$ (as in logistic regression) or the probit $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the standard Gaussian CDF.

Example 1.2.2 (Poisson regression)**.** When modeling count data $y \in \mathbb{N}$, the likelihood model might be Poisson, $p(y \,|\, \mu) = \mu^y e^{-\mu}/y!$, and $g(\mu) = \log(\mu)$ is the typical log link.

Example 1.2.3 (Robust regression)**.** For robust regression of real-valued data $y \in \mathbb{R}$, the identity link $g(\mu) = \mu$ can be paired with the Laplace likelihood

$$p(y \,|\, \mu) = \frac{1}{2b} e^{-|y-\mu|/b},$$

the Cauchy likelihood

$$p(y \,|\, \mu) = \frac{1}{\pi b \Big(1 + \frac{(y-\mu)^2}{b^2}\Big)},$$

the "Huber" log-likelihood

$$\log p(y \,|\, \mu) = \begin{cases} -\frac{1}{2}(y-\mu)^2 & |s-y| \le b \\ -b|y-\mu| + \frac{1}{2}b^2 & \text{otherwise,} \end{cases}$$

or the "smoothed Huber" log-likelihood

$$\log p(y \,|\, \mu) = -b^2 \left(\sqrt{1 + \frac{(y-\mu)^2}{b^2}} - 1 \right),$$

where in each case b serves as a scale parameter.

Example 1.2.4 (Gamma regression)**.** For positive data $y \in \mathbb{R}_+$, the likelihood model might be gamma, $p(y \,|\, \mu) = \Gamma(\nu)^{-1}(\nu/\mu)^\nu y^{\nu-1} e^{-\nu y/\mu}$, paired with the log link.

If we place a prior $\pi_0(\mathrm{d}\boldsymbol{\theta})$ on the parameters, then a full Bayesian analysis aims to approximate the (typically intractable) GLM posterior distribution

$$\pi_{\mathcal{D}}(\mathrm{d}\boldsymbol{\theta}) := \frac{p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta})\, \pi_0(\mathrm{d}\boldsymbol{\theta})}{\mathcal{E}_{\mathcal{D}}},$$

where $\mathcal{E}_{\mathcal{D}} := \int p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta})\, \pi_0(\mathrm{d}\boldsymbol{\theta})$ is the marginal likelihood (a.k.a. the model evidence). The *maximum a posteriori* (MAP) solution gives a point estimate of the parameter:

$$\boldsymbol{\theta}_{\mathrm{MAP}} := \arg\max_{\boldsymbol{\theta} \in \Theta} \pi_{\mathcal{D}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta} \in \Theta} \log \pi_0(\boldsymbol{\theta}) + \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}), \tag{1.1}$$

where $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) := \log p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})$ is the data log-likelihood. The MAP problem strictly generalizes finding the maximum likelihood estimate (MLE), since the MAP solution equals the MLE when using the (possibly improper) prior $\pi_0(\boldsymbol{\theta}) = 1$.

1.3 Computation and exponential families

In large part due to the high-dimensional integral implicit in the normalizing constant, approximating the posterior via, e.g., MCMC or VB, is often prohibitively expensive. Approximating this integral will typically require many evaluations of the (log-)likelihood, or its gradient, and each evaluation may require $\Omega(N)$ time.

Computation is much more efficient, though, if the model is in an *exponential family* (EF). In the EF case, there exist functions $\mathbf{t} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^m$ and $\boldsymbol{\eta} : \Theta \to \mathbb{R}^m$, such that[1]

$$\log p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = \mathbf{t}(y_n, \mathbf{x}_n) \cdot \boldsymbol{\eta}(\boldsymbol{\theta}) =: \mathcal{L}_{\mathcal{D},\mathrm{EF}}(\boldsymbol{\theta}; \mathbf{t}(y_n, \mathbf{x}_n)).$$

Thus, if the observations are conditionally independent, we can rewrite the log-likelihood as

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \textstyle\sum_{n=1}^{N} \mathcal{L}_{\mathcal{D},\mathrm{EF}}(\boldsymbol{\theta}; \mathbf{t}(y_n, \mathbf{x}_n)) =: \mathcal{L}_{\mathcal{D},\mathrm{EF}}(\boldsymbol{\theta}; \mathbf{t}(\mathcal{D})),$$

where $\mathbf{t}(\mathcal{D}) := \sum_{n=1}^{N} \mathbf{t}(y_n, \mathbf{x}_n)$. The *sufficient statistics* $\mathbf{t}(\mathcal{D})$ can be calculated in $O(N)$ time, after which each evaluation of $\mathcal{L}_{\mathcal{D},\mathrm{EF}}(\boldsymbol{\theta}; \mathbf{t}(\mathcal{D}))$ or $\nabla \mathcal{L}_{\mathcal{D},\mathrm{EF}}(\boldsymbol{\theta}; \mathbf{t}(\mathcal{D}))$ requires only $O(1)$ time. Thus, instead of K passes over N data points (requiring $O(NK)$ time), only $O(N + K)$ time is needed. Even for moderate values of N, the time savings can be substantial when K is large.

The Poisson distribution is an illustrative example of a one-parameter exponential family with $\mathbf{t}(y) = (1, y, \log y!)$ and $\boldsymbol{\eta}(\theta) = (\theta, \log \theta, 1)$. Thus, if we have data $\mathbf{y}$ (there are no covariates), $\mathbf{t}(\mathbf{y}) = (N, \sum_n y_n, \sum \log y_n!)$. In this case it is easy to calculate the maximum likelihood estimate of θ from $\mathbf{t}(\mathbf{y})$ as $t_1(\mathbf{y})/t_0(\mathbf{y}) = N^{-1} \sum_n y_n$.

Unfortunately, GLMs rarely belong to an exponential family – even if the outcome distribution is in an exponential family, the use of a link destroys the EF structure. In logistic regression, we write (overloading the ϕ notation) $\log p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = \phi_{\mathsf{logit}}(y_n \mathbf{x}_n \cdot \boldsymbol{\theta})$, where $\phi_{\mathsf{logit}}(s) := -\log(1 + e^{-s})$. For Poisson regression with log link, $\log p(y_n \mid \mathbf{x}_n, \boldsymbol{\theta}) = \phi_{\mathsf{Poisson}}(y_n, \mathbf{x}_n \cdot \boldsymbol{\theta})$, where $\phi_{\mathsf{Poisson}}(y, s) := ys - e^s - \log y!$. In both cases, we cannot express the log-likelihood as an inner product between a function solely of the data and a function solely of the parameter.

[1] Our presentation is slightly different from the standard textbook account because we have implicitly absorbed the base measure and log-partition function into $\mathbf{t}$ and $\boldsymbol{\eta}$.

1.4 Approximate Bayesian inference via likelihood approximation

The methods I present in this thesis can be viewed as ways to approximate a non-exponential family model log-likelihood with *approximate* sufficient statistics that are of low-dimensionality compared to the amount of data. Sufficient statistics are a classical and powerful technique in statistics and the methods in Chapters 2 and 3 are ways to adapt these venerable ideas to the modern statistical and computational landscape. Toward this end, we wish to find functions $\tilde{\mathbf{t}} : \mathcal{Y}^N \times \mathcal{X}^N \to \mathbb{R}^m$ and $\tilde{\boldsymbol{\eta}} : \Theta \to \mathbb{R}^m$, for some $m \ll N$, such that

$$\log p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta}) \approx \tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X}) \cdot \tilde{\boldsymbol{\eta}}(\boldsymbol{\theta}).$$

We call $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X})$ a set of *approximate sufficient statistics for* $\mathcal{D}$. We would like to calculate $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X})$ and construct $\tilde{\boldsymbol{\eta}}$ efficiently and in such a way that $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta})$ can be calculated in $O(md)$ time.

In Chapter 2, we will view the data itself as a set of sufficient statistics, which we will then attempt to approximate. For $i = 1, 2, \ldots, m$, let $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta})_i := \log p(y_{n_i} \,|\, \mathbf{x}_{n_i}, \boldsymbol{\theta})$ for some $n_i \in [N]$ and take $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X}) \in \mathbb{R}^m_+$. Together these provide a coreset approximation to the data, where $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X})_i$ denotes the weight given to the i-th datapoint in the coreset $\tilde{\mathcal{D}} = \{(y_{n_i}, \mathbf{x}_{n_i})\}_{i=1}^m$. We will show how to efficiently choose the indices n_i and the weights $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X})$. In Chapter 3, we take a rather different approach. We will let $\tilde{\mathbf{t}}(\mathbf{y}, \mathbf{X}) := \sum_{n=1}^N \tilde{\mathbf{t}}(y_n, \mathbf{x}_n)$, where for $i = 1, 2, \ldots, m$, $\tilde{\mathbf{t}}(y_n, \mathbf{x}_n)_i$ will be a low-degree polynomial in $y_n, x_{n1}, x_{n2}, \ldots, x_{nd}$. Similarly, $\tilde{\boldsymbol{\eta}}(\boldsymbol{\theta})_i$ is a low-degree polynomial in $\theta_1, \theta_2, \ldots, \theta_d$. In Chapter 4 we will develop some general tools for controlling the error in posterior approximations based on likelihood approximations. In particular, we will bound the Wasserstein distance between the exact and approximate posteriors in terms of the (expected) error between the gradients of the exact and approximate log-likelihoods. Finally, in Chapter 5 I will develop proxies for the kernel Stein discrepancies (KSDs) described in Section 1.1.3 that can be computed in time almost linear (rather than quadratic) in the number of samples, while still retaining the theoretical soundness of KSDs.

Chapter 2

Logistic Regression Coresets

In this chapter I leverage data redundancy to develop a scalable Bayesian inference framework that modifies the *dataset* instead of the common practice of modifying the inference algorithm. As I discussed earlier, while much of the data may be redundant, some portion of it may be more distinctive. For example, in a large document corpus, one news article about a hockey game may serve as an excellent representative of hundreds or thousands of other similar pieces about hockey games. However, there may only be a few articles about luge, so it is also important to include at least one article about luge. Similarly, one individual's genetic information may serve as a strong representative of other individuals from the same ancestral population admixture, though some individuals may be genetic outliers. My method, which can be thought of as a preprocessing step, constructs a *coreset* – a small, weighted subset of the data that approximates the full dataset [2, 50] – that can be used in many standard inference procedures to provide posterior approximations with guaranteed quality. The scalability of posterior inference with a coreset thus simply depends on the coreset's growth with the full dataset size. To the best of my knowledge, coresets have not previously been used in a Bayesian setting.

The concept of coresets originated in computational geometry (e.g. [2]), but then became popular in theoretical computer science as a way to efficiently solve clustering problems such as k-means and PCA (see [50, 52] and references therein). Coreset research in the machine learning community has focused on scalable clustering in the optimization setting [10, 11, 52, 87], with the exception of Feldman et al. [51], who developed a coreset algorithm for Gaussian mixture models. Coreset-like ideas have previously been explored for maximum likelihood-learning of logistic regression models, though these methods either lack rigorous justification or have only asymptotic guarantees (see [68] and references therein as well as [90], which develops a methodology applicable beyond logistic regression).

The job of the coreset in the Bayesian setting is to provide an approximation of the full data log-likelihood up to a multiplicative error uniformly over the parameter space. I will begin with a theoretical analysis of the quality of the posterior distribution obtained from such an approximate log-likelihood. The remainder of the chapter develops the efficient construction of small coresets for Bayesian logistic regression. I develop a coreset construction algorithm, the output of which uniformly approxi-

mates the full data log-likelihood over parameter values in a ball with a user-specified radius. The approximation guarantee holds for a given dataset with high probability. I also obtain results showing that the boundedness of the parameter space is necessary for the construction of a nontrivial coreset, as well as results characterizing the algorithm's expected performance under a wide class of data-generating distributions. Our proposed algorithm is applicable in both the streaming and distributed computation settings, and the coreset can then be used by any inference algorithm which accesses the (gradient of the) log-likelihood as a black box. Although our coreset algorithm is specifically for logistic regression, our approach is broadly applicable to other Bayesian generative models.

Experiments on a variety of synthetic and real-world datasets validate our approach and demonstrate robustness to the choice of algorithm hyperparameters. An empirical comparison to random subsampling shows that, in many cases, coreset-based posteriors are orders of magnitude better in terms of maximum mean discrepancy, including on a challenging 100-dimensional real-world dataset. Crucially, our coreset construction algorithm adds negligible computational overhead to the inference procedure.

2.1 Bayesian Coresets

Our aim is to construct a weighted dataset $\tilde{\mathcal{D}} = \{(\gamma_m, \tilde{\mathbf{x}}_m, \tilde{y}_m)\}_{m=1}^M$ with $M \ll N$ such that the weighted log-likelihood $\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{m=1}^M \gamma_m \log p(\tilde{y}_m \,|\, \tilde{\mathbf{x}}_m, \boldsymbol{\theta})$ satisfies

$$|\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta})| \leq \varepsilon |\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})|, \quad \forall \boldsymbol{\theta} \in \Theta. \tag{2.1}$$

If $\tilde{\mathcal{D}}$ satisfies Eq. (2.1), it is called an *ε-coreset of $\mathcal{D}$*, and the approximate posterior

$$\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{\exp(\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta})}{\tilde{\mathcal{E}}_{\mathcal{D}}}, \qquad \text{with} \quad \tilde{\mathcal{E}}_{\mathcal{D}} = \int \exp(\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta},$$

has a marginal likelihood $\tilde{\mathcal{E}}_{\mathcal{D}}$ which approximates the true marginal likelihood $\mathcal{E}_{\mathcal{D}}$, shown by Proposition 2.1.1. Thus, from a Bayesian perspective, the ε-coreset is a useful notion of approximation.

Proposition 2.1.1. *Let $\mathcal{L}(\boldsymbol{\theta})$ and $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ be arbitrary non-positive log-likelihood functions that satisfy $|\mathcal{L}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}(\boldsymbol{\theta})| \leq \varepsilon |\mathcal{L}(\boldsymbol{\theta})|$ for all $\boldsymbol{\theta} \in \Theta$. Then for any prior $\pi_0(\boldsymbol{\theta})$ such that the marginal likelihoods*

$$\mathcal{E} = \int \exp(\mathcal{L}(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta} \qquad \textit{and} \qquad \tilde{\mathcal{E}} = \int \exp(\tilde{\mathcal{L}}(\boldsymbol{\theta}))\pi_0(\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{\theta}$$

are finite, the marginal likelihoods satisfy

$$|\log \mathcal{E} - \log \tilde{\mathcal{E}}| \leq \varepsilon |\log \mathcal{E}|.$$

Algorithm 1 Construction of logistic regression coreset

Require: Data $\mathcal{D}$, k-clustering $\mathcal{Q}$, radius $R > 0$, tolerance $\varepsilon > 0$, failure rate $\delta \in (0, 1)$

for $n = 1, \ldots, N$ **do** ▷ calculate sensitivity upper bounds using the k-clustering
  $m_n \leftarrow \frac{N}{1+\sum_{i=1}^{k} |G_i^{(-n)}| e^{-R\|\bar{\mathbf{z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2}}$
end for
$\bar{m}_N \leftarrow \frac{1}{N}\sum_{n=1}^{N} m_n$
$M \leftarrow \left\lceil \frac{c\bar{m}_N^2}{\varepsilon^2}[(d+1) + \log(1/\delta)] \right\rceil$ ▷ coreset size; c is from proof of Theorem A.2.1
for $n = 1, \ldots, N$ **do**
  $p_n \leftarrow \frac{m_n}{N\bar{m}_N}$ ▷ importance weights of data
end for
$(K_1, \ldots, K_N) \sim \mathsf{Multi}(M, (p_n)_{n=1}^N)$ ▷ sample data for coreset
for $n = 1, \ldots, N$ **do** ▷ calculate coreset weights
  $\gamma_n \leftarrow \frac{K_n}{p_n M}$
end for
$\tilde{\mathcal{D}} \leftarrow \{(\gamma_n, \mathbf{x}_n, y_n) \,|\, \gamma_n > 0\}$ ▷ only keep data points with non-zero weights
return $\tilde{\mathcal{D}}$

2.2 Coresets for Logistic Regression

2.2.1 Coreset Construction

Recall that in logistic regression, the covariates are real feature vectors $\mathbf{x}_n \in \mathbb{R}^d$, the observations are labels $y_n \in \{-1, 1\}$, $\Theta \subseteq \mathbb{R}^d$, and the likelihood is defined as

$$p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) = p_{logistic}(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) := \frac{1}{1 + \exp(-y_n \mathbf{x}_n \cdot \boldsymbol{\theta})}.$$

The analysis in this work allows any prior $\pi_0(\boldsymbol{\theta})$; common choices are the Gaussian, Cauchy [57], and spike-and-slab [59, 98]. For notational brevity, define $\mathbf{z}_n := y_n \mathbf{x}_n$, and let $\phi(s) := \log(1 + \exp(-s))$. Choosing the optimal ϵ-coreset is not computationally feasible, so I take a less direct approach. I design a coreset construction algorithm and prove its correctness using a quantity $\sigma_n(\Theta)$ called the *sensitivity* [50], which quantifies the redundancy of a particular data point n – the larger the sensitivity, the less redundant. In the setting of logistic regression, we have that the sensitivity is

$$\sigma_n(\Theta) := \sup_{\boldsymbol{\theta} \in \Theta} \frac{N\, \phi(\mathbf{z}_n \cdot \boldsymbol{\theta})}{\sum_{\ell=1}^{N} \phi(Z_\ell \cdot \boldsymbol{\theta})}.$$

Intuitively, $\sigma_n(\Theta)$ captures how much influence data point n has on the log-likelihood $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ when varying the parameter $\boldsymbol{\theta} \in \Theta$, and thus data points with high sensitivity should be included in the coreset. Evaluating $\sigma_n(\Theta)$ exactly is not tractable, however, so an upper bound $m_n \geq \sigma_n(\Theta)$ must be used in its place. Thus, the key challenge is

to efficiently compute a tight upper bound on the sensitivity.

For the moment I will consider $\Theta = \mathbb{B}_R$ for any $R > 0$, where $\mathbb{B}_R := \{\boldsymbol{\theta} \in \mathbb{R}^d \,|\, \|\boldsymbol{\theta}\|_2 \leq R\}$; I discuss the case of $\Theta = \mathbb{R}^d$ shortly. Choosing the parameter space to be a Euclidean ball is reasonable since data is usually preprocessed to have mean zero and variance 1 (or, for sparse data, to be between -1 and 1), so each component of $\boldsymbol{\theta}$ is typically in a range close to zero (e.g. between -4 and 4) [57].

The idea behind our sensitivity upper bound construction is that we would expect data points that are bunched together to be redundant while data points that are far from from other data have a large effect on inferences. Clustering is an effective way to summarize data and detect outliers, so I will use a *k-clustering* of the data $\mathcal{D}$ to construct the sensitivity bound. A k-clustering is given by k cluster centers $\mathcal{Q} = \{Q_1, \ldots, Q_k\}$. Let $G_i := \{\mathbf{z}_n \,|\, i = \arg\min_j \|Q_j - \mathbf{z}_n\|_2\}$ be the set of vectors closest to center Q_i and let $G_i^{(-n)} := G_i \setminus \{\mathbf{z}_n\}$. Define $\mathbf{Z}_{G,i}^{(-n)}$ to be a uniform random vector from $G_i^{(-n)}$ and let $\bar{\mathbf{Z}}_{G,i}^{(-n)} := \mathbb{E}[\mathbf{Z}_{G,i}^{(-n)}]$ be its mean. The following lemma uses a k-clustering to establish an efficiently computable upper bound on $\sigma_n(\mathbb{B}_R)$:

Lemma 2.2.1. *For any k-clustering $\mathcal{Q}$,*

$$\sigma_n(\mathbb{B}_R) \leq m_n := \frac{N}{1 + \sum_{i=1}^k |G_i^{(-n)}| e^{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2}}. \tag{2.2}$$

Furthermore, m_n can be calculated in $O(k)$ time.

The bound in Eq. (2.2) captures the intuition that if the data forms tight clusters (that is, each $\mathbf{z}_n$ is close to one of the cluster centers), we expect each cluster to be well-represented by a small number of typical data points. For example, if $\mathbf{z}_n \in G_i$, $\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2$ is small, and $|G_i^{(-n)}| = \Theta(N)$, then $\sigma_n(\mathbb{B}_R) = O(1)$. I use the (normalized) sensitivity bounds obtained from Lemma 2.2.1 to form an importance distribution $(p_n)_{n=1}^N$ from which to sample the coreset. I sample $\mathbf{z}_n$, then assign it weight γ_n proportional to $1/p_n$. The size of the coreset depends on the mean sensitivity bound, the desired error ε, and a quantity closely related to the VC dimension of $\boldsymbol{\theta} \mapsto \phi(\boldsymbol{\theta} \cdot \mathbf{Z})$, which I show is $d + 1$. Combining these pieces we obtain Algorithm 1, which constructs an ε-coreset with high probability by Theorem 2.2.2.

Theorem 2.2.2. *Fix $\varepsilon > 0$, $\delta \in (0, 1)$, and $R > 0$. Consider a dataset $\mathcal{D}$ with k-clustering $\mathcal{Q}$. With probability at least $1 - \delta$, Algorithm 1 with inputs $(\mathcal{D}, \mathcal{Q}, R, \varepsilon, \delta)$ constructs an ε-coreset of $\mathcal{D}$ for logistic regression with parameter space $\Theta = \mathbb{B}_R$. Furthermore, Algorithm 1 runs in $O(Nk)$ time.*

Remark 2.2.3. The coreset algorithm is efficient with an $O(Nk)$ running time. However, the algorithm requires a k-clustering, which must also be constructed. A high-quality clustering can be obtained cheaply via k-means++ in $O(Nk)$ time [7], although a coreset algorithm could also be used.

Examining Algorithm 1, we see that the coreset size M is of order $\bar{m}_N \log \bar{m}_N$, where $\bar{m}_N = \frac{1}{N}\sum_n m_n$. So for M to be smaller than N, at a minimum, $\bar{m}_N$ should satisfy $\bar{m}_N = \tilde{o}(N)$,[1] and preferably $\bar{m}_N = O(1)$. Indeed, for the coreset size to be

[1] Recall that the tilde notation suppresses logarithmic terms.

small, it is critical that (a) Θ is chosen such that most of the sensitivities satisfy $\sigma_n(\Theta) \ll N$ (since N is the maximum possible sensitivity), (b) each upper bound m_n is close to $\sigma_n(\Theta)$, and (c) ideally, that $\bar{m}_N$ is bounded by a constant. In Section 2.2.2, I address (a) by providing sensitivity lower bounds, thereby showing that the constraint $\Theta = \mathbb{B}_R$ is necessary for nontrivial sensitivities even for "typical" (i.e. non-pathological) data. I then apply our lower bounds to address (b) and show that our bound in Lemma 2.2.1 is nearly tight. In Section 2.2.3, I address (c) by establishing the expected performance of the bound in Lemma 2.2.1 for a wide class of data-generating distributions.

2.2.2 Sensitivity Lower Bounds

I now develop lower bounds on the sensitivity to demonstrate that essentially we must limit ourselves to bounded Θ,[2] thus making our choice of $\Theta = \mathbb{B}_R$ a natural one, and to show that the sensitivity upper bound from Lemma 2.2.1 is nearly tight.

I begin by showing that in both the worst case and the average case, for all n, $\sigma_n(\mathbb{R}^d) = N$, the maximum possible sensitivity – even when the $\mathbf{z}_n$ are arbitrarily close. Intuitively, the reason for the worst-case behavior is that if there is a separating hyperplane between a data point $\mathbf{z}_n$ and the remaining data points, and $\boldsymbol{\theta}$ is in the direction of that hyperplane, then when $\|\boldsymbol{\theta}\|_2$ becomes very large, $\mathbf{z}_n$ becomes arbitrarily more important than any other data point.

Theorem 2.2.4. *For any $d \geq 3$, $N \in \mathbb{N}$ and $0 < \epsilon' < 1$, there exists $\epsilon > 0$ and unit vectors $\mathbf{z}_1, \ldots, \mathbf{z}_N \in \mathbb{R}^d$ such that for all pairs n, n', $\mathbf{z}_n \cdot \mathbf{z}_{n'} \geq 1 - \epsilon'$ and for all $R > 0$ and n,*

$$\sigma_n(\mathbb{B}_R) \geq \frac{N}{1 + (N-1)e^{-R\epsilon\sqrt{\epsilon'}/4}}, \qquad \text{and hence} \qquad \sigma_n(\mathbb{R}^d) = N.$$

The proof of Theorem 2.2.4 is based on choosing N distinct unit vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^{d-1}$ and setting $\epsilon = 1 - \max_{n \neq n'} \mathbf{v}_n \cdot \mathbf{v}_{n'} > 0$. But what is a "typical" value for ϵ? In the case of the vectors being uniformly distributed on the unit sphere, we have the following scaling for ϵ as N increases:

Proposition 2.2.5. *If $\mathbf{v}_1, \ldots, \mathbf{v}_N$ are independent and uniformly distributed on the unit sphere $\mathbb{S}^d := \{v \in \mathbb{R}^d \mid \|v\| = 1\}$ with $d \geq 2$, then with high probability*

$$1 - \max_{n \neq n'} \mathbf{v}_n \cdot \mathbf{v}_{n'} \geq C_d N^{-4/(d-1)},$$

where C_d is a constant depending only on d.

Furthermore, N can be exponential in d even with ϵ remaining very close to 1:

Proposition 2.2.6. *For $N = \lfloor \exp((1-\epsilon)^2 d/4)/\sqrt{2} \rfloor$, and $\mathbf{v}_1, \ldots, \mathbf{v}_N$ i.i.d. such that $v_{ni} = \pm\frac{1}{\sqrt{d}}$ with probability $\frac{1}{2}$, then with probability at least $\frac{1}{2}$, $1 - \max_{n \neq n'} \mathbf{v}_n \cdot \mathbf{v}_{n'} \geq \epsilon$.*

[2] Certain pathological datasets allow us to use unbounded Θ, but I do not assume we are given such data.

Propositions 2.2.5 and 2.2.6 demonstrate that the data vectors $\mathbf{z}_n$ found in Theorem 2.2.4 are, in two different senses, "typical" vectors and should not be thought of as worst-case data only occurring in some "negligible" or zero-measure set. These three results thus demonstrate that it is necessary to restrict attention to bounded Θ. One can also use Theorem 2.2.4 to show that our sensitivity upper bound is nearly tight.

Corollary 2.2.7. *For the data $\mathbf{z}_1, \ldots, \mathbf{z}_N$ from Theorem 2.2.4,*

$$\frac{N}{1 + (N-1)e^{-R\epsilon\sqrt{\epsilon'}/4}} \leq \sigma_n(\mathbb{B}_R) \leq \frac{N}{1 + (N-1)e^{-R\sqrt{2\epsilon'}}}.$$

2.2.3 k-Clustering Sensitivity Bound Performance

While Lemma 2.2.1 and Corollary 2.2.7 provide an upper bound on the sensitivity given a fixed dataset, we would also like to understand how the expected mean sensitivity increases with N. We might expect it to be finite since the logistic regression likelihood model is parametric; the coreset would thus be acting as a sort of approximate finite sufficient statistic. Proposition 2.2.8 characterizes the expected performance of the upper bound from Lemma 2.2.1 under a wide class of generating distributions. This result demonstrates that, under reasonable conditions, the expected value of $\bar{m}_N$ is bounded for all N. As a concrete example, Corollary 2.2.9 specializes Proposition 2.2.8 to data with a single shared Gaussian generating distribution.

Proposition 2.2.8. *Let $\mathbf{x}_n \overset{indep}{\sim} \mathcal{N}(\mu_{L_n}, \Sigma_{L_n})$, where $L_n \overset{indep}{\sim} \mathsf{Multi}(\pi_1, \pi_2, \ldots)$ is the mixture component responsible for generating $\mathbf{x}_n$. For $n = 1, \ldots, N$, let $y_n \in \{-1, 1\}$ be conditionally independent given $\mathbf{x}_n$ and set $\mathbf{z}_n = y_n \mathbf{x}_n$. Select $0 < r < 1/2$, and define $\eta_i = \max(\pi_i - N^{-r}, 0)$. The clustering of the data implied by $(L_n)_{n=1}^N$ results in the expected sensitivity bound*

$$\mathbb{E}[\bar{m}_N] \leq \frac{1}{N^{-1} + \sum_i \eta_i e^{-R\sqrt{A_i N^{-1} \eta_i^{-1} + B_i}}} + \sum_{i:\eta_i > 0} N e^{-2N^{1-2r}} \overset{N\to\infty}{\to} \frac{1}{\sum_i \pi_i e^{-R\sqrt{B_i}}},$$

where

$$\begin{aligned} A_i &:= \mathrm{Tr}[\Sigma_i] + \big(1 - \bar{y}_i^2\big)\mu_i^T \mu_i, \\ B_i &:= \textstyle\sum_j \pi_j \big(\mathrm{Tr}[\Sigma_j] + \bar{y}_j^2 \mu_i^T \mu_i - 2\bar{y}_i \bar{y}_j \mu_i^T \mu_j + \mu_j^T \mu_j\big), \end{aligned}$$

and $\bar{y}_j = \mathbb{E}[y_1 | L_1 = j]$.

Corollary 2.2.9. *In the setting of Proposition 2.2.8, if $\pi_1 = 1$ and all data is assigned to a single cluster, then there is a constant C such that for sufficiently large N,*

$$\mathbb{E}[\bar{m}_N] \leq C e^{R\sqrt{\mathrm{Tr}[\Sigma_1] + (1 - \bar{y}_1^2)\mu_1^T \mu_1}}.$$

2.2.4 Streaming and Parallel Settings

Algorithm 1 is a batch algorithm, but it can easily be used in parallel and streaming computation settings using standard methods from the coreset literature, which are based on the following two observations (cf. [51, Section 3.2]):

1. If $\tilde{\mathcal{D}}_i$ is an ε-coreset for $\mathcal{D}_i$, $i = 1, 2$, then $\tilde{\mathcal{D}}_1 \cup \tilde{\mathcal{D}}_2$ is an ε-coreset for $\mathcal{D}_1 \cup \mathcal{D}_2$.
2. If $\tilde{\mathcal{D}}$ is an ε-coreset for $\mathcal{D}$ and $\tilde{\mathcal{D}}'$ is an ε'-coreset for $\tilde{\mathcal{D}}$, then $\tilde{\mathcal{D}}'$ is an ε''-coreset for $\mathcal{D}$, where $\varepsilon'' := (1+\varepsilon)(1+\varepsilon') - 1$.

We can use these observations to merge coresets that were constructed either in parallel, or sequentially, in a binary tree. Coresets are computed for two data blocks, merged using observation 1, then compressed further using observation 2. The next two data blocks have coresets computed and merged/compressed in the same manner, then the coresets from blocks 1&2 and 3&4 can be merged/compressed analogously. We continue in this way and organize the merge/compress operations into a binary tree. Then, if there are B data blocks total, only $\log B$ blocks ever need be maintained simultaneously. In the streaming setting we would choose blocks of constant size, so $B = O(N)$, while in the parallel setting B would be the number of machines available.

2.3 Experiments

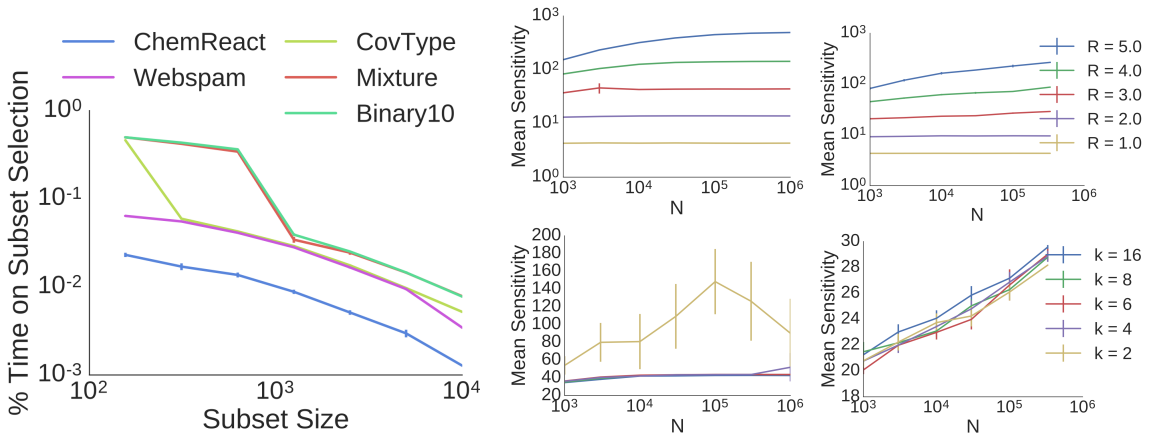
I evaluated the performance of the logistic regression coreset algorithm on a number of synthetic and real-world datasets. I used a maximum dataset size of 1 million examples because I wanted to be able to calculate the true posterior, which would be infeasible for extremely large datasets. The datasets I used are summarized in Table 2.1.

Synthetic Data. I generated synthetic binary data according to the model $X_{ni} \overset{\text{indep}}{\sim} \mathsf{Bern}(p_i), i = 1, \ldots, d$ and $y_n \overset{\text{indep}}{\sim} p_{logistic}(\cdot \,|\, \mathbf{x}_n, \boldsymbol{\theta})$. The idea is to simulate data in which there are a small number of rarely occurring but highly predictive features, which is a common real-world phenomenon. I thus took

$$\mathbf{p} = (1, .2, .3, .5, .01, .1, .2, .007, .005, .001) \qquad \text{and}$$
$$\boldsymbol{\theta} = (-3, 1.2, -.5, .8, 3, -1., -.7, 4, 3.5, 4.5)$$

for the $d = 10$ experiments (BINARY10) and the first 5 components of $\mathbf{p}$ and $\boldsymbol{\theta}$ for the $d = 5$ experiments (BINARY5). The generative model is the same one used by Scott et al. [123] and the first 5 components of $\mathbf{p}$ and $\boldsymbol{\theta}$ correspond to those used in the Scott et al. experiments (given in [123, Table 1b]). I generated a synthetic mixture dataset with continuous covariates (MIXTURE) using a model similar to that of Han et al. [68]: $y_n \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(1/2)$ and $\mathbf{x}_n \overset{\text{indep}}{\sim} \mathcal{N}(\mu_{y_n}, I)$, where $\mu_{-1} = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$ and $\mu_1 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$.

Real-world Data. The CHEMREACT dataset consists of $N = 26{,}733$ chemicals, each with $d = 100$ properties. The goal is to predict whether each chemical is reactive. The WEBSPAM corpus consists of $N = 350{,}000$ web pages, approximately 60% of



(a) (b) Binary10 (c) Webspam

Figure 2-1: **(A)** Percentage of time spent creating the coreset relative to the total inference time (including 10,000 iterations of MCMC). Except for very small coreset sizes, coreset construction is a small fraction of the overall time. **(B,C)** The mean sensitivities for varying choices of R and k. When R varies $k = 6$ and when k varies $R = 3$. The mean sensitivity increases exponentially in R, as expected, but is robust to the choice of k.

which are spam. The covariates consist of the $d = 127$ features that each appear in at least 25 documents. The cover type (CovType) dataset consists of $N = 581{,}012$ cartographic observations with $d = 54$ features. The task is to predict the type of trees that are present at each observation location.

2.3.1 Scaling Properties of the Coreset Construction Algorithm

Constructing Coresets. In order for coresets to be a worthwhile preprocessing step, it is critical that the time required to construct the coreset is small relative to the time needed to complete the inference procedure. I implemented the logistic regression coreset algorithm in Python.[6] More details on our implementation are provided in Section 2.3.3. In Fig. 2-1a, I plot the relative time to construct the coreset for each type of dataset ($k = 6$) versus the total inference time, including 10,000 iterations of the MCMC procedure described in Section 2.3.2. Except for very small coreset sizes, the time to run MCMC dominates.

[3]dataset `ds1.100` from `http://komarix.org/ac/ds/`.

[4]Available from `http://www.cc.gatech.edu/projects/doi/WebbSpamCorpus.html`

[5]dataset `covtype.binary` from `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.

[6]Code to recreate all of our experiments is available at `https://bitbucket.org/jhhuggins/lrcoresets`.

Table 2.1: datasets used for experiments

Name	N	d	positive examples	k
Low-dimensional Synthetic Binary	1M	5	9.5%	4
Higher-dimensional Synthetic Binary	1M	10	8.9%	4
Synthetic Balanced Mixture	1M	10	50%	4
Chemical Reactivity[3]	26,733	100	3%	6
Webspam[4]	350K	127	60%	6
Cover type[5]	581,012	54	51%	6

Sensitivity. An important question is how the mean sensitivity $\bar{m}_N$ scales with N, as it determines how the size of the coreset scales with the data. Furthermore, ensuring that mean sensitivity is robust to the number of clusters k is critical since needing to adjust the algorithm hyperparameters for each dataset could lead to an unacceptable increase in computational burden. I also seek to understand how the radius R affects the mean sensitivity. Figs. 2-1b and 2-1c show the results of our scaling experiments on the BINARY10 and WEBSPAM data. The mean sensitivity is essentially constant across a range of dataset sizes. For both datasets the mean sensitivity is robust to the choice of k and scales exponentially in R, as I would expect from Lemma 2.2.1.

2.3.2 Posterior Approximation Quality

Since the ultimate goal is to use coresets for Bayesian inference, the key empirical question is how well a posterior formed using a coreset approximates the true posterior distribution. I compared the coreset algorithm to random subsampling of data points, since that is the approach used in many existing scalable versions of variational inference and MCMC [14, 15, 71, 81]. Indeed, coreset-based importance sampling could be used as a drop-in replacement for the random subsampling used by these methods, though I leave the investigation of this idea for future work.

Experimental Setup. I used adaptive Metropolis-adjusted Langevin algorithm (MALA) [66, 118] for posterior inference. For each dataset, I ran the coreset and random subsampling algorithms 20 times for each choice of subsample size M. I ran adaptive MALA for 100,000 iterations on the full dataset and each subsampled dataset. The subsampled datasets were fixed for the entirety of each run, in contrast to subsampling algorithms that resample the data at each iteration. For the synthetic datasets, which are lower dimensional, I used $k = 4$ while for the real-world datasets, which are higher dimensional, I used $k = 6$. I used a heuristic to choose R as large as was feasible while still obtaining moderate total sensitivity bounds. For a clustering $\mathcal{Q}$ of data $\mathcal{D}$, let $\mathcal{I} := N^{-1} \sum_{i=1}^{k} \sum_{Z \in G_i} \|Z - Q_i\|^2$ be the normalized k-means score. I chose $R = a/\sqrt{\mathcal{I}}$, where a is a small constant. The idea is that, for $i \in [k]$ and $\mathbf{z}_n \in G_i$, we want $R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2 \approx a$ on average, so the term $\exp\{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2\}$ in Eq. (2.2) is not too small and hence $\sigma_n(\mathbb{B}_R)$ is not too large. Our experiments used

$a = 3$. I obtained similar results for $4 \leq k \leq 8$ and $2.5 \leq a \leq 3.5$, indicating that the logistic regression coreset algorithm has some robustness to the choice of these hyperparameters. I used negative test log-likelihood and maximum mean discrepancy (MMD) with a 3rd degree polynomial kernel as comparison metrics (so smaller is better).

Synthetic Data Results. Figures 2-2a-2-2c show the results for synthetic data. In terms of test log-likelihood, coresets did as well as or outperformed random subsampling. In terms of MMD, the coreset posterior approximation typically outperformed random subsampling by 1-2 orders of magnitude and never did worse. These results suggest much can be gained by using coresets, with comparable performance to random subsampling in the worst case.

Real-world Data Results. Figures 2-2d-2-2f show the results for real data. Using coresets led to better performance on ChemReact for small subset sizes. Because the dataset was fairly small and random subsampling was done without replacement, coresets were worse for larger subset sizes. Coreset and random subsampling performance was approximately the same for Webspam. On Webspam and CovType, coresets either outperformed or did as well as random subsampling in terms MMD and test log-likelihood on almost all subset sizes. The only exception was that random subsampling was superior on Webspam for the smallest subset set. I suspect this is due to the variance introduced by the importance sampling procedure used to generate the coreset.

For both the synthetic and real-world data, in many cases I am able to obtain a high-quality logistic regression posterior approximation using a coreset that is many orders of magnitude smaller than the full dataset – sometimes just a few hundred data points. Using such a small coreset represents a substantial reduction in the memory and computational requirements of the Bayesian inference algorithm that uses the coreset for posterior inference. I expect that the use of coresets could lead similar gains for other Bayesian models. Designing coreset algorithms for other widely-used models is an exciting direction for future research.

2.3.3 Implementation Details

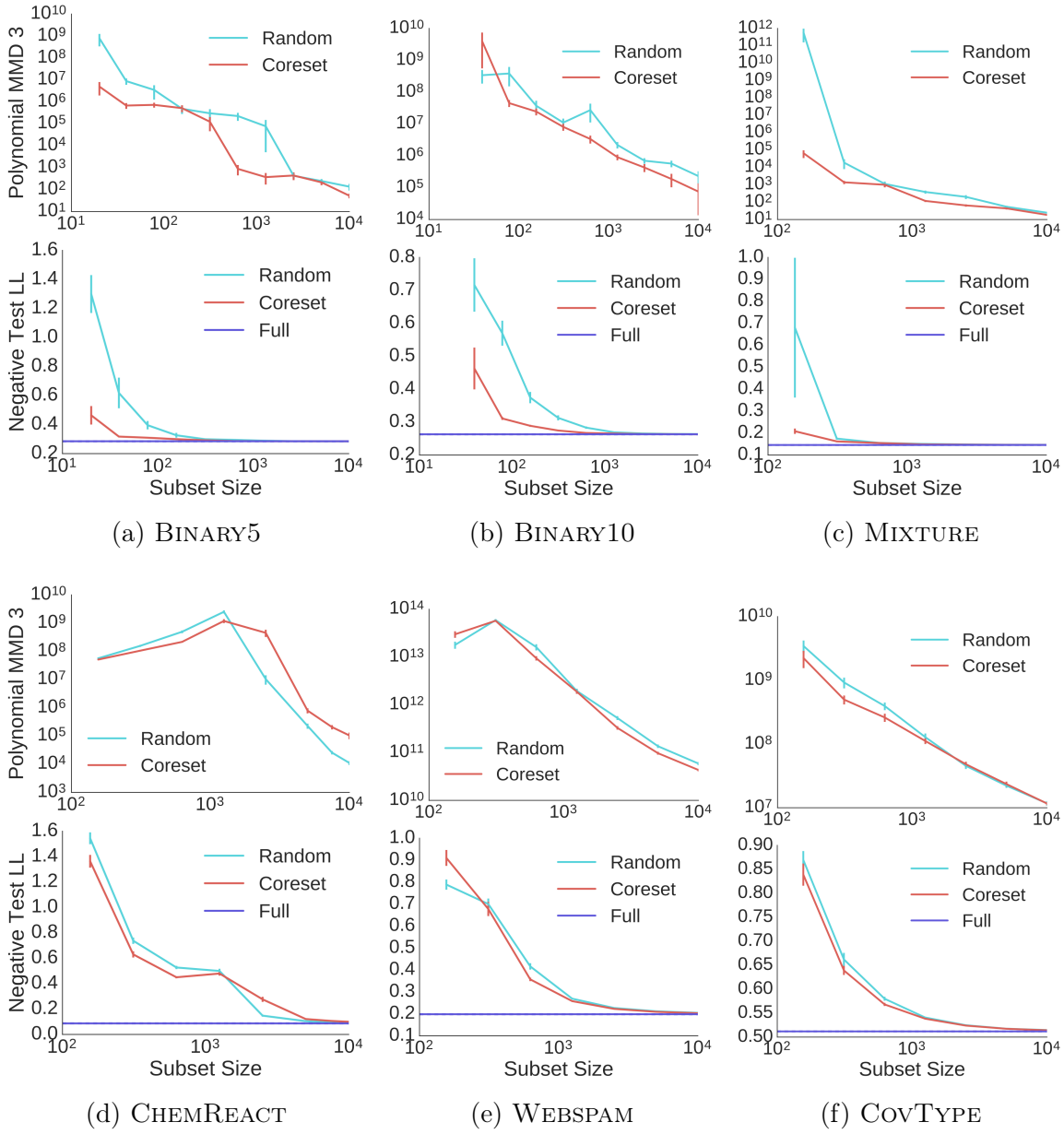
Implementing Algorithm 1. One time-consuming part of creating the coreset is calculating the adjusted centers $\bar{\mathbf{Z}}_{G,i}^{(-n)}$. I instead used the original centers Q_i. Since I use small k values and N in large, each cluster is large. Thus, the difference between $\bar{\mathbf{Z}}_{G,i}^{(-n)}$ and Q_i was negligible in practice, resulting at most a 1% change in the sensitivity while resulting in an order of magnitude speed-up in the algorithm. In order to speed up the clustering step, I selected a random subset of the data of size $L = \min(1000k, 0.025N)$ and ran the `sklearn` implementation of k-means++ to obtain k cluster centers. I then calculated the clustering and the normalized k-means score $\mathcal{I}$ for the full dataset. Notice that L is chosen to be independent of N as N becomes large but is never more than a construct fraction of the full dataset when N is small.[7] Thus, calculating a clustering only takes a small amount of time that is

[7]Note that I use data subsampling here *only* to choose the cluster centers. I still calculate sensitivity upper bounds across the entire data set and thereby are still able to capture rare but

comparable to the time required to run our implementation of Algorithm 1.

Posterior Inference Procedure. I used the adaptive Metropolis-adjusted Langevin algorithm [66, 118], where I adapted the overall step size and targeted an acceptance rate of 0.574 [116]. It T iterations were used in total, adaptation was done for the first $T/2$ iterations while the remaining iterations were used as approximate posterior samples. For the subsampling experiments, for a subsample size M, an approximate dataset $\tilde{\mathcal{D}}$ of size M was obtained either using random sampling or Algorithm 1. The dataset $\tilde{\mathcal{D}}$ was then fixed for the full MCMC run.

influential data patterns. Indeed, I expect influential data points to be far from cluster centers chosen either with or without subsampling, and I thereby expect to pick up these data points with high probability during the coreset sampling procedure in Algorithm 1.



(a) Binary5 (b) Binary10 (c) Mixture

(d) ChemReact (e) Webspam (f) CovType

Figure 2-2: Polynomial MMD and negative test log-likelihood of random sampling and the logistic regression coreset algorithm for synthetic and real data with varying subset sizes (lower is better for all plots). For the synthetic data, $N = 10^6$ total data points were used and 10^3 additional data points were generated for testing. For the real data, 2,500 (resp. 50,000 and 29,000) data points of the ChemReact (resp. Webspam and CovType) dataset were held out for testing. One standard deviation error bars were obtained by repeating each experiment 20 times.

Chapter 3

Polynomial Approximate Sufficient Statistics

In this chapter we propose to construct approximate sufficient statistics via a very simple *polynomial approximation* for generalized linear models. We therefore call our method *polynomial approximate sufficient statistics for generalized linear models* (PASS-GLM). PASS-GLM satisfies all of the criteria laid of above. It provides a posterior approximation with theoretical guarantees. It is scalable since is requires only a single pass over the data and can be applied to streaming and distributed data. And by increasing the number of approximate sufficient statistics, PASS-GLM can produce arbitrarily good approximations to the posterior.

We construct our novel polynomial approximation and specify our PASS-GLM algorithm in Section 3.1. We will see that streaming and distributed computation are trivial for our algorithm and do not compound error. In Section 3.2.1, we demonstrate finite-sample guarantees on the quality of the MAP estimate arising from our algorithm, with the maximum likelihood estimate (MLE) as a special case. In Section 3.2.2, we prove guarantees on the Wasserstein distance between the exact and approximate posteriors—and thereby bound both posterior-derived point estimates and uncertainty estimates. In Section 3.3, we demonstrate the efficacy of our approach in practice by focusing on logistic regression. We demonstrate experimentally that PASS-GLM can be scaled with almost no loss of efficiency to multi-core architectures. We show on a number of real-world datasets—including a large, high-dimensional advertising dataset (40 million examples with 20,000 dimensions)—that PASS-GLM provides an attractive trade-off between computation and accuracy.

Related work. The Laplace approximation [136] and variational methods with a Gaussian approximation family [74, 82] may be seen as polynomial (quadratic) approximations in the log-likelihood space. But we note that the VB variants suffer the issues described in Chapter 1. A Laplace approximation relies on a Taylor series expansion of the log-likelihood around the *maximum a posteriori* (MAP) solution, which requires first calculating the MAP—an expensive multi-pass optimization in the large-scale data setting. Neither Laplace nor VB offers the simplicity of sufficient statistics, including in streaming and distributed computations. The recent work of Stephanou et al. [131] is similar in spirit to ours, though they address a different

Algorithm 2 PASS-GLM inference

Require: data $\mathcal{D}$, GLM mapping function $\phi : \mathbb{R} \to \mathbb{R}$, degree M, polynomial basis $(\psi_m)_{m\in\mathbb{N}}$ with base measure ς

Calculate basis coefficients $b_m \leftarrow \int \phi\psi_m \mathrm{d}\varsigma$ using numerical integration for $m = 0, \ldots, M$
Calculate polynomial coefficients $b_m^{(M)} \leftarrow \sum_{k=m}^{M} \alpha_{k,m} b_m$ for $m = 0, \ldots, M$
for $\mathbf{k} \in \mathbb{N}^d$ with $\sum_j k_j \leq M$ **do**
    Initialize $t_{\mathbf{k}} \leftarrow 0$
end for
for $n = 1, \ldots, N$ **do** ▷ Can be done with any combination of batch, parallel, or streaming
    for $\mathbf{k} \in \mathbb{N}^d$ with $\sum_j k_j \leq M$ **do**
        Update $t_{\mathbf{k}} \leftarrow t_{\mathbf{k}} + (y_n \mathbf{x}_n)^{\mathbf{k}}$
    end for
end for
Form approximate log-likelihood $\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{\mathbf{k}\in\mathbb{N}^d : \sum_j k_j \leq m} \binom{m}{\mathbf{k}} b_m^{(M)} t_{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}}$
Use $\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta})$ to construct approximate posterior $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$

statistical problem: they construct sequential quantile estimates using Hermite polynomials. PASS-GLM has certain advantages over the coreset approach of the previous chapter. While coresets provide theoretical guarantees on the quality of inference via the model evidence, the resulting guarantees are better suited to approximate optimization and do not translate to guarantees on typical Bayesian desiderata, such as the accuracy of posterior mean and uncertainty estimates. Moreover, while coresets do admit streaming and distributed constructions, the approximation error is compounded across computations.

3.1 PASS-GLM

Recall from Section 1.2 that the log-likelihood in a GLM is given by

$$\log p(y_n \,|\, g^{-1}(\mathbf{x}_n \cdot \boldsymbol{\theta})) = \phi(y_n, \mathbf{x}_n \cdot \boldsymbol{\theta}),$$

where $\phi(y, s)$ is called the *mapping function*. Since exact sufficient statistics are not available for GLMs, we propose to construct approximate sufficient statistics by approximating ϕ with an order-M polynomial ϕ_M. We first illustrate our method in the logistic regression case, where $\log p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) = \phi_{\mathsf{logit}}(y_n \mathbf{x}_n \cdot \boldsymbol{\theta})$. Let $b_0^{(M)}, b_1^{(M)} \ldots, b_M^{(M)}$ be constants such that

$$\phi_{\mathsf{logit}}(s) \approx \phi_M(s) := \sum_{m=0}^{M} b_m^{(M)} s^m.$$

Let $\mathbf{v}^{\mathbf{k}} := \prod_{j=1}^{d} v_j^{k_j}$ for vectors $\mathbf{v}, \mathbf{k} \in \mathbb{R}^d$. Taking $s = y\mathbf{x} \cdot \boldsymbol{\theta}$, we obtain

$$\begin{aligned}\phi_{\mathsf{logit}}(y\mathbf{x} \cdot \boldsymbol{\theta}) \approx \phi_M(y\mathbf{x} \cdot \boldsymbol{\theta}) &= \textstyle\sum_{m=0}^{M} b_m^{(M)} (y\mathbf{x} \cdot \boldsymbol{\theta})^m = \sum_{m=0}^{M} b_m^{(M)} \sum_{\mathbf{k} \in \mathbb{N}^d \atop \sum_j k_j = m} \binom{m}{\mathbf{k}} (y\mathbf{x})^{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}} \\ &= \textstyle\sum_{m=0}^{M} \sum_{\mathbf{k} \in \mathbb{N}^d : \sum_j k_j = m} a(\mathbf{k}, m, M) (y\mathbf{x})^{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}},\end{aligned}$$

where $\binom{m}{\mathbf{k}}$ is the multinomial coefficient and $a(\mathbf{k}, m, M) := \binom{m}{\mathbf{k}} b_m^{(M)}$. Thus, ϕ_M is an M-degree polynomial approximation to $\phi_{\mathsf{logit}}(y\mathbf{x} \cdot \boldsymbol{\theta})$ with the $\binom{d+M}{d}$ monomials of degree at most M serving as sufficient statistics derived from $y\mathbf{x}$. Specifically, we have a exponential family model with

$$\mathbf{t}(y\mathbf{x}) = ([y\mathbf{x}]^{\mathbf{k}})_{\mathbf{k}} \qquad \text{and} \qquad \boldsymbol{\eta}(\boldsymbol{\theta}) = (a(\mathbf{k}, m, M) \boldsymbol{\theta}^{\mathbf{k}})_{\mathbf{k}},$$

where $\mathbf{k}$ is taken over all $\mathbf{k} \in \mathbb{N}^d$ such that $\sum_j k_j \leq M$. We next discuss the calculation of the $b_m^{(M)}$ and the choice of M.

We can generalize the setup just described to cover a wide range of GLMs by assuming the log-likelihood is of the form

$$\log p(y \,|\, \mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} y^{\alpha_k} \phi_{(k)}(y^{\beta_k} \mathbf{x} \cdot \boldsymbol{\theta} - a_k y),$$

where typically $\alpha_k, \beta_k, a_k \in \{0, 1\}$. We consider the $K = 1$ case and drop the k subscripts since the extension to $K > 1$ is trivial and serves only to introduce extra notational clutter. Letting $\phi_M(s) = \sum_{m=0}^{M} b_m^{(M)} s^m$ be the order M polynomial approximation to $\phi(s) = \phi_{(1)}(s)$, we have that

$$\begin{aligned}\log p(y \,|\, \mathbf{x}, \boldsymbol{\theta}) &\approx y^{\alpha} \phi_M(y^{\beta} \mathbf{x} \cdot \boldsymbol{\theta} - ay) \\ &= y^{\alpha} \sum_{m=0}^{M} b_m^{(M)} (y^{\beta} \mathbf{x} \cdot \boldsymbol{\theta} - ay)^m \\ &= y^{\alpha} \sum_{m=0}^{M} b_m^{(M)} \sum_{i=0}^{m} \binom{m}{i} (y^{\beta} \mathbf{x} \cdot \boldsymbol{\theta})^i (ay)^{m-i} \\ &= \sum_{i=0}^{M} (y^{\beta} \mathbf{x} \cdot \boldsymbol{\theta})^i y^{\alpha} \sum_{m=i}^{M} b_m^{(M)} \binom{m}{i} (ay)^{m-i} \\ &= \sum_{i=0}^{M} \sum_{\mathbf{k} \in \mathbb{N}^d \atop \sum_j k_j = i} a'(\mathbf{k}, i, M, y) (y^{\beta} \mathbf{x})^{\mathbf{k}} \boldsymbol{\theta}^{\mathbf{k}},\end{aligned}$$

where $a'(\mathbf{k}, \bar{k}, M, y) := y^{\alpha} \binom{\bar{k}}{\mathbf{k}} \sum_{m=i}^{M} b_m^{(M)} \binom{m}{\bar{k}} (ay)^{m-\bar{k}}$. Thus, we have an exponential family model with

$$\mathbf{t}(\mathbf{x}, y) = \Big(a'\Big(\mathbf{k}, \textstyle\sum_j k_j, M, y\Big) \mathbf{x}^{\mathbf{k}}\Big)_{\mathbf{k}} \qquad \text{and} \qquad \boldsymbol{\eta}(\boldsymbol{\theta}) = (\boldsymbol{\theta}^{\mathbf{k}})_{\mathbf{k}},$$

where $\mathbf{k}$ is taken over all $\mathbf{k} \in \mathbb{N}^d$ such that $\sum_j k_j \leq M$.

The following examples show how a variety of GLM models fit into our framework. Throughout, let $s = \mathbf{x}_n \cdot \boldsymbol{\theta}$.

Example 3.1.1 (Robust regression)**.** For robust regression, the log-likelihood is in the form $\delta(s - y)$, where δ is a choice of "distance" function. So we have $\phi_{(1)} = \delta$, $a_1 = 1$, $\alpha_1 = \beta_1 = 0$.

Example 3.1.2 (Poisson regression)**.** For Poisson regression the log-likelihood is $ys - e^s$, so $\phi_{(1)}(s) = s$, $\phi_{(2)}(s) = -e^s$, $\alpha_1 = 1$, and $\beta_1 = a_1 = \alpha_2 = \beta_2 = a_2 = 0$.

Example 3.1.3 (Gamma regression)**.** For gamma regression, the log-likelihood is $-\nu s - \nu y e^{-s} + c(y, \nu)$ if using the log link, where ν is a scale parameter. We can ignore the $c(y, \nu)$ term since it does not depend on $\boldsymbol{\theta}$. Thus, $\phi_{(1)}(s) = -\nu s$, $\phi_{(2)}(s) = -\nu e^{-s}$, $\alpha_2 = 1$, and $\beta_1 = a_1 = \alpha_1 = \beta_2 = a_2 = 0$.

Example 3.1.4 (Probit regression)**.** For probit regression, $\mathcal{Y} = \{0, 1\}$, and the log-likelihood is

$$\begin{cases} \ln(1 - \Phi(s)) & y = 0 \\ \ln(\Phi(s)) & y = 1 \end{cases},$$

where Φ denotes the standard normal CDF. Thus, $\phi_{(1)}(s) = \ln(1 - \Phi(s))$, $\phi_{(2)}(s) = \ln(\Phi(s)) - \ln(1 - \Phi(s))$, $\alpha_2 = 1$, and $\beta_1 = a_1 = \alpha_1 = \beta_2 = a_2 = 0$.

Choosing the polynomial approximation. To calculate the coefficients $b_m^{(M)}$, we choose a polynomial basis $(\psi_m)_{m \in \mathbb{N}}$ orthogonal with respect to a base measure ς, where ψ_m is degree m [133]. That is,

$$\psi_m(s) = \textstyle\sum_{j=0}^{m} \alpha_{m,j} s^j$$

for some $\alpha_{m,j}$, and

$$\int \psi_m \psi_{m'} \mathrm{d}\varsigma = \delta_{mm'},$$

where $\delta_{mm'} = 1$ if $m = m'$ and zero otherwise. If $b_m := \int \phi \psi_m \mathrm{d}\varsigma$, then

$$\phi(s) = \textstyle\sum_{m=0}^{\infty} b_m \psi_m(s)$$

and the approximation $\phi_M(s) = \sum_{m=0}^{M} b_m \psi_m(s)$. Conclude that $b_m^{(M)} = \sum_{k=m}^{M} \alpha_{k,m} b_m$. The complete PASS-GLM framework appears in Algorithm 2.

Choices for the orthogonal polynomial basis include Chebyshev, Hermite, Leguerre, and Legendre polynomials [133]. We choose Chebyshev polynomials since they provide a uniform quality guarantee on a finite interval, e.g. $[-R, R]$ for some $R > 0$ in what follows. If ϕ is smooth, the choice of Chebyshev polynomials (scaled appropriately, along with the base measure ς, based on the choice of R) yields error exponentially small in M: $\sup_{s \in [-R,R]} |\phi(s) - \phi_M(s)| \leq C\rho^M$ for some $0 < \rho < 1$ and $C > 0$

[93]. We show in Appendix B.1 that the error in the approximate derivative ϕ'_M is also exponentially small in M: $\sup_{s\in[-R,R]} |\phi'(s) - \phi'_M(s)| \leq C'\rho^M$, where $C' > C$.

Choosing the polynomial degree. For fixed d, the number of monomials is $O(M^d)$ while for fixed M the number of monomials is $O(d^M)$. The number of approximate sufficient statistics can remain manageable when either M or d is small but becomes unwieldy if M and d are both large. Since our experiments (Section 3.3) generally have large d, we focus on the small M case here.

In our experiments we further focus on the choice of logistic regression as a particularly popular GLM example with $p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) = \phi_{\mathsf{logit}}(y_n \mathbf{x}_n \cdot \boldsymbol{\theta})$, where

$$\phi_{\mathsf{logit}}(s) := -\log(1 + e^{-s}).$$

In general, the smallest and therefore most compelling choice of M *a priori* is 2, and we demonstrate the reasonableness of this choice empirically in Section 3.3 for a number of large-scale data analyses. In addition, in the logistic regression case, $M = 6$ is the next usable choice beyond $M = 2$. This is because $b^{(M)}_{2k+1} = 0$ for all integer $k \geq 1$ with $2k + 1 \leq M$. So any approximation beyond $M = 2$ must have $M \geq 4$. Also, $b^{(M)}_{4k} > 0$ for all integers $k \geq 1$ with $4k \leq M$. So choosing $M = 4k$, $k \geq 1$, leads to a pathological approximation of ϕ_{logit} where the log-likelihood can be made arbitrarily large by taking $\|\boldsymbol{\theta}\|_2 \to \infty$. Thus, a reasonable polynomial approximation for logistic regression requires $M = 2 + 4k$, $k \geq 0$. We have discussed the relative drawbacks of other popular quadratic approximations, including the Laplace approximation and variational methods, at the beginning of the chapter.

3.2 Theoretical Results

We next establish quality guarantees for PASS-GLM. We first provide finite-sample and asymptotic guarantees on the MAP (point estimate) solution, and therefore on the MLE, in Section 3.2.1. We then provide guarantees on the Wasserstein distance between the approximate and exact posteriors, and show these bounds translate into bounds on the quality of posterior mean and uncertainty estimates, in Section 3.2.2. See Appendix B.2 for extended results, further discussion, and all proofs.

3.2.1 MAP approximation

In Appendix B.2, we state and prove Theorem B.2.1, which provides guarantees on the quality of the MAP estimate for an arbitrary approximation $\tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta})$ to the log-likelihood $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$. The approximate MAP (i.e., the MAP under $\tilde{\mathcal{L}}_{\mathcal{D}}$) is (cf. Eq. (1.1))

$$\tilde{\boldsymbol{\theta}}_{\mathrm{MAP}} := \arg\max_{\boldsymbol{\theta}\in\Theta} \log \pi_0(\boldsymbol{\theta}) + \tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}).$$

Roughly, we find in Theorem B.2.1 that the error in the MAP estimate naturally depends on the error of the approximate log-likelihood as well as the peakedness of the posterior near the MAP. In the latter case, if $\log \pi_{\mathcal{D}}$ is very flat, then even a small

error from using $\tilde{\mathcal{L}}_{\mathcal{D}}$ in place of $\mathcal{L}_{\mathcal{D}}$ could lead to a large error in the approximate MAP solution. We measure the peakedness of the distribution in terms of the strong convexity constant[1] of $-\log \pi_{\mathcal{D}}$ near $\boldsymbol{\theta}_{\text{MAP}}$.

We apply Theorem B.2.1 to PASS-GLM for logistic regression and robust regression. We require the assumption that

$$\phi_M(t) \leq \phi(t) \ \forall t \notin [-R, R], \tag{3.1}$$

which in the cases of logistic regression and smoothed Huber regression, we conjecture holds for $M = 2 + 4k$, $k \in \mathbb{N}$. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_2$ denotes its spectral norm.

Corollary 3.2.1. *For the logistic regression model, assume that*

$$\|(\nabla^2 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{MAP}))^{-1}\|_2 \leq cd/N$$

for some constant $c > 0$ and that $\|\mathbf{x}_n\|_2 \leq 1$ for all $n = 1, \ldots, N$. Let ϕ_M be the order-M Chebyshev approximation to ϕ_{logit} on $[-R, R]$ such that Eq. (3.1) holds. Let $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ denote the posterior approximation obtained by using ϕ_M with a log-concave prior. Then there exist numbers $r = r(R) > 1$, $\varepsilon = \varepsilon(M) = O(r^{-M})$, and $\alpha^* \geq \frac{27}{\varepsilon d^3 c^3 + 54}$, such that if $R - \|\boldsymbol{\theta}_{MAP}\|_2 \geq 2\sqrt{\frac{cd\varepsilon}{\alpha^*}}$, then

$$\|\boldsymbol{\theta}_{MAP} - \tilde{\boldsymbol{\theta}}_{MAP}\|_2^2 \leq \frac{4cd\varepsilon}{\alpha^*} \leq \frac{4}{27} c^4 d^4 \varepsilon^2 + 8cd\varepsilon.$$

The main takeaways from Corollary 3.2.1 are that (1) the error decreases exponentially in M thanks to the ε term, (2) the error does not depend on the amount of data, and (3) in order for the bound on the approximate MAP solution to hold, the norm of the true MAP solution must be sufficiently smaller than R.

Remark 3.2.2. Some intuition for the assumption on the Hessian of $\mathcal{L}_{\mathcal{D}}$, i.e., $\nabla^2 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \phi''_{\mathsf{logit}}(y_n \mathbf{x}_n \cdot \boldsymbol{\theta}) \mathbf{x}_n \mathbf{x}_n^\top$, is as follows. Typically for $\boldsymbol{\theta}$ near $\boldsymbol{\theta}_{\text{MAP}}$, the minimum eigenvalue of $\nabla^2 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ is at least $N/(cd)$ for some $c > 0$. The minimum eigenvalue condition in Corollary 3.2.1 holds if, for example, a constant fraction of the data satisfies $0 < b \leq \|x_n\|_2 \leq B < \infty$ and that subset of the data does not lie too close to any $(d-1)$-dimensional hyperplane. This condition essentially requires the data not to be degenerate and is similar to ones used to show asymptotic consistency of logistic regression [137, Ex. 5.40].

The approximate MAP error bound in the robust regression case using, for example, the smoothed Huber loss (Example 3.1.1), is quite similar to the logistic regression result.

Corollary 3.2.3. *For robust regression with smoothed Huber loss, assume that a constant fraction of the data satisfies $|\mathbf{x}_n \cdot \boldsymbol{\theta}_{MAP} - y_n| \leq b/2$ and that $\|\mathbf{x}_n\|_2 \leq 1$ for all $n = 1, \ldots, N$. Let ϕ_M be the order M Chebyshev approximation to ϕ_{Huber}*

[1] Recall that a twice-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is ϱ-strongly convex at $\boldsymbol{\theta}$ if the minimum eigenvalue of the Hessian of f evaluated at $\boldsymbol{\theta}$ is at least $\varrho > 0$.

on $[-R, R]$ *such that Eq.* (3.1) *holds. Let* $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ *denote the posterior approximation obtained by using* ϕ_M *with a log-concave prior. Then if* $R \gg \|\boldsymbol{\theta}_{MAP}\|_2$*, there exists* $r > 1$ *such that for* M *sufficiently large,* $\|\boldsymbol{\theta}_{MAP} - \tilde{\boldsymbol{\theta}}_{MAP}\|_2^2 = O(dr^{-M})$.

3.2.2 Posterior approximation

We next establish guarantees on how close the approximate and exact posteriors are in *Wasserstein distance*, $d_{\mathcal{W}}$. For distributions P and Q on $\mathbb{R}^d$,

$$d_{\mathcal{W}}(P, Q) := \sup_{f:\|f\|_L \leq 1} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right|,$$

where $\|f\|_L$ denotes the Lipschitz constant of f.[2] This choice of distance is particularly useful since, if $d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) \leq \delta$, then $\tilde{\pi}_{\mathcal{D}}$ can be used to estimate any function with bounded gradient with error at most $\delta \sup_{\mathbf{w}} \|\nabla f(\mathbf{w})\|_2$. Wasserstein error bounds therefore give bounds on the mean estimates (corresponding to $f(\boldsymbol{\theta}) = \theta_i$) as well as uncertainty estimates such as mean absolute deviation (corresponding to $f(\boldsymbol{\theta}) = |\bar{\theta}_i - \theta_i|$, where $\bar{\theta}_i$ is the expected value of θ_i).

Our general result (Theorem B.2.3) is stated and proved in Appendix B.2. Similar to Theorem B.2.1, the result primarily depends on the peakedness of the approximate posterior and the error of the approximate gradients. If the gradients are poorly approximated then the error can be large while if the (approximate) posterior is flat then even small gradient errors could lead to large shifts in expected values of the parameters and hence large Wasserstein error.

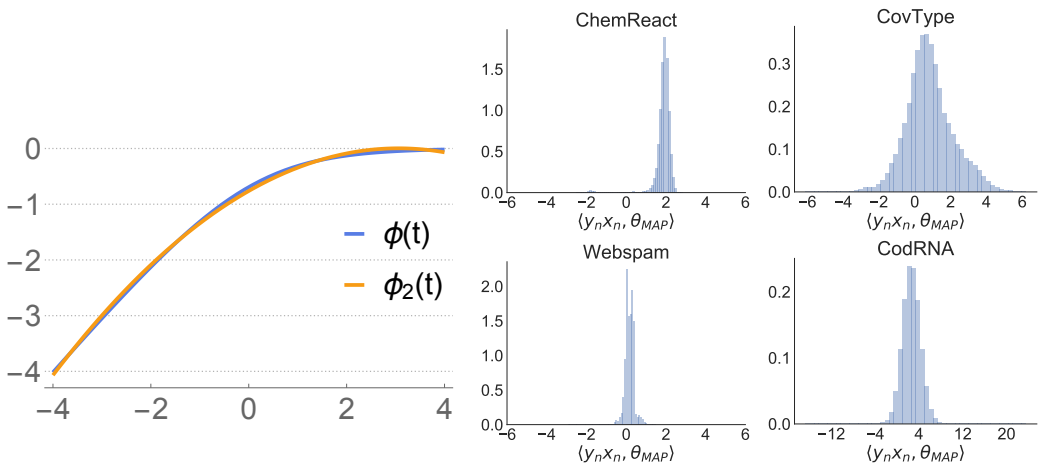
We apply Theorem B.2.3 to PASS-GLM for logistic regression and Poisson regression. We give simplified versions of these corollaries in the main text and defer the more detailed versions to Appendix B.2. For logistic regression we assume $M = 2$ and $\Theta = \mathbb{R}^d$ since this is the setting we use for our experiments. The result is similar in spirit to Corollary 3.2.1, though more straightforward since $M = 2$. Critically, we see in this result how having small error depends on $|y_n \mathbf{x}_n \cdot \bar{\boldsymbol{\theta}}| \leq R$ with high probability. Otherwise the second term in the bound will be large.

Corollary 3.2.4. *Let* ϕ_2 *be the second-order Chebyshev approximation to* ϕ_{logit} *on* $[-R, R]$ *and let* $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \tilde{\boldsymbol{\theta}}_{MAP}, \tilde{\boldsymbol{\Sigma}})$ *denote the posterior approximation obtained by using* ϕ_2 *with a Gaussian prior* $\pi_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$*. Let* $\bar{\boldsymbol{\theta}} := \int \boldsymbol{\theta} \pi_{\mathcal{D}}(\mathrm{d}\boldsymbol{\theta})$*, let* $\delta_1 := N^{-1} \sum_{n=1}^N \langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}} \rangle$*, and let* σ_1 *be the subgaussianity constant of the random variable* $\langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}} \rangle - \delta_1$*, where* $n \sim \mathsf{Unif}\{1, \ldots, N\}$*. Assume that* $|\delta_1| \leq R$*, that* $\|\tilde{\boldsymbol{\Sigma}}\|_2 \leq cd/N$*, and that* $\|\mathbf{x}_n\|_2 \leq 1$ *for all* $n = 1, \ldots, N$*. Then with* $\sigma_0^2 := \|\boldsymbol{\Sigma}_0\|_2$*, we have*

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) = O\Big(dR^4 + d\sigma_0 \exp\Big(\sigma_1^2 \sigma_0^{-2} - \sqrt{2}\, \sigma_0^{-1}(R - |\delta_1|)\Big)\Big).$$

The main takeaway from Corollary 3.2.4 is that if (a) for most n, $|\langle \mathbf{x}_n, \bar{\boldsymbol{\theta}} \rangle| < R$, so that ϕ_2 is a good approximation to ϕ_{logit}, and (b) the approximate posterior

[2]The Lipschitz constant of function $f : \mathbb{R}^d \to \mathbb{R}$ is $\|f\|_L := \sup_{\mathbf{v}, \mathbf{w} \in \mathbb{R}^d} \frac{\|\phi(\mathbf{v}) - \phi(\mathbf{w})\|_2}{\|\mathbf{v} - \mathbf{w}\|_2}$.



(a) (b)

Figure 3-1: Validating the use of PASS-GLM with $M = 2$. **(a)** The second-order Chebyshev approximation to $\phi = \phi_{\text{logit}}$ on $[-4, 4]$ is very accurate, with error of at most 0.069. **(b)** For a variety of datasets, the inner products $\langle y_n \mathbf{x}_n, \boldsymbol{\theta}_{\text{MAP}} \rangle$ are mostly in the range of $[-4, 4]$.

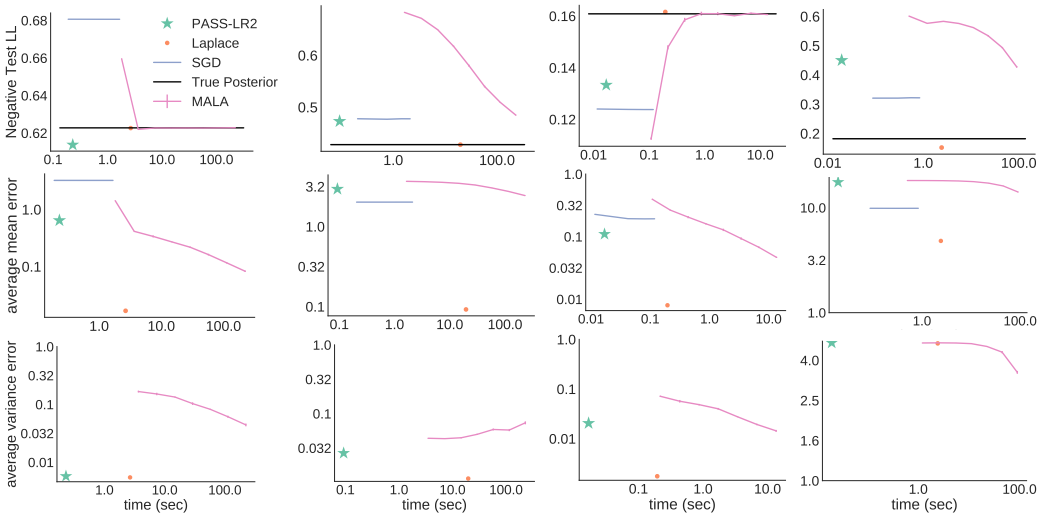
concentrates quickly, then we get a high-quality approximate posterior. This result matches up with the experimental results (see Section 3.3 for further discussion).

For Poisson regression, we return to the case of general M. Recall that in the Poisson regression model that the expectation of y_n is $\mu = e^{\mathbf{x}_n \cdot \boldsymbol{\theta}}$. If y_n is bounded and has non-trivial probability of being greater than zero, we lose little by restricting $\mathbf{x}_n \cdot \boldsymbol{\theta}$ to be bounded. Thus, we will assume that the parameter space is bounded. As in Corollaries 3.2.1 and 3.2.3, the error is exponentially small in M and, as long as $\| \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top \|_2$ grows linearly in N, does not depend on the amount of data.

Corollary 3.2.5. *Let $f_M(s)$ be the order-M Chebyshev approximation to e^t on the interval $[-R, R]$, and let $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ denote the posterior approximation obtained by using the approximation $\log \tilde{p}(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) := y_n \mathbf{x}_n \cdot \boldsymbol{\theta} - f_M(\mathbf{x}_n \cdot \boldsymbol{\theta}) - \log y_n!$ with a log-concave prior on $\Theta = \mathbb{B}_R(\mathbf{0})$. If $\inf_{s \in [-R,R]} f_M''(s) \geq \tilde{\varrho} > 0$, $\| \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top \|_2 = \Omega(N/d)$, and $\|\mathbf{x}_n\|_2 \leq 1$ for all $n = 1, \ldots, N$, then*

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) = O\big(d \tilde{\varrho}^{-1} M^2 e^R 2^{-M}\big).$$

Note that although $\tilde{\varrho}^{-1}$ does depend on R and M, as M becomes large it converges to e^R. Observe that if we truncate a prior on $\mathbb{R}^d$ to be on $\mathbb{B}_R(\mathbf{0})$, by making R and M sufficiently large, the Wasserstein distance between $\pi_{\mathcal{D}}$ and the PASS-GLM posterior approximation $\tilde{\pi}_{\mathcal{D}}$ can be made arbitarily small. Similar results could be shown for other GLM likelihoods.



(a) WEBSPAM (b) COVTYPE (c) CHEMREACT (d) CODRNA

Figure 3-2: Batch inference results. In all metrics smaller is better.

3.3 Experiments

In our experiments, we focus on logistic regression, a particularly popular GLM example.[3] As discussed in Section 3.1, we choose $M = 2$ and call our algorithm PASS-LR2. Empirically, we observe that $M = 2$ offers a high-quality approximation of ϕ on the interval $[-4, 4]$ (Fig. 3-1a). In fact $\sup_{s\in[-4,4]} |\phi_2(s) - \phi(s)| < 0.069$. Moreover, we observe that for many datasets, the inner products $y_n \mathbf{x}_n \cdot \boldsymbol{\theta}_{\text{MAP}}$ tend to be concentrated within $[-4, 4]$, and therefore a high-quality approximation on this range is sufficient for our analysis. In particular, Fig. 3-1b shows histograms of $y_n \mathbf{x}_n \cdot \boldsymbol{\theta}_{\text{MAP}}$ for four datasets from our experiments. In all but one case, over 98% of the data points satisfy $|y_n \mathbf{x}_n \cdot \boldsymbol{\theta}_{\text{MAP}}| \leq 4$. In the remaining dataset (CODRNA), only ∼80% of the data satisfy this condition, and this is the dataset for which PASS-LR2 performed most poorly (cf. Corollary 3.2.4).

3.3.1 Large dataset experiments

In order to compare PASS-LR2 to other approximate Bayesian methods, we first restrict our attention to datasets with fewer than 1 million data points. We compare to the Laplace approximation and the adaptive Metropolis-adjusted Langevin algorithm (MALA). We also compare to stochastic gradient descent (SGD) although SGD provides only a point estimate and no approximate posterior. In all experiments, no method performs as well as PASS-LR2 given the same (or less) running time.

Datasets. The CHEMREACT dataset consists of $N = 26{,}733$ chemicals, each with $d = 100$ properties. The goal is to predict whether each chemical is reactive. The

[3]Code is available at `https://bitbucket.org/jhhuggins/pass-glm`.

WEBSPAM corpus consists of $N = 350{,}000$ web pages and the covariates consist of the $d = 127$ features that each appear in at least 25 documents. The cover type (COVTYPE) dataset consists of $N = 581{,}012$ cartographic observations with $d = 54$ features. The task is to predict the type of trees that are present at each observation location. The CODRNA dataset consists of $N = 488{,}565$ and $d = 8$ RNA-related features. The task is to predict whether the sequences are non-coding RNA.

Fig. 3-2 shows average errors of the posterior mean and variance estimates as well as negative test log-likelihood for each method versus the time required to run the method. SGD was run for between 1 and 20 epochs. The true posterior was estimated by running three chains of adaptive MALA for 50,000 iterations, which produced Gelman-Rubin statistics well below 1.1 for all datasets.

Speed. For all four datasets, PASS-LR2 was an order of magnitude faster than SGD and 2–3 orders of magnitude faster than the Laplace approximation.

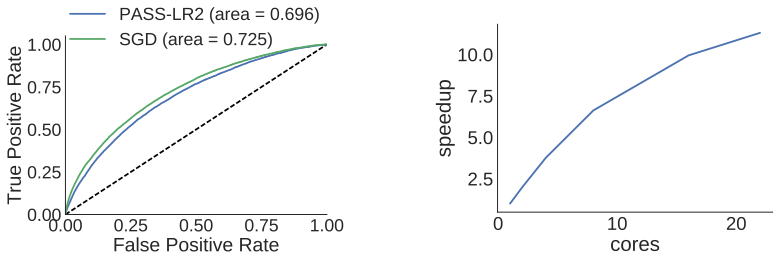
Mean and variance estimates. For CHEMREACT, WEBSPAM, and COVTYPE, PASS-LR2 was superior to or competitive with SGD, with MALA taking 10–100x longer to produce comparable results. Laplace again outperformed all other methods. Critically, on all datasets the PASS-LR2 variance estimates were competitive with Laplace and MALA.

Test log-likelihood. For CHEMREACT and WEBSPAM, PASS-LR2 produced results competitive with all other methods. MALA took 10–100x longer to produce comparable results. For COVTYPE, PASS-LR2 was competitive with SGD but took a tenth of the time, and MALA took 1000x longer for comparable results. Laplace outperformed all other methods, but was orders of magnitude slower than PASS-LR2. CODRNA was the only dataset where PASS-LR2 performed poorly. However, this performance was expected based on the $y_n \mathbf{x}_n \cdot \boldsymbol{\theta}_{\text{MAP}}$ histogram (Fig. 3-1a).

3.3.2 Very large dataset experiments using streaming and distributed PASS-GLM

We next test PASS-LR2, which is streaming without requiring any modifications, on a subset of 40 million data points from the Criteo terabyte ad click prediction dataset (CRITEO). The covariates are 13 integer-valued features and 26 categorical features. After one-hot encoding, on the subset of the data we considered, $d \approx 3$ million. For tractability we used sparse random projections [83] to reduce the dimensionality to 20,000. At this scale, comparing to the other fully Bayesian methods from Section 3.3.1 was infeasible; we compare only to the predictions and point estimates from SGD. PASS-LR2 performs slightly worse than SGD in AUC (Fig. 3-3a), but outperforms SGD in negative test log-likelihood (0.07 for SGD, 0.045 for PASS-LR2). Since PASS-LR2 estimates a full covariance, it was about 10x slower than SGD. A promising approach to speeding up and reducing memory usage of PASS-LR2 would be to use a low-rank approximation to the second-order moments.

To validate the efficiency of distributed computation with PASS-LR2, we compared running times on 6M examples with dimensionality reduced to 1,000 when using 1–22 cores. As shown in Fig. 3-3b, the speed-up is close to optimal: K cores produces a



(a) (b)

Figure 3-3: **(a)** ROC curves for streaming inference on 40 million CRITEO data points. SGD and PASS-LR2 had negative test log-likelihoods of, respectively, 0.07 and 0.045. **(b)** Cores vs. speedup (compared to one core) for parallelization experiment on 6 million examples from the CRITEO dataset.

speedup of about $K/2$ (baseline 3 minutes using 1 core). We used Ray to implement the distributed version of PASS-LR2 [104].[4]

3.4 Discussion

We have presented PASS-GLM, a novel framework for scalable parameter estimation and Bayesian inference in generalized linear models. Our theoretical results provide guarantees on the quality of point estimates as well as approximate posteriors derived from PASS-GLM. We validated our approach empirically with logistic regression and a quadratic approximation. We showed competitive performance on a variety of real-world data, scaling to 40 million examples with 20,000 covariates, and trivial distributed computation with no compounding of approximation error.

There a number of important directions for future work. The first is to use randomization methods along the lines of random projections and random feature mappings [83, 114] to scale to larger M and d. We conjecture that the use of randomization will allow experimentation with other GLMs for which quadratic approximations are insufficient.

[4] `https://github.com/ray-project/ray`

Chapter 4

Approximate Diffusions

So far we have examined a number of approaches to replacing an exact log-likelihood $\mathcal{L}_{\mathcal{D}}$ that requires $\Omega(N)$ time to evaluate with an approximation $\tilde{\mathcal{L}}_{\mathcal{D}}$ that requires $o(N)$ time to evaluate. As we have seen, this is a practical and general-purpose approach to scaling Bayesian inference to very large datasets. In this chapter we step back to develop some general-purpose theoretical tools for evaluating the quality of an approximate posterior $\tilde{\pi}_{\mathcal{D}}$ derived from a likelihood approximation $\tilde{\mathcal{L}}_{\mathcal{D}}$. These results are used Appendix B to prove the accuracy guarantees of the PASS-GLM methodology developed in Chapter 3. We provide further applications of our results below as well.

We tackle the question of how close $\tilde{\pi}_{\mathcal{D}}$ is to the true target distribution $\pi_{\mathcal{D}}$ from the perspective of Markov chains and their continuous-time counterpart, diffusion processes — these stochastic processes are ubiquitous in machine learning and statistics, forming a core component of the inference and modeling toolkit. Since faster convergence enables more efficient sampling and inference, a large and fruitful literature has investigated how quickly these stochastic processes converge to equilibrium. The large-data setting, however, leads us to develop stochastic processes that can be simulated from efficiently — by replacing $\mathcal{L}_{\mathcal{D}}$ with $\tilde{\mathcal{L}}_{\mathcal{D}}$ — while remaining accurate (as measured by fast convergence to the target distribution). Consider an MCMC algorithm that employs $\tilde{\mathcal{L}}_{\mathcal{D}}$. A central question of both theoretical and practical importance is how to quantify the deviation between the equilibrium distribution that the approximate chain converges to and the desired distribution targeted by the original chain. Moreover, we would like to understand, given a fixed computational budget, how to design approximate chains that generate the most accurate samples.

Our contributions. In this chaper, we develop general results to quantify the accuracy of approximate diffusions and Markov chains and apply these results to characterize the computational–statistical trade-off in specific algorithms. Our starting point is continuous-time diffusion processes because these are the objects which are discretized to construct many sampling algorithms, such as the unadjusted and Metropolis-adjusted Langevin algorithms [118] and Hamiltonian Monte Carlo [103]. Given two diffusion processes, we bound the deviation in their equilibrium distributions in terms of the deviation in their drifts and the rate at which the diffusion mixes (Theorem 4.2.1). Moreover, we show that this bound is tight for certain Gaussian

target distributions. These characterizations of diffusions are novel and are likely of more general interest beyond the inferential settings we consider. We apply our general results to derive a finite-sample error bound on a specific unadjusted Langevin dynamics algorithm (Theorem 4.4.1). Under computational constraint, the relevant trade-off here is between computing the exact log-likelihood gradient for few iterations or computing an approximate gradient for more iterations. We characterize settings where the approximate Langevin dynamics produce more accurate samples from the true posterior. We illustrate our analyses with simulation results. In addition, we apply our approach to quantify the accuracy of approximations to the zig-zag process, a recently-developed non-reversible sampling scheme.

Chapter outline. We introduce the basics of diffusion processes and other preliminaries in Section 4.1. Section 4.2 discusses the main results on bounding the error between an exact and perturbed diffusion. We describe the main ideas behind our analyses in Section 4.3; all the detailed proofs are deferred to the Appendix C. Section 4.4 applies the main results to derive finite sample error bounds for unadjusted Langevin dynamics and illustrates the computational–statistical trade-off. Section 4.5 extends our main results to quantify the accuracy of approximate piecewise deterministic Markov processes, including the zig-zag process. Numerical experiments to complement the theory are provided in Section 5.5. We conclude with a discussion of how our results connect to the relevant literature and suggest directions for further research.

4.1 Diffusions and preliminaries

In this chapter we adopt different notation than the other chapters to follow the standard notation used in the stochastic processes literature. Let $\mathcal{X} = \mathbb{R}^d$ be the parameter space and let π be a probability density over $\mathbb{R}^d$ (e.g. it can be the posterior distribution of some latent parameters given data). A Langevin diffusion is characterized by the stochastic differential equation

$$\mathrm{d}X_t = \nabla \log \pi(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

where $X_t \in \mathbb{R}^d$ and W_t is a standard Brownian motion. The intuition is that X_t undergoes a biased random walk in which it is more likely to move in directions that increase the density. Under appropriate regularity conditions, as $t \to \infty$, the distribution of X_t converges to π. Thus, simulating the Langevin diffusion provides a powerful framework to sample from the target π. To implement such a simulation, we need to discretize the continuous diffusion into finite-width time steps. For our main results, we focus on analyzing properties of the underlying diffusion processes. This allows us to obtain general results which are independent of any particular discretization scheme.

Beyond Langevin dynamics, more general diffusions can take the form

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \tag{4.1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ is the drift and is not necessarily the gradient of some log-density.[1] Furthermore, we can analyze other continuous-time Markov processes such as piecewise deterministic Markov processes (PDMPs). For example, Hamiltonian Monte Carlo [103] can be viewed as approximating a PDMP and the zig-zag process is a recently-developed non-reversible PDMP designed for large Bayesian inference (see Section 4.5).

In many large-data settings, computing the drift $b(X_t)$ in Eq. (4.1) can be expensive; for example, computing $b(X_t) = \nabla \log \pi(X_t)$ requires using all of the data and may involve evaluating a complex function such as a differential equation solver. Hence, it is desirable to replace b with an approximation $\tilde{b}$. Such an approximation changes the underlying diffusion process to

$$\mathrm{d}\tilde{X}_t = \tilde{b}(\tilde{X}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{W}_t, \tag{4.2}$$

where $\tilde{W}_t$ is a standard Brownian motion. In order to understand the quality of different approximations, we need to quantify how the equilibrium distribution of Eq. (4.1) differs from the equilibrium distribution of Eq. (4.2). We use the standard *Wasserstein metric* to measure this distance. We recall the definition previously given in Chapter 3:

Definition. The *Wasserstein distance* between distributions π and $\tilde{\pi}$ is

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) = \sup_{\phi \in C_L(\mathbb{R}^d)} |E_\pi[\phi] - E_{\tilde{\pi}}[\phi]|,$$

where $C_L(\mathbb{R}^d)$ is the set of continuous functions $\phi : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant $\|\phi\|_L \le 1$.[2]

The distance between π and $\tilde{\pi}$ should depend on how good the drift approximation is, which can be quantified by $\|b - \tilde{b}\|_2$.[3] It is also natural for the distance to depend on how quickly the original diffusion with drift b mixes, since the faster it mixes, the less time there is for the error to accumulate. Geometric contractivity is a useful property which quantifies fast-mixing diffusions. For each $x \in \mathbb{R}^d$, let $\mu_{x,t}$ denote the law of $X_t \,|\, X_0 = x$.

Assumption 4.A (**Geometric contractivity**). *There exist constants $C > 0$ and $0 < \rho < 1$ such that for all $x, x' \in \mathbb{R}^d$,*

$$d_{\mathcal{W}}(\mu_{x,t}, \mu_{x',t}) \le C\|x - x'\|_2 \rho^t.$$

[1] All of our results can be extended to more general diffusions on a domain $\mathcal{X} \subseteq \mathbb{R}^d$, $\mathrm{d}X_t = b(X_t) + \Sigma\,\mathrm{d}W_t - n_t L(\mathrm{d}t)$, where Σ is the covariance of the Brownian motion, and $n_t L$ captures the reflection forces at the boundary $\partial\mathcal{X}$. To keep the exposition simple, we focus on the simpler diffusion in the main text.

[2] Recall that the Lipschitz constant of function $\phi : \mathbb{R}^d \to \mathbb{R}$ is $\|\phi\|_L := \sup_{x,y \in \mathbb{R}^d} \frac{\|\phi(x) - \phi(y)\|_2}{\|x - y\|_2}$.

[3] For a function $\phi : \mathbb{R}^n \to \mathbb{R}^m$, define $\|\phi\|_2 := \sup_{x \in \mathbb{R}^n} \|\phi(x)\|_2$.

Geometric contractivity holds in many natural settings. Recall that a twice continuously-differentiable function ϕ is *k-strongly concave* if for all $x, x' \in \mathbb{R}^d$

$$(\nabla\phi(x) - \nabla\phi(x')) \cdot (x - x') \leq -k\|x - x'\|_2^2. \tag{4.3}$$

When $b = \nabla \log \pi$ and $\log \pi$ is k-strongly concave, the diffusion is exponentially ergodic with $C = 1$ and $\rho = e^{-k}$ (this can be shown using standard coupling arguments [21]). In fact, exponential contractivity also follows if Eq. (4.3) is satisfied when x and x' are far apart and $\log \pi$ has "bounded convexity" when x and x' are close together [46]. Alternatively, Hairer et al. [67] provides a Lyapunov function-based approach to proving exponential contractivity.

To ensure that the diffusion and the approximate diffusion are well-behaved, we impose some standard regularity properties.

Assumption 4.B (**Regularity conditions**)**.** *Let π and $\tilde{\pi}$ denote the stationary distributions of the diffusions in Eq.* (4.1) *and Eq.* (4.2)*, respectively.*

1. *The target density satisfies $\pi \in C^2(\mathbb{R}^d, \mathbb{R})$ and $\int x^2 \pi(\mathrm{d}x) < \infty$. The drift satisfies $b \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and $\|b\|_L < \infty$.*
2. *The approximate drift satisfies $\tilde{b} \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ and $\|\tilde{b}\|_L < \infty$.*
3. *If a function $\phi \in C(\mathbb{R}^d, \mathbb{R})$ is π-integrable then it is $\tilde{\pi}$-integrable.*

Here $C^k(\mathbb{R}^m, \mathbb{R}^n)$ denotes the set of k-times continuously differentiable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$ and $C(\mathbb{R}^m, \mathbb{R}^n)$ is the set of all Lebesgue-measurable function from $\mathbb{R}^m$ to $\mathbb{R}^n$. The only notable regularity condition is (3). In Appendix C.3, we discuss how to verify it and why it can safely be treated as a mild technical condition. Furthermore, it is worth mentioning that the Lipschitz conditions can easily be weakened [62]

4.2 Main results

We can now state our main result, which quantifies the deviation in the equilibrium distributions of the two diffusions in terms of the mixing rate and the difference between the diffusions' drifts.

Theorem 4.2.1 (Error induced by approximate drift)**.** *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eq.* (4.1) *and Eq.* (4.2)*, respectively. If the diffusion Eq.* (4.1) *is exponentially ergodic with parameters C and ρ, the regularity conditions of Assumption 4.B hold, and $\|b - \tilde{b}\|_2 \leq \epsilon$, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{\log(1/\rho)}. \tag{4.4}$$

Remark 4.2.2 (*Coherency of the error bound*). To check that the error bound of Eq. (4.4) has coherent dependence on its parameters, consider the following thought experiment. Suppose we change the time scale of the diffusion from t to $s = at$

for some $a > 0$. We are simply *speeding up* or *slowing down* the diffusion process depending on whether $a > 1$ or $a < 1$. Changing the time scale does not affect the equilibrium distribution and hence $d_{\mathcal{W}}(\pi, \tilde{\pi})$ remains unchanged. After time s has passed, the exponential contraction is ρ^{at} and hence the effective contraction constant is ρ^a instead of ρ. Moreover, the drift at each location is also scaled by a and hence the drift error is ϵa. The scaling a thus cancels out in the error bound, which is desirable since the error should be independent of how we set the time scale. □

Remark 4.2.3 (*Tightness of the error bound*). We can choose b and $\tilde{b}$ such that the bound in Eq. (4.4) is an equality, thus showing that, under the assumptions considered, Theorem 4.2.1 cannot be improved. Let $\pi(x) = \mathcal{N}(x; \mu, \sigma^2 I)$ be the Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\sigma^2 I$ and let $\tilde{\pi}(x) = \mathcal{N}(x; \tilde{\mu}, \sigma^2 I)$. The Wasserstein distance between two Gaussians with the same covariance is the distance between their means, so $d_{\mathcal{W}}(\pi, \tilde{\pi}) = \|\mu - \tilde{\mu}\|_2$. Consider the corresponding diffusions where $b = \nabla \log \pi$ and $\tilde{b} = \nabla \log \tilde{\pi}$. We have that for any $x \in \mathbb{R}^d$, $\|b(x) - \tilde{b}(x)\|_2 = \sigma^{-2}\|\mu - \tilde{\mu}\|_2 =: \epsilon$. Furthermore, the Hessian is $(\nabla^2 \log \pi)(x) = -\sigma^{-2} I$, which implies that b is σ^{-2}-strongly concave. Therefore, per the discussion in Section 4.1, exponential contractivity holds with $C = 1$ and $\rho = e^{-\sigma^{-2}}$. We thus conclude that

$$\frac{C\epsilon}{\log(1/\rho)} = \frac{\sigma^{-2}\|\mu - \tilde{\mu}\|_2}{\sigma^{-2}} = \|\mu - \tilde{\mu}\|_2 = d_{\mathcal{W}}(\pi, \tilde{\pi}).$$

and hence the bound of Theorem 4.2.1 is tight in this setting. □

Theorem 4.2.1 assumes that the approximate drift is a deterministic function and that the error in the drift is uniformly bounded. We can generalize the results of Theorem 4.2.1 to allow for the approximate diffusion to use stochastic drift with non-uniform drift error. We will see that only the expected magnitude of the drift bias affects the final error bound. Let $\tilde{b}(\tilde{X}_t, \tilde{Y}_t)$ denote the approximate drift, which is now a function of both the current location $\tilde{X}_t$ and an independent diffusion $\tilde{Y}_t \in \mathbb{R}^\ell$:

$$\begin{aligned} \mathrm{d}\tilde{X}_t &= (\tilde{b}(\tilde{X}_t, \tilde{Y}_t))\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\tilde{W}_t^X \\ \mathrm{d}\tilde{Y}_t &= b_{aux}(\tilde{Y}_t)\,\mathrm{d}t + \Sigma\,\mathrm{d}\tilde{W}_t^Y, \end{aligned} \tag{4.5}$$

where Σ is an $\ell \times \ell$ matrix and the notation $\tilde{W}_t^X$ and $\tilde{W}_t^Y$ highlights that the Brownian motions in $\tilde{X}_t$ and $\tilde{Y}_t$ are independent. Let $\tilde{\pi}_Z$ denote the stationary distribution of $\tilde{Z}_t := (\tilde{X}_t, \tilde{Y}_t)$. For measure μ and function f, we write $\mu(f) := \int f(x)\mu(\mathrm{d}x)$ to reduce clutter. We can now state a generalization of Theorem 4.2.1.

Theorem 4.2.4 (Error induced by stochastic approximate drift)**.** *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eqs.* (4.1) *and* (4.5)*, respectively. Assume that there exists a measurable function $\epsilon \in C(\mathbb{R}^d, \mathbb{R}_+)$ such that for $(\tilde{X}, \tilde{Y}) \sim \tilde{\pi}_Z$ and for all $x \in \mathbb{R}^d$,*

$$\|b(x) - \mathbb{E}[\tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X} = x]\|_2 \leq \epsilon(x).$$

If the diffusion Eq. (4.1) *is exponentially ergodic and the regularity conditions of*

Assumption 4.B hold, then

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\,\tilde{\pi}(\epsilon)}{\log(1/\rho)}.$$

Whereas the bound of Theorem 4.2.1 is proportional to the deterministic drift error ϵ, the bound for the diffusion with a stochastic approximate drift is proportional to the expected drift error bound $\tilde{\pi}(\epsilon)$. The bound of Theorem 4.2.4 thus takes into account how the drift error varies with the location of the drift. Our results match the asymptotic behavior for stochastic gradient Langevin dynamics documented in Teh et al. [135]: in the limit of the step size going to zero, they show that the stochastic gradient has no effect on the equilibrium distribution.

Example. Suppose $\tilde{Y}_t$ is an Ornstein–Uhlenbeck process with $\ell = d$, the dimensionality of $\tilde{X}_t$. That is, for some $\alpha, v > 0$, $\mathrm{d}\tilde{Y}_t = -\alpha\tilde{Y}_t\mathrm{d}t + \sqrt{2v}\,\mathrm{d}\tilde{W}_t^Y$. Then the equilibrium distribution of $\tilde{Y}_t$ is that of a Gaussian with covariance $\sigma^2 I$, where $\sigma^2 := v/\alpha$. Let $\tilde{b}(x, y) = b(x) + y$, so $\mathbb{E}[\tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X} = x] = b(x)$ and hence $d_{\mathcal{W}}(\pi, \tilde{\pi}) = 0$. □

While exponential contractivity is natural and applies in many settings, it is useful to have bounds on the Wasserstein distance of approximations when the diffusion process mixes more slowly. We can prove the analogous guarantee of Theorem 4.2.1 when a weaker, polynomial contractivity condition is satisfied.

Assumption 4.C (**Polynomial contractivity**). *There exist constants $C > 0$, $\alpha > 1$, and $\beta > 0$ such that for all $x, x' \in \mathbb{R}^d$,*

$$d_{\mathcal{W}}(\mu_{x,t}, \mu_{x',t}) \leq C\|x - x'\|_2(t + \beta)^{-\alpha}.$$

The parameters α and β determines how quickly the diffusion converges to equilibrium. Polynomial contractivity can be certified using, for example, the techniques from Butkovsky [28] (see also the references therein).

Theorem 4.2.5 (Error induced by approximate drift, polynomial contractivity). *Let π and $\tilde{\pi}$ denote the invariant distributions of the diffusions in Eq.* (4.1) *and Eq.* (4.2)*, respectively. If the diffusion Eq.* (4.1) *is polynomially ergodic with parameters C, α, and β, the regularity conditions of Assumption 4.B hold, and $\|b - \tilde{b}\|_2 \leq \epsilon$, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{(\alpha - 1)\beta^{\alpha-1}}. \tag{4.6}$$

Remark 4.2.6 (*Coherency of the error bound*). The error bound of Eq. (4.6) has a coherent dependence on its parameters, just like Eq. (4.4). If we change the time scale of the diffusion from t to $s = at$ for some $a > 0$, the polynomial contractivity constants C, α, and β become, respectively, $C/a^\alpha, \alpha$, and β/a. Making these substitutions and replacing ϵ by ϵa, one can check that the scaling a cancels out in the error bound, so the error is independent of how we set the time scale. □

4.3 Overview of analysis techniques

We use Stein's method [13, 119, 130] to bound the Wasserstein distance between π and $\tilde{\pi}$ as a function of a bound on $\|b - \tilde{b}\|_2$ and the mixing time of π. We describe the analysis ideas for the setting when $\|b - \tilde{b}\|_2 < \epsilon$ (Theorem 4.2.1); the analysis with stochastic drift (Theorem 4.2.4) or assuming polynomial contractivity (Theorem 4.2.5) is similar. All of the details are in Appendix C.2.

For a diffusion $(X_t)_{t\geq 0}$ with drift b, the corresponding infinitesimal generator satisfies

$$\mathcal{A}_b\phi(x) = b(x) \cdot \nabla\phi(x) + \Delta\phi(x)$$

for any function ϕ that is twice continuously differentiable and vanishing at infinity. See, e.g., Ethier and Kurtz [48] for an introduction to infinitesimal generators. Under quite general conditions, the invariant measure π and the generator $\mathcal{A}_b$ satisfy

$$\pi(\mathcal{A}_b\phi) = 0.$$

For any measure ν on $\mathbb{R}^d$ and set of test functions $\mathcal{F} \subseteq C^2(\mathbb{R}^d, \mathbb{R})$, we can define the *Stein discrepancy* as:

$$\mathcal{S}(\nu, \mathcal{A}_b, \mathcal{F}) := \sup_{\phi\in\mathcal{F}} |\pi(\mathcal{A}_b\phi) - \nu(\mathcal{A}_b\phi)| = \sup_{\phi\in\mathcal{F}} |\nu(\mathcal{A}_b\phi)|.$$

The Stein discrepancy quantifies the difference between ν and π in terms of the maximum difference in the expected value of a function (belonging to the transformed test class $\{\mathcal{A}_b\phi \,|\, \phi \in \mathcal{F}\}$) under these two distributions. We can analyze the Stein discrepancy between π and $\tilde{\pi}$ as follows. Consider a test set $\mathcal{F}$ such that $\|\nabla\phi\|_2 \leq 1$ for all $\phi \in \mathcal{F}$, which is equivalent to having $\|\phi\|_L \leq 1$. We have that

$$\begin{aligned}
\mathcal{S}(\tilde{\pi}, \mathcal{A}_b, \mathcal{F}) &= \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b\phi)| = \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b\phi - \mathcal{A}_{\tilde{b}}\phi)| \\
&= \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\nabla\phi \cdot b - \nabla\phi \cdot \tilde{b})| \\
&\leq \sup_{\phi\in\mathcal{F}} |\tilde{\pi}(\|\nabla\phi\|_2 \|b - \tilde{b}\|_2)| \leq \epsilon,
\end{aligned}$$

where we have used the definition of Stein discrepancy, that $\tilde{\pi}(\mathcal{A}_{\tilde{b}}\phi) = 0$, the definition of the generator, the Cauchy-Schwartz inequality, that $\|\nabla\phi\|_2 \leq 1$, and the assumption $\|b - \tilde{b}\|_2 \leq \epsilon$. It remains to show that the Wasserstein distance satisfies $d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq C_\pi \mathcal{S}(\tilde{\pi}, \mathcal{A}_b, \mathcal{F})$ for some constant C_π that may depend on π. This would then allow us to conclude that $d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq C_\pi\epsilon$. To obtain C_π, for each 1-Lipschitz function h, we construct the solution u_h to the differential equation

$$h - \pi(h) = \mathcal{A}_g u \tag{4.7}$$

and show that $\|\nabla u_h\|_2 \leq C_\pi \|\nabla h\|_2$.

4.4 Application: computational–statistical trade-offs

As an application of our results we analyze the behavior of the *unadjusted Langevin Monte Carlo algorithm* (ULA) [118] when approximate gradients of the log-likelihood are used. ULA uses a discretization of the continuous-time Langevin diffusion to approximately sample from the invariant distribution of the diffusion. We prove conditions under which we can obtain more accurate samples by using an approximate drift derived from a Taylor expansion of the exact drift.

For the diffusion $(X_t)_{t\geq 0}$ driven by drift b as defined in Eq. (4.1) and a non-increasing sequence of step sizes $(\gamma_i)_{i\geq 1}$, the associated ULA Markov chain is

$$X'_{i+1} = X'_i + \gamma_{i+1}\, b(X'_i) + \sqrt{2\gamma_{i+1}}\,\xi_{i+1}, \qquad \xi_{i+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{4.8}$$

Recently, substantial progress has been made in understanding the approximation accuracy of ULA [27, 39, 43]. These analyses show, as a function of the discretization step size γ_i, how quickly the distribution of X'_i converges to the desired target distribution.

In many big data settings, however, computing $b(X'_i)$ exactly at every step is computationally expensive. Given a fixed computational budget, one option is to compute $b(X'_i)$ precisely and run the discretized diffusion for a small number of steps to generate samples. Alternatively, we could replace $b(X'_i)$ with an approximate drift $\tilde{b}(X'_i)$ which is cheaper to compute and run the discretized approximate diffusion for a larger number of steps to generate samples. While approximating the drift can introduce error, running for more steps can compensate by sampling from a better mixed chain. Thus, our objective is to compare the ULA chain using an exact drift initialized at some point $x^\star \in \mathbb{R}^d$ to a ULA chain using an approximate drift initialized at the same point. We denote the exact and approximate drift chains by $X'_{x^\star,i}$ and $\tilde{X}'_{x^\star,i}$, respectively, and denote laws of these chains by $\mu_i^\star$ and $\tilde{\mu}_i^\star$.

For concreteness, we analyze generalized linear models with unnormalized log-densities of the form

$$\mathcal{L}(x) := \log \pi_0(x) + \sum_{i=1}^{N} \phi_i(x \cdot y_i),$$

where $y_1, \ldots, y_N \in \mathbb{R}^d$ is the data and x is the parameter. In this setting the drift is $b(x) = \nabla \mathcal{L}(x)$. We take $x^\star = \arg\max_x \mathcal{L}(x)$ and approximate the drift with a Taylor expansion around $x^\star$:

$$\tilde{b}(x) := (\nabla^2 \log \pi_0)(x^\star)(x - x^\star) + \sum_{i=1}^{N} \phi_i''(x^\star \cdot y_i) y_i y_i^\top (x - x^\star), \tag{4.9}$$

where ∇^2 is the Hessian operator. The quadratic approximation of Eq. (4.9) basically corresponds to taking a Laplace approximation of the log-likelihood. In practice, higher-order Taylor truncation or other approximations can be used, and our analysis can be extended to quantify the trade-offs in those cases as well. Here we focus on the

second-order approximation as a simple illustration of the computational–statistical trade-off.

In order for the Taylor approximation to be well-behaved, we require the prior π_0 and link functions ϕ_i to satisfying some regularity conditions, which are usually easy to check in practice.

Assumption 4.D (Concavity, smoothness, and asymptotic behavior of data)**.**

1. *The function* $\log \pi_0 \in C^3(\mathbb{R}^d, \mathbb{R})$ *is strongly concave,* $\|\nabla \log \pi_0\|_L < \infty$*, and* $\|\nabla^2[\partial_j \log \pi_0]\|_2 < \infty$ *for* $j = 1, \dots, d$*, where* $\|\cdot\|_2$ *denotes the matrix spectral norm.*

2. *For* $i = 1, \dots, N$*, the function* $\phi_i \in C^3(\mathbb{R}, \mathbb{R})$ *is strongly concave,* $\|\phi_i'\|_L < \infty$*, and* $\|\phi_i'''\|_\infty < \infty$*.*

3. *The data satisfies* $\|\sum_{i=1}^N y_i y_i^\top\|_2 = \Theta(N)$*.*

We measure computational cost by the number of d-dimensional inner products performed. Running ULA with the original drift b for T steps costs TN because each step needs to compute $x \cdot y_i$ for each of the N y_i's. Running ULA with the Taylor approximation $\tilde{b}$, we need to compute $\sum_{i=1}^N \phi_i''(x^\star \cdot y_i) y_i y_i^\top$ once up front, which costs Nd, and then for each step we just multiply this d-by-d matrix with $x - x^\star$, which costs d. So the total cost of running approximate ULA for $\tilde{T}$ steps is $(\tilde{T} + N)d$.

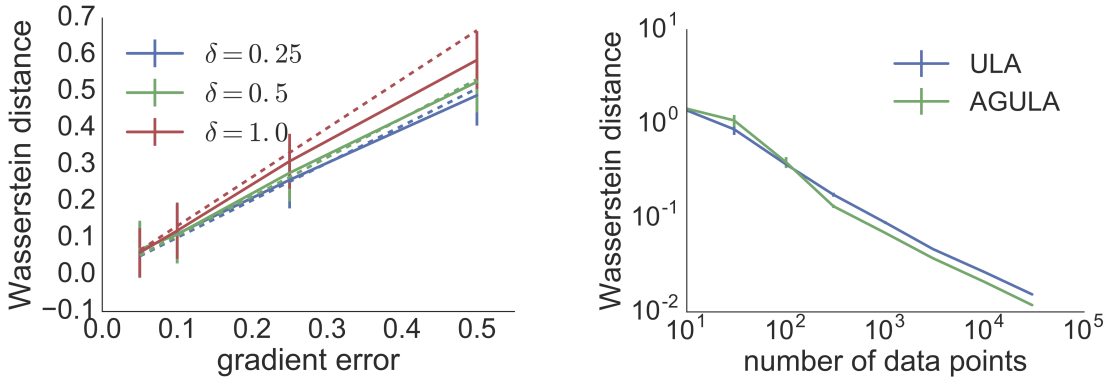
Theorem 4.4.1 (Computational–statistical trade-off for ULA)**.** *Set the step size* $\gamma_i = \gamma_1 i^{-\alpha}$ *for fixed* $\alpha \in (0, 1)$ *and suppose the ULA of Eq.* (4.8) *is run for* $T > d$ *steps. If Assumption 4.D holds and* $\tilde{T}$ *is chosen such that the computational cost of the second-order approximate ULA using drift Eq.* (4.9) *equals that of the exact ULA, then* γ_1 *may be chosen such that*

$$d_{\mathcal{W}}^2(\mu_T^\star, \pi) = \tilde{O}\left(\frac{d}{TN}\right) \qquad \text{and} \qquad d_{\mathcal{W}}^2(\tilde{\mu}_{\tilde{T}}^\star, \pi) = \tilde{O}\left(\frac{d^2}{N^2 T} + \frac{d^3}{N^2}\right).$$

The ULA procedure of Eq. (4.8) has Wasserstein error decreasing like $1/N$ for data size N. Because approximate ULA can be run for more steps at the same computational cost, its error decreases as $1/N^2$. Thus, for large N and fixed T and d, approximate ULA with drift $\tilde{b}$ achieves more accurate sampling than ULA with b. A conceptual benefit of our results is that we can cleanly decompose the final error into the discretization error and the equilibrium bias due to approximate drift. Our theorems in Section 4.2 quantifies the equilibrium bias, and we can apply existing techniques to bound the discretization error.

4.5 Extension: piecewise deterministic Markov processes

We next demonstrate the generality of our techniques by providing a perturbation analysis of piecewise deterministic Markov processes (PDMPs), which are continuous-time processes that are deterministic except at random jump times. Originating with



(a) (b)

Figure 4-1: **(a)** Gradient error ϵ versus the Wasserstein distance between π_δ and $\tilde{\pi}_{\delta,\epsilon}$, the stationary distribution of the diffusion with approximate drift $\tilde{b}_{\delta,\epsilon}(x) = \nabla \log \pi_\delta(x) + \epsilon$. The solid lines are the simulation results and the dotted lines are the theoretical upper bounds obtained from Theorem 4.2.1. The simulation results closely match the theoretical bounds and show linear growth in ϵ, as predicted by the theory. Due to Monte Carlo error the simulation estimates sometimes slightly exceed the theoretical bounds. **(b)** The y-axis measures the Wasserstein distance between the true posterior distribution and the finite-time distribution of the exact gradient ULA (ULA) and the approximate gradient ULA (AGULA). Except for when the number of data points $N < 100$, AGULA shows superior performance, in agreement with the analysis of Theorem 4.4.1. For all experiments the Wasserstein distance was estimated 10 times, each time using 1,000 samples from each distribution.

the work of Davis [41], there is now a rich literature on the ergodic and convergence properties of PDMPs [8, 16, 36, 53, 99]. They have been used to model a range of phenomena including communication networks, neuronal activity, and biologic population models (see [8] and references therein). Recently, PDMPs have also been used to design novel MCMC inference schemes. zig-zag processes (ZZPs) [18–20] are a class of PDMPs that are particularly promising for inference. ZZPs can be simulated exactly (making Metropolis-Hastings corrections unnecessary) and are non-reversible, which can potentially lead to more efficient sampling [97, 102].

Our techniques can be readily applied to analyze the accuracy of approximate PDMPs. For concreteness we demonstrate the results for ZZPs in detail and defer the general treatment of PDMPs, which includes an idealized version of Hamiltonian Monte Carlo, to Appendix C.4. The ZZP is defined on the space $E = \mathbb{R}^d \times \mathcal{B}$, where $\mathcal{B} := \{-1, +1\}^d$. Densities on $\mathcal{B}$ are with respect to the counting measure.

Informally, the behavior of a ZZP can be described as follows. The trajectory is X_t and its velocity is Θ_t, so $\frac{d}{dt} X_t = \Theta_t$. At random times, a single coordinate of Θ_t flips signs. In between these flips, the velocity is a constant and the trajectory is a straight line (hence the name "zig-zag"). The rate at which Θ_t flips a coordinate is time inhomogeneous. The i-th component of Θ switches at rate $\lambda_i(X_t, \Theta_t)$. By

choosing the switching rates appropriately, the ZZP can be made to sample from the desired distribution. More precisely, the ZZP $(X_t, \Theta_t)_{t\geq 0}$ is determined by the switching rate $\lambda \in C^0(E, \mathbb{R}^d_+)$ and has generator

$$\mathcal{A}_\lambda \phi(x,\theta) = \theta \cdot \nabla_x \phi(x,\theta) + \lambda(x,\theta) \cdot \nabla_\theta \phi(x,\theta) \tag{4.10}$$

for any sufficiently regular $\phi : E \to \mathbb{R}$. Here $\nabla_x \phi$ denotes the gradient of ϕ with respect to x and $\nabla_\theta \phi$ is the discrete differential operator.[4] Let $(a)^+ := \max(0, a)$ denote the positive part of $a \in \mathbb{R}$ and $\partial_i \phi := \frac{\partial \phi}{\partial x_i}$. The following result shows how to construct a ZZP with invariant distribution π.

Theorem 4.5.1 (Bierkens et al. [20, Theorem 2.2, Proposition 2.3])**.** *Suppose* $\log \pi \in C^1(\mathbb{R}^d)$ *and* $\gamma \in C^0(E, \mathbb{R}^d_+)$ *satisfies* $\gamma_i(x,\theta) = \gamma_i(x, R_i\theta)$*. Let*

$$\lambda_i(x,\theta) = (-\theta_i \partial_i \log \pi(x))^+ + \gamma_i(x,\theta).$$

Then the Markov process with generator $\mathcal{A}_\lambda$ *has invariant distribution* $\pi_E(\mathrm{d}x, \theta) = 2^{-d}\pi(\mathrm{d}x)$*.*

Analogously to the approximate diffusion setting, we compare $(X_t, \Theta_t)_{t\geq 0}$ to an approximating ZZP $(\tilde{X}_t, \tilde{\Theta}_t)_{t\geq 0}$ with switching rate $\tilde{\lambda} \in C^0(E, \mathbb{R}^d_+)$. For example, if $\tilde{\pi}$ is an approximating density, the approximate switching rate could be chosen as

$$\tilde{\lambda}_i(x,\theta) = (-\theta_i \partial_i \log \tilde{\pi}(x))^+ + \gamma_i(x,\theta). \tag{4.11}$$

To relate the errors in the switching rates to the Wasserstein distance in the final distributions, we use the same strategy as before: apply Stein's method to the ZZP generator in Eq. (4.10). We rely on ergodicity and regularity conditions that are analogous to those for diffusions. We write $(X_{x,\theta,t}, \Theta_{x,\theta,t})$ to denote the version of the ZZP satisfying $(X_{x,\theta,0}, \Theta_{x,\theta,0}) = (x,\theta)$ and denote its law by $\mu_{x,\theta,t}$.

Assumption 4.E (**ZZP polynomial ergodicity**)**.** *There exist constants* $C > 0$*,* $\alpha > 1$*, and* $\beta > 0$ *such that for all* $x \in \mathbb{R}^d$*,* $\theta \in \mathcal{B}$*, and* $i \in [d]$*,*

$$d_{\mathcal{W}}(\mu_{x,\theta,t}, \mu_{x,R_i\theta,t}) \leq C(t+\beta)^{-\alpha}.$$

The ZZP polynomial ergodicity condition is looser than that used for diffusions. Indeed, we only need a quantitative bound on the ergodicity constant when the chains are started with the same x value. Together with the fact that $\mathcal{B}$ is compact, this simplifies verification of the condition, which can be done using well-developed coupling techniques from the PDMP literature [8, 16, 53, 99] as well as more general Lyapunov function-based approaches [67].

Our main result of this section bounds the error in the invariant distributions due to errors in the ZZP switching rates. It is more natural to measure the error between λ and $\tilde{\lambda}$ in terms of the ℓ^1 norm.

[4] $\nabla_\theta \phi := (\partial_{\theta,1}\phi, \ldots, \partial_{\theta,d}\phi)$, where $\partial_{\theta,i}\phi(x,\theta) := \phi(x, R_i\theta) - \phi(x,\theta)$ and for $i \in [d]$, the reversal function $R_i : \mathcal{B} \to \mathcal{B}$ is given by $(R_i\theta)_j := \begin{cases} -\theta_j & j = i \\ \theta_j & j \neq i. \end{cases}$

Theorem 4.5.2 (ZZP error induced by approximate switching rate)**.** *Assume the ZZP with switching rate λ (respectively $\tilde{\lambda}$) has invariant distribution π (resp. $\tilde{\pi}$). Also assume that $\int_E x^2 \pi(\mathrm{d}x, \mathrm{d}\theta) < \infty$ and if a function $\phi \in C(E, \mathbb{R})$ is π-integrable then it is $\tilde{\pi}$-integrable. If the ZZP with switching rate λ is polynomially ergodic with constants C, α, and β and $\|\lambda - \tilde{\lambda}\|_1 \leq \epsilon$, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{(\alpha - 1)\beta^{\alpha - 1}}.$$

Remark 4.5.3. If the approximate switching rate takes the form of Eq. (4.11), then $\|\nabla \log \pi - \nabla \log \tilde{\pi}\|_1 \leq \epsilon$ implies $\|\lambda - \tilde{\lambda}\|_1 \leq \epsilon$. □

4.6 Experiments

We used numerical experiments to investigate whether our bounds capture the true behavior of approximate diffusions and their discretizations.

Approximate Diffusions. For our theoretical results to be a useful guide in practice, we would like the Wasserstein bounds to be reasonably tight and have the correct scaling in the problem parameters (e.g., in $\|b - \tilde{b}\|_2$). To test our main result concerning the error induced from using an approximate drift (Theorem 4.2.1), we consider mixtures of two Gaussian densities of the form

$$\pi_\delta(x) = \frac{1}{2(2\pi)^{d/2}} \Big(e^{-\|x - \delta/2\|_2^2/2} + e^{-\|x + \delta/2\|_2^2/2} \Big),$$

where $\delta \in \mathbb{R}^d$ parameterizes the difference between the means of the Gaussians. If $\|\delta\|_2 < 2$, then π_δ is $(1 - \|\delta\|_2/4)$-strongly log-concave; if $\|\delta\|_2 = 2$, then π_δ is log-concave; and if $\|\delta\|_2 > 2$, then π_δ is not log-concave, but is log-concave in the tails. Thus, for all choices of δ, the diffusion with drift $b_\delta(x) := \nabla \log \pi_\delta(x)$ is exponentially ergodic. Importantly, this class of Gaussian mixtures allows us to investigate a range of practical regimes, from strongly unimodal to highly multi-modal distributions. For $d = 1$ and a variety of choices of δ, we generated 1,000 samples from the target distribution π_δ (which is the stationary distribution of a diffusion with drift $b_\delta(x)$) and from $\tilde{\pi}_{\delta,\epsilon}$ (which is the stationary distribution of the approximate diffusion with drift $\tilde{b}_{\delta,\epsilon}(x) := b_\delta(x) + \epsilon$) for $\epsilon = 0.05, 0.1, 0.25, 0.5$. We then calculated the Wasserstein distance between the empirical distribution of the target and the empirical distribution of each approximation. Fig. 4-1a shows the empirical Wasserstein distance (solid lines) for $\delta = 0.25, 0.5, 1.0$ along with the corresponding theoretical bounds from Theorem 4.2.1 (dotted lines). The two are in close agreement. We also investigated larger distances for $\delta = 1.0, 2.0, 3.0$. Here the exponential contractivity constants that can be derived from Eberle [46] are rather loose. Importantly, however, for all values of δ considered, the Wasserstein distance grows linearly in ϵ, as predicted by our theory. Results for $d > 1$ show similar linear behavior in ϵ, though we omit the plots.

Computational–statistical trade-off. We illustrate the computational–statistical trade-off of Theorem 4.4.1 in the case of logistic regression. This corresponds to

$\phi_i(t) = \phi_{lr}(t) := -\log(1 + e^{-t})$. We generate data $y_1, y_2, \ldots$ according to the following process:

$$z_i \sim \mathsf{Bern}(.5), \qquad \zeta_i \sim \mathcal{N}(\mu_{z_i}, I), \qquad y_i = (2z_i - 1)\zeta_i,$$

where $\mu_0 = (0, 0, 1, 1)$ and $\mu_1 = (1, 1, 0, 0)$. We restrict the domain $\mathcal{X}$ to a ball of radius 3, $\mathcal{X} = \{x \in \mathbb{R}^4 \,|\, \|x\|_2 \leq 3\}$, and add a projection step to the ULA algorithm [27], replacing Z_i' with $\arg\min_{z \in \mathcal{X}} \|Z_i' - z\|_2$. While Theorem 4.4.1 assumes $\mathcal{X} = \mathbb{R}^4$, the numerical results here on the bounded domain still illustrate our key point: for the same computational budget, computing fast approximate gradients and running the ULA chain for longer can produce a better sampler. Fig. 4-1b shows that except for very small N, the approximate gradient ULA (AGULA), which uses the approximation in Eq. (4.9), produces better performance than exact gradient ULA (ULA) with the same budget. For each data-set size (N), the true posterior distribution was estimated by running an adaptive Metropolis-Hastings (MH) sampler for 100,000 iterations. ULA and AGULA were each run 1,000 times to empirically estimate the approximate posteriors. We then calculated the Wasserstein distance between the ULA and AGULA empirical distributions and the empirical distribution obtained from the MH sampler.

4.7 Discussion

Related Work. Recent theoretical work on scalable MCMC algorithms has yielded numerous insights into the regimes in which such methods produce computational gains [5, 76, 108, 120]. Many of these works focused on approximate Metropolis-Hastings algorithms, rather than gradient-based MCMC. Moreover, the results in these papers are for discrete chains, whereas our results also apply to continuous diffusions as well as other continuous-time Markov processes such as the zig-zag process. Perhaps the closest to our work is that of Rudolf and Schweizer [120] and Gorham et al. [62]. The former studies general perturbations of Markov chains and includes an application to stochastic Langevin dynamics. They also rely on a Wasserstein contraction condition, like our Assumption 4.A, in conjunction with a Lyapunov condition on the perturbed chain. However, our more specialized analysis is particularly transparent and leads to tighter bounds in terms of the contraction constant ρ: the bound of Rudolf and Schweizer [120] is proportional to $(1 - \rho)^{-1}$ whereas our bound is proportional to $-(\log \rho)^{-1}$. Another advantage of our approach is that our results are more straightforward to apply since we do not need to directly analyze the Lyapunov potential and the perturbation ratios as in Rudolf and Schweizer [120]. Our techniques also apply to the weaker polynomial contraction setting. Gorham et al. [62] have results of similar flavor to ours and also rely on Stein's method, but their assumptions and target use cases differ from ours. Our results in Section 4.4, which apply when ULA is used with a deterministic approximation to the drift, complement the work of Teh et al. [135] and Vollmer et al. [138], which provides (non-)asymptotic analysis when the drift is approximated stochastically at each iteration.

Conclusion. We have established general results on the accuracy of diffusions with approximate drifts. As an application, we show how this framework can quantify the computational–statistical trade-off in approximate gradient ULA. The example in Section 5.5 illustrates how the log-concavity constant can be estimated in practice and how theory provides reasonably precise error bounds. We expect our general framework to have many further applications. In particular, an interesting direction is to extend our framework to analyze the trade-offs in subsampling Hamiltonian Monte Carlo algorithms and stochastic Langevin dynamics.

Chapter 5

Fast Generalized Maximum Mean Discrepancies

5.1 Introduction

Maximum mean discrepancies [MMDs, 63] like the kernel Stein discrepancy [KSD, 35, 61, 86] provide a principled and convenient way to compute the distance between probability distributions. They have been applied to a range of problems in machine learning and statistics, including measuring the quality of samples from approximate Bayesian inference algorithms [61], goodness-of-fit testing [35, 75, 86], two-sample testing [34, 63], and approximate Bayesian inference [85].

Unfortunately, the computation of an MMD requires evaluating a kernel function at every pair of sample points, making its use prohibitive for large sample sizes. Thus, substantial efforts have been devoted to approximating MMDs, with the objective of obtaining $o(N^2)$ running times given N sample points.[1] Popular approaches approximate the MMD in $\Theta(NM)$ time using a finite feature expansion of size M based on random Fourier features [72, 113, 128, 132, 149], other types of randomly sampled features [9, 25, 34, 38, 75], or the Nyström method [101, 144, 146]. However, existing analyses only guarantee $O_P(M^{-1/2})$-precision estimates of the reference MMD and require $\Theta(N^2)$ time to achieve the typical $\Theta(N^{-1/2})$ precision of the reference MMD.[2] The aim of this work is to provide practical prescriptions for improving this computation–accuracy trade-off.

To this end, in Section 5.3, we first introduce a new class of kernel-based discrepancy measures, *generalized MMDs* (GMMDs), and establish upper and lower bounds in terms of standard reference MMDs. We then develop cheap stochastic approximations, *fast GMMDs* (fGMMDs), that closely approximate each GMMD. In Section 5.4, we derive high probability relative error bounds for fGMMDs and, for any $\gamma > 0$, show how to compute $O_P(N^{-1/2})$-precision estimates of numerous GMMDs and many common MMDs in $O(N^{1+\gamma})$ (near-linear) time when the GMMD or MMD precision is $\Omega(N^{-1/2})$.

[1] In the two-sample setting, we assume each sample has size N.

[2] See Section 5.4.2 for a more detailed discussion of MMD precision and its import.

In the context of hypothesis testing, we further derive the asymptotic distribution of the fGMMD when sample points are drawn i.i.d. and develop an asymptotically exact and full-power test of goodness of fit. We validate the benefits of fGMMDs in Section 5.5 by using them to select hyperparameters of biased Markov chain Monte Carlo (MCMC) samplers and to conduct fast goodness-of-fit tests. We obtain high-quality results using only 10-25 features.

Notation For a (signed) measure μ, we abuse notation and write

$$\mu(f) := \int f(x)\mu(\mathrm{d}x).$$

If μ_i is a measure on $\mathcal{X}_i$ and $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$,

$$(\mu_1 \times \mu_2)(f) := \int\int f(x_1, x_2)\mu_1(\mathrm{d}x_1)\mu_2(\mathrm{d}x_2).$$

If μ is a measure on $\mathcal{X}$ and $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then μ applied to f is assumed to operate on the first argument of f: $(\mu f)(x') := \int f(x, x')\mu(\mathrm{d}x)$. For an operator, if there is ambiguity we use a subscript to indicate the variable it operates on. For example, $\nabla_x f(x, y)$ denotes the gradient of $f(x, y)$ with respect to its first argument while $\partial_{x_d} f(x, y)$ denotes the partial derivative of $f(x, y)$ with respect to x_d. We denote the generalized Fourier transform of a function f by $\hat{f}$ or $\mathcal{F}(f)$ and the inverse Fourier transform by $\mathcal{F}^{-1}(f)$. The convolution of functions f and g is denoted $(f * g)(x) = \int f(y)g(x - y)\,\mathrm{d}y$. For $r \geq 1$ and a measure (or density) μ, we write $L^r(\mu)$ to denote the space of functions with μ-integrable r-th moment. The corresponding norm is $\|f\|_{L^r(\mu)} := \left(\int |f(x)|^r \mu(\mathrm{d}x)\right)^{1/r}$. We write L^r when μ is the Lebesgue measure. We let $\overset{\mathcal{D}}{\Longrightarrow}$ and $\overset{P}{\to}$ denote convergence in distribution and in probability, respectively. For a complex number a, $\overline{a}$ denotes its complex conjugate. For $J \in \mathbb{N}$, define $[J] := \{1, \ldots, J\}$.

5.2 Maximum mean discrepancies

In sample quality measurement and goodness-of-fit testing, our aim is to quantify how well a sample $Q_N = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}$ with sample points $x_1, \ldots, x_N \in \mathcal{X} \subseteq \mathbb{R}^D$ approximates a fixed target distribution P on $\mathcal{X}$. It is common to frame this comparison in terms of an integral probability metric [100] measuring the maximum discrepancy between sample and target expectations over a class of test functions. When the class of test functions is the unit ball of a reproducing kernel Hilbert space (RKHS), one recovers the maximum mean discrepancy (MMD) [63],

$$\begin{aligned}\mathrm{MMD}^2_{k_0}(Q_N, P) &= (\Delta_N \times \Delta_N)k_0 \\ &= \frac{1}{N^2}\sum_{n=1}^{N}\sum_{n'=1}^{N} k_0(x_n, x_{n'}) - \frac{2}{N}\sum_{n=1}^{N}(Pk_0)(x_n) + (P \times P)k_0,\end{aligned}$$

where $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel* (i.e., a symmetric, positive definite function) and Δ_N is the signed measure $Q_N - P$.

An important special case of an MMD is the kernel Stein discrepancy (KSD) [35, 61, 86], which avoids explicit integration under P by employing a *Stein kernel*

$$k_0(x, y) = \sum_{d=1}^{D} \frac{1}{p(x)p(y)} \partial_{x_d} \partial_{y_d} (p(x)k(x, y)p(y)) \tag{5.1}$$

designed to have zero mean under P. Here, p is the (Lebesgue) density of P, and k is any continuously differentiable base reproducing kernel.

A principal drawback of the generic MMD is its quadratic computational cost in the number of sample points. In the sections to follow, we will show how to construct alternative kernel-based discrepancy measures that retain the theoretical and practical benefits of a reference MMD at a reduced cost.

5.3 Generalized MMDs

Beginning with a reference MMD, we derive a family of *generalized MMDs* (GMMDs) which upper bound the MMD and, in certain special cases, are exactly equal to the reference MMD. We then show how to efficiently approximate GMMDs using importance sampling. We call the resulting stochastic discrepancy measures *fast GMMDs*.

We begin by restricting our focus to kernels of the form

$$k_0(x, y) = \sum_{j=1}^{J} \int k_j^{1/2}(x, \omega) \overline{k_j^{1/2}(y, \omega)} \rho(\omega) \, \mathrm{d}\omega, \tag{5.2}$$

where $J \in \mathbb{N}$, ρ nonnegative, and $k_j^{1/2}(x, \cdot) \in L^2(\rho)$ for all $x \in \mathcal{X}$. This class is broad enough to cover commonly employed kernels such as translation invariant kernels

$$k_0(x, y) = \Phi(x - y) = \int e^{i\langle \omega, x \rangle} \overline{e^{i\langle \omega, y \rangle}} \hat{\Phi}(\omega) \, \mathrm{d}\omega,$$

polynomial kernels

$$k_0(x, y) = (\langle x, y \rangle + c)^b = \sum_{j=1}^{J} A_j(x) A_j(y),$$

and Stein kernels of the form

$$\begin{aligned} &\sum_{d=1}^{D} \frac{1}{p(x)p(y)} \partial_{x_d} \partial_{y_d} (p(x)A(x)\Phi(x - y)A(y)p(y)) \\ &= \sum_{d=1}^{D} \int \frac{\partial_{x_d}(p(x)A(x)e^{i\langle \omega, x \rangle})}{p(x)} \overline{\frac{\partial_{y_d}(p(y)A(y)e^{i\langle \omega, y \rangle})}{p(y)}} \hat{\Phi}(\omega) \, \mathrm{d}\omega. \end{aligned}$$

The kernel decomposition Eq. (5.2) allows us to upper bound MMD_{k_0} using the generalized Hölder's inequality and the Babenko-Beckner inequality. For any $r \in$

$[1, 2]$, $s = r/(r-1)$, and $t = r/(2-r)$, we have

$$\begin{aligned}\mathrm{MMD}^2_{k_0}(Q_N, P) &= \textstyle\sum_{j=1}^J \int |(\Delta_N k_j^{1/2})(\omega)|^2 \rho(\omega) d\omega \\ &\leq \|\rho\|_{L^t} \textstyle\sum_{j=1}^J \|\Delta_N k_j^{1/2}\|^2_{L^s} \\ &\leq c_{r,d}^2 \|\rho\|_{L^t} \mathrm{GMMD}^2_{\boldsymbol{k}^{1/2},r}(Q_N, P).\end{aligned} \tag{5.3}$$

Here, $c_{r,d} := (r^{1/r}/s^{1/s})^{d/2} \leq 1$ and

$$\mathrm{GMMD}^2_{\boldsymbol{k}^{1/2},r}(Q_N, P) := \textstyle\sum_{j=1}^J \|\mathcal{F}^{-1}\Delta_N k_j^{1/2}\|^2_{L^r}$$

is what we term a *generalized MMD* (GMMD). Indeed, when $r = 2$, the GMMD is itself an instance of an MMD, and when $\rho \equiv 1$, $\mathrm{GMMD}_{\boldsymbol{k}^{1/2},2} = \mathrm{MMD}_{k_0}$.

The only Q_N-dependent term in the bound in Eq. (5.3) is the GMMD term, which we can approximate using importance sampling under a sampling density ν; we call the resulting stochastic discrepancy measure a *fast GMMD* (fGMMD):

$$\begin{aligned}&\mathrm{fGMMD}^2_{\boldsymbol{k}^{1/2},r,\nu,M}(Q_N, P) \\ &:= \sum_{j=1}^J \left(\frac{1}{M} \sum_{m=1}^M \frac{|(\mathcal{F}^{-1}\Delta_N k_j^{1/2})(Z_m)|^r}{\nu(Z_m)} \right)^{2/r}\end{aligned}$$

for $Z_1, \ldots, Z_M \overset{\text{i.i.d.}}{\sim} \nu$. Crucially, fGMMDs can be computed in $O(MN)$ time. We will show in Section 5.4 that, with appropriate practical choices of $\boldsymbol{k}^{1/2}, r$, and ν, the fGMMD dominates its reference MMD and GMMD with high probability even when M grows sublinearly in N. We will occasionally omit the Q_N and P arguments to MMD, GMMD, and fGMMD when they are clear from context.

5.3.1 Special cases

A number of existing MMD approximations and stochastic kernel discrepancies can be recovered as special cases of fGMMDs. If $J = 1$, $k_0(x, y) = \Phi(x - y)$,

$$\mathcal{F}^{-1}(k_1^{1/2}(x, \cdot))(z) = e^{-i\langle z, x\rangle} \hat{\Phi}(z)^{1/2},$$

and $\nu \propto \hat{\Phi}$, then $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},2,\nu,M}$ is the random Fourier feature (RFF) approximation to MMD_{k_0} [113]. Chwialkowski et al. [34, Prop. 1] showed that the RFF approximation can be a poor choice of discrepancy measure, as there exist uncountably many pairs of distinct distributions that, with high probability, cannot be distinguished by the RFF approximation.

Chwialkowski et al. [34] introduced two alternative stochastic kernel discrepancies, based on an arbitrary continuous sampling density ν, to overcome this limitation of RFFs. Their smooth characteristic function metric can be realized as $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},2,\nu,M}$ with $J = 1$ and $\mathcal{F}^{-1}(k_1^{1/2}(x, \cdot))(z) = e^{-i\langle z, x\rangle} \hat{\kappa}(x) \nu(z)^{1/2}$ for κ integrable, analytic, and positive definite. Their mean embedding metric is recovered by $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},2,\nu,M}$ when

$J = 1$ and $\mathcal{F}^{-1}(k_1^{1/2}(x,\cdot))(z) = f(x,z)\nu(z)^{1/2}$ for f a real analytic and characteristic reproducing kernel. The related random finite set Stein discrepancy [FSSD-rand, 75] is an $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},2,\nu,M}$ with $J = D$ and, for each $d \in [D]$, $\mathcal{F}^{-1}(k_d^{1/2}(x,\cdot))(z) = \frac{\partial_{x_d}(p(x)f(x,z))}{p(x)}\nu(z)^{1/2}$ for f a real analytic and C_0-universal [30, Def. 4.1] reproducing kernel. In each case, the fGMMD construction exposes a relationship to an underlying MMD_{k_0} with $\rho \equiv 1$. In the sequel, we will leverage this relationship to establish high-probability convergence-determining properties – or the lack thereof – for different classes of fGMMD.

5.4 Theoretical guarantees

In practice, we would like to select an fGMMD that (i) detects when a sample sequence is not converging to P, (ii) detects when a sample sequence is converging to P, and (iii) maintains subquadratic sample complexity. Our strategy is to first select a convergence-determining reference MMD and then, for any $\gamma > 0$, choose an associated fGMMD that satisfies the relative error bound

$$\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},r,\nu,M} \geq \tfrac{1}{2}\,\mathrm{GMMD}_{\boldsymbol{k}^{1/2},r} \geq C'\,\mathrm{MMD}_{k_0}$$

with high probability whenever $M = \Omega(N^\gamma)$. This ensures that (with probability 1 by Borel-Cantelli) if $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},r,\nu,M}(Q_N, P) \xrightarrow{P} 0$ then $\mathrm{MMD}_{k_0}(Q_N, P) \to 0$ and hence $Q_N \xRightarrow{\mathcal{D}} P$. We detail sufficient conditions for these properties to hold in Sections 5.4.1 and 5.4.2 and explicit examples in Section 5.4.3. In Section 5.4.4 we provide complementary finite-sample upper bounds on $\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},r,\nu,M}$ and $\mathrm{GMMD}_{\boldsymbol{k}^{1/2},r}$ in terms of MMD_{k_0}. Finally, in Section 5.4.5 we adopt a hypothesis testing perspective and derive the asymptotic distribution of $N\,\mathrm{fGMMD}^2$ when sample points x_n are i.i.d. draws from a distribution Q. This enables us to conduct asymptotically exact hypothesis tests based on fGMMD in Section 5.5.

5.4.1 Selecting a reference MMD

While many MMDs based on standard C_0 kernel functions exactly metrize weak convergence and hence detect non-convergence [see, e.g., 127, Prop. 71], few KSDs are known to determine weak convergence on $\mathbb{R}^d$. A notable exception is the KSD with inverse multiquadric (IMQ) base kernel $k(x,y) = \Psi_{c,\beta}^{\mathrm{IMQ}}(x-y) := (c^2 + \|x-y\|_2^2)^\beta$ for $c > 0$ and $\beta \in (-1, 0)$. Gorham and Mackey [61, Thm. 8] proved that these IMQ KSDs determine weak convergence on $\mathbb{R}^d$ whenever $P \in \mathcal{P}$, the set of distantly dissipative distributions with Lipschitz $\nabla \log p$. We say P satisfies *distant dissipativity* [46, 62] if $\kappa_0 := \liminf_{r\to\infty} \kappa(r) > 0$ for

$$\kappa(r) = \inf\{-2\tfrac{\langle \nabla \log p(x) - \nabla \log p(y), x-y\rangle}{\|x-y\|_2^2} : \|x-y\|_2 = r\}.$$

An IMQ KSD will serve as the convergence-determining reference MMD for our L^r IMQ fGMMDs developed in Example 5.4.2.

Our next result, proved in Appendix D.1, shows that tilted base kernels of the form $A(x)\Phi(x-y)A(y)$ also determine convergence to $\mathcal{P}$.

Theorem 5.4.1 (Tilted KSDs determine convergence)**.** *Suppose that $P \in \mathcal{P}$ and that $k(x,y) = A(x)\Phi(x-y)A(y)$ for $\Phi \in C^2$ with $F(u) = \sup_{\omega\in\mathbb{R}^d} e^{-\|\omega\|_2^2/(2u^2)}/\hat{\Phi}(\omega)$ finite for all $u > 0$ and $A \in C^1$ with $A > 0$, $1/A \in L^2$, and bounded-Lipschitz $\nabla \log A$. Then, for any sequence of probability measures $(\mu_n)_{n=1}^\infty$, $\mu_n \overset{\mathcal{D}}{\Longrightarrow} P$ whenever $\mathrm{KSD}_k(\mu_n, P) \to 0$.*

Theorem 5.4.1 motivates the new convergence-determining tilted hyperbolic secant kernels introduced in Example 5.4.1.

5.4.2 Relative error bounds

We next turn to developing high-probability relative error bounds for fGMMDs with sub-quadratic sample complexity. Our strategy is to show that the second moment of each fGMMD feature, $w_j(Z, \Delta_N) := |(\mathcal{F}^{-1}\Delta_N k_j^{1/2})(Z)|^r/\nu(Z)$, is bounded by a power of its mean:

Definition 5.4.2 ((C,γ) second moments)**.** Fix a target distribution P and a family of distributions $\mathcal{Q}$. For $Z \sim \nu$ and $j \in [J]$, let $Y_j := w_j(Z, \Delta_N)$. If for some $C > 0$ and $\gamma \in [0,2]$, for all $Q_N \in \mathcal{Q}$,

$$\mathbb{E}[Y_j^2] \le C\mathbb{E}[Y_j]^{2-\gamma},$$

then we say $(\boldsymbol{k}^{1/2}, r, \nu)$ yields (C,γ) *second moments for P and $\mathcal{Q}$.*

The next proposition, proved in Appendix D.2, demonstrates the value of this second moment property.

Proposition 5.4.3. *If $(\boldsymbol{k}^{1/2}, r, \nu)$ yields (C,γ) second moments for P and $\mathcal{Q}$ and $M \ge 2C\mathbb{E}[Y_j]^{-\gamma}\log(J/\delta)/\epsilon^2$ for all $j \in [J]$, then, for any $Q_N \in \mathcal{Q}$, with probability at least $1-\delta$,*

$$\mathrm{fGMMD}_{\boldsymbol{k}^{1/2},r,\nu,M} \ge (1-\epsilon)^{1/r}\,\mathrm{GMMD}_{\boldsymbol{k}^{1/2},r}\,.$$

When $\|\rho\|_{L^s} < \infty$ and $\mathrm{MMD}_{k_0}^2(Q_N, P) = \Omega(N^{-1})$, Proposition 5.4.3 and the MMD-GMMD inequality Eq. (5.3) imply that a (C,γ) fGMMD dominates its reference MMD with high probability whenever the importance sample size $M = \Omega(N^{\gamma r/2})$. Note that $\mathrm{MMD}_{k_0}^2(Q_N, P) = \Omega_P(N^{-1})$ whenever the sample points $x_1, \dots, x_N$ are drawn i.i.d. from a distribution Q, since the scaled V-statistic $N\,\mathrm{MMD}_{k_0}^2(Q_N, P)$ diverges when $Q \neq P$ and converges in distribution to a non-zero limit when $Q = P$ [124, Thm. 32]. Moreover, working in a hypothesis testing framework of shrinking alternatives, Gretton et al. [63, Thm. 13] showed that $\mathrm{MMD}_{k_0}^2(Q, P) = \Theta(N^{-1})$ was the smallest local departure distinguishable by an asymptotic MMD test.

We next turn to establishing practical sufficient conditions that imply (C,γ) second moments. Notably, any improvements on these second moment bounds translate directly into improved sample complexity bounds for fGMMDs.

Our first result, proved in Appendix D.3, yields $(C, 1)$ moments whenever functions $w_j(z, Q_N)$ are bounded. Let $\mathcal{Q}(\boldsymbol{k}^{1/2}, \nu, C') := \{Q_N \mid \sup_{z,j} w_j(z, Q_N) < C'\}$.

Proposition 5.4.4. *Fix any $C' > 0$. If $\sup_{z,j} w_j(z, P) < \infty$, then for some $C > 0$, $(\boldsymbol{k}^{1/2}, r, \nu)$ yields $(C, 1)$ second moments for P and $\mathcal{Q}(\boldsymbol{k}^{1/2}, \nu, C')$.*

A more refined analysis provides sufficient conditions for (C, γ) second moments for any $\gamma > 0$. The key condition is a structural assumption about $k_j^{1/2}$:

Assumption 5.F. *For all $j \in [J]$, $k_j^{1/2}(x, \omega) = \mathcal{F}(\mathcal{O}_{j,x} f_j(x - \cdot))(\omega)$, where $\mathcal{O}_{j,x}$ is a linear operator and f_j is a symmetric function.*

Let

$$k_j(x, y) := \int k_j^{1/2}(x, \omega) \overline{k_j^{1/2}(y, \omega)} \rho(\omega) \, \mathrm{d}\omega,$$

so if k_0 is of the form Eq. (5.2), $k_0 = \sum_{j=1}^J k_j$. As shown in Appendix D.4, Assumption 5.F leads to a convenient rewriting of the kernel components k_j in terms of the linear operators acting on a stationary kernel:

Proposition 5.4.5. *Under Assumption 5.F,*

$$k_j(x, y) = \mathcal{O}_{j,x} \mathcal{O}_{j,y} \Phi_j(x - y),$$

where $\hat{\Phi}_j := \widehat{f_j}^2 \rho$. Hence

$$(\Delta_N \times \Delta_N) k_0 = \textstyle\sum_{j=1}^J \left\| \int \mathcal{O}_{j,x} \Phi_j(x - \cdot) \Delta_N(\mathrm{d}x) \right\|_{\Phi_j}^2.$$

We also require two additional conditions, which concern how f_j relates to $\hat{\Phi}_j$ and ν relates to f_j. Specifically, f_j must be sufficiently smooth:

Assumption 5.G. *There exists a* smoothness parameter $\overline{\lambda} \in (1/2, 1]$ *such that if $\lambda \in (1/2, \overline{\lambda})$, then $\widehat{f_j} / \hat{\Phi}_j^{\lambda/2} \in L^2$.*

Requiring that $\widehat{f_j} / \hat{\Phi}_j^{\lambda/2} \in L^2$ is equivalent to requiring that f_j belong to the reproducing kernel Hilbert space $\mathcal{K}_\lambda$ induced by the kernel $\mathcal{F}^{-1}(\hat{\Phi}_j^\lambda)$. The smoothness of the functions in $\mathcal{K}_\lambda$ increases as λ increases. Hence $\overline{\lambda}$ quantifies the smoothness of f_j relative to Φ_j.

The importance distribution ν must also be heavy-tailed relative to f_j:

Assumption 5.H. *There exists a* tail parameter $\xi \in (0, 1)$ *such that for $j \in [J]$, $\nu^{-1} \leq C_{\nu,j} f_j^{-\xi r}$ for some $C_{\nu,j} > 0$.*

Assumption 5.H results in over-dispersed features relative to f_j, which ensures that regions of mismatch between P and Q_N are picked up by the fGMMD. In addition to Assumptions 5.F, 5.G and 5.H, our result (proved in Appendix D.5) relies on several other regularity conditions. The statement of these conditions is deferred to Appendix D.5.

Theorem 5.4.6. *Assume that $P \in \mathcal{P}$ and that Assumptions 5.F, 5.G, 5.H, E.14, E.15 and E.16 hold. Let $\mathcal{Q}(b, C_{B,j}, C_{\mathcal{O},j})$ denote the family of distributions for which Assumptions E.17 and E.18 hold for some uniform choice of the constants $b, C_{B,j}$, and $C_{\mathcal{O},j}$. For $\alpha > 2(1-\overline{\lambda})$, let $\gamma_\alpha := \alpha + (2-\alpha)\xi/(2-b-\xi)$. Then for a constant $C_\alpha > 0$, $(\boldsymbol{k}^{1/2}, r, \nu)$ yields $(C_\alpha, \gamma_\alpha)$ second moments for P and $\mathcal{Q}(b, C_{B,j}, C_{\mathcal{O},j})$.*

Theorem 5.4.6 suggests a strategy for improving the importance sample growth rate γ of an fGMMD: increase the smoothness $\overline{\lambda}$ of f_j and decrease the tail parameter ξ to increase the over-dispersion of ν relative to f_j.

5.4.3 Explicit examples

The results of Sections 5.4.1 and 5.4.2 allow us to develop explicit convergence-determining fGMMDs with (C, γ) second moments for any $\gamma > 0$. We give two examples, corresponding to the fGMMDs used in our experiments. In each example, we first fix target smoothness and tail parameters $\overline{\lambda}$ and ξ; these define a range of achievable importance sample growth rates γ via Theorem 5.4.6. We next introduce a convergence-determining reference KSD with Stein kernel Eq. (5.1) written in the canonical form Eq. (5.2); together with a choice of r, the decomposition into $\boldsymbol{k}^{1/2}$ and ρ fully determines the associated GMMD. Finally, we prove both that the assumptions of Theorem 5.4.6 are met when all the f_j are chosen equal and $\nu \propto f_1^{\xi r}$, guaranteeing (C, γ) second moments, and that $\rho = \hat{\Phi}_1 / \widehat{f_1}^2$ has bounded L^t norm, guaranteeing that the GMMD upper bounds the KSD by Eq. (5.3).

Example 5.4.1 (L^2 tilted sech fGMMD)**.** Take $\overline{\lambda} = 1$, fix any $\xi \in (0, 1)$, and choose an inverse scale parameter $a > 0$. Recall that the hyperbolic secant (sech) function is given by $\operatorname{sech}(u) = \frac{2}{e^u + e^{-u}}$. For $x \in \mathbb{R}^D$, define the sech kernel $\Phi_a^{\text{sech}}(x) := \prod_{d=1}^D \operatorname{sech}\big(\sqrt{\frac{\pi}{2}} a x_d\big)$.

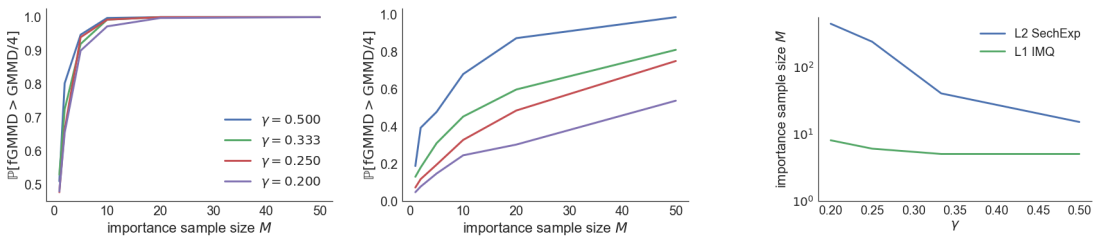
Consider a reference KSD with tilted sech base kernel $k(x, y) = A(x)\Phi_a^{\text{sech}}(x - y)A(y)$, where A is any positive, differentiable function. The induced Stein kernel k_0 from Eq. (5.1) has canonical form Eq. (5.2) with $J = D$ and, for each $j \in [J]$, $f_j = \Phi_{2a}^{\text{sech}}$, $k_j^{1/2}(x, \omega) = \mathcal{F}(\mathcal{T}_{j,x} A(x) f_j(x - \cdot))(\omega)$, and $\rho = (\hat{\Phi}_{2a}^{\text{sech}})^2 / \widehat{f_j}^2 = 1$. As shown in Appendix D.9, with the choice of $r = 2$ and $\nu(z) \propto \Phi_{2a}^{\text{sech}}(z)^{2\xi}$, the resulting L^2 IMQ fGMMD meets all of our relative error criteria.

Theorem 5.4.7 (L^2 tilted sech fGMMD properties)**.** *Under the settings of Example 5.4.1, let $\mathcal{Q}^{\text{sech}}(C)$ denote the family of distributions for which*

$$Q_N((1 + \|\cdot\|_1) A(\cdot) e^{\sqrt{\frac{\pi}{2}} a \|\cdot\|_1}) \leq C.$$

Assume that $P \in \mathcal{P}$ and that for some constants $C_1, \ldots, C_4 \geq 0$,

$$\|\nabla \log p(x)\|_1 \leq C_1 + C_2 \|x\|_1 \qquad \text{and}$$
$$\|\nabla \log A(x)\|_1 \leq C_3 + C_4 \|x\|_1.$$



(a) Efficiency of L1 IMQ (b) Efficiency of L2 SechExp (c) M needed for $\frac{\text{stdev(fGMMD)}}{\text{GMMD}} < \frac{1}{2}$

Figure 5-1: Efficiency of fGMMDs. The L1 IMQ fGMMD displays exceptional efficiency.

Then, in the notation of Theorem 5.4.6, for any $b > 0$, $(\boldsymbol{k}^{1/2}, r, \nu)$ yields $(C_\alpha, \gamma_\alpha)$ second moments for P and $\mathcal{Q}^{\text{sech}}(C)$. Moreover, $\text{MMD}^2_{k_0} = \text{GMMD}^2_{\boldsymbol{k}^{1/2},2}$.

Example 5.4.2 (L^r IMQ fGMMD)**.** Fix any $\overline{\lambda} \in (1/2, 1)$, $\underline{\xi} \in (0, 1/2)$, and $\xi \in (\underline{\xi}, 1)$. Consider a reference KSD with IMQ base kernel $k(x, y) = \Psi^{\text{IMQ}}_{c,\beta}(x - y)$ with $c > 0$ and $\beta \in [-D/2, 0)$. For $c' = \overline{\lambda}c/2$ and any $\beta' \in [-D/(2\underline{\xi}), -\beta/(2\underline{\xi}) - D/(2\underline{\xi}))$, the induced Stein kernel k_0 from Eq. (5.1) has canonical form Eq. (5.2) with $J = D$, and, for each $j \in [J]$, $f_j = \Psi^{\text{IMQ}}_{c',\beta'}$, $k_j^{1/2}(x, \omega) = \mathcal{F}(\mathcal{T}_{j,x} f_j(x - \cdot))(\omega)$, and $\rho = \hat{\Psi}^{\text{IMQ}}_{c,\beta} / \widehat{f_j}^2$. As shown in Appendix D.10, with the choice of $r = -D/(2\beta'\underline{\xi})$ and $\nu(z) \propto \Psi^{\text{IMQ}}_{c',\beta'}(z)^{\xi r}$, the resulting L^r IMQ fGMMD meets all of our relative error criteria.

Theorem 5.4.8 (L^r IMQ fGMMD properties)**.** *Under the settings of Example 5.4.2, let $\mathcal{Q}^{\text{IMQ}}(C)$ denote the family of distributions for which*

$$Q_N(\|\cdot\|_2^{1-2\beta'}) \leq C.$$

Assume that $P \in \mathcal{P}$ and that for some constants $C_1, C_2 \geq 0$,

$$\|\nabla \log p(x)\|_2 \leq C_1 + C_2 \|x\|_2.$$

Then, in the notation of Theorem 5.4.6, for $b = 0$, $(\boldsymbol{k}^{1/2}, r, \nu)$ yields $(C_\alpha, \gamma_\alpha)$ second moments for P and $\mathcal{Q}^{\text{IMQ}}(C)$. Moreover, there exists a constant $C' > 0$ such that for any Q_N

$$\text{MMD}^2_{k_0}(Q_N, P) \leq C' \, \text{GMMD}^2_{\boldsymbol{k}^{1/2},r}(Q_N, P).$$

A particularly simple setting is given by $\beta' = -D/(2\underline{\xi})$ which yields $r = 1$.

5.4.4 Upper bounds on the GMMD and fGMMD

While it is most important that the GMMD and fGMMD upper bound a reference MMD, it is also worthwhile to verify that as $Q_N \to P$, GMMD and fGMMD $\to 0$. We accomplish this by upper bounding the GMMD and fGMMD by the reference

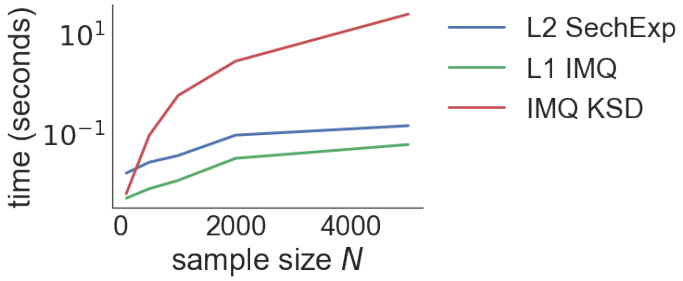


Figure 5-2: Speed of fGMMDs using $M = 10$ importance samples compared to the IMQ KSD. All data had dimension $D = 10$. Even for moderate dataset sizes, fGMMDs are orders of magnitude faster than the KSD.

MMD. We obtain these upper bounds under a subset of the assumptions required by Theorem 5.4.6.

Theorem 5.4.9. *Assume that* $\int_{\mathbb{B}_{\|\cdot\|}(R)^\complement} f_j(z)^{r(1-b)} \mathrm{d}z < G(R)$ *for a decreasing function* G *and that Assumptions 5.F, 5.G and E.15 hold. Let* $V(R) := \mathrm{vol}(\mathbb{B}_{\|\cdot\|}(R))$ *and* $W(R) := G(R)/V(R)$*. Then for any* $\lambda \in (1/2, \overline{\lambda})$ *there exists a* $C, C' > 0$ *such that*

$$\mathrm{GMMD}^2_{\boldsymbol{k}^{1/2},r} \leq C \sum_{j=1}^J G\Big(W^{-1}\big(C' \,\mathrm{MMD}_{k_j}^{r(2\lambda-1)}\big)\Big)^{2/r}.$$

Theorem 5.4.10. *Assume that Assumptions 5.F, 5.G and E.18 hold. Then for any* $\lambda \in (1/2, \overline{\lambda})$ *there exists a* $C > 0$ *such that*

$$\mathrm{fGMMD}^2_{\boldsymbol{k}^{1/2},r,\nu,M} \leq c(\nu, Z_{1:M})^{2/r} \sum_{j=1}^J \mathrm{MMD}_{k_j}^{4\lambda-2},$$

where $c(\nu, Z_{1:M}) := CM^{-1} \sum_{m=1}^M \nu(Z_m)^{-1}$.

The proofs of these upper bound are in, respectively, Appendices D.7 and D.8.

5.4.5 Asymptotics

Recall that in goodness-of-fit testing we have an empirical distribution

$$Q_N = N^{-1} \sum_{n=1}^N \delta_{X_n},$$

where $X_n \overset{\text{i.i.d.}}{\sim} Q$. We wish to determine whether the null hypothesis $H_0 : P = Q$ or alternative hypothesis $H_1 : P \neq Q$ holds. In order to do so, we need to estimate the distribution of the statistic $F_{r,N} := \mathrm{fGMMD}^2_{\boldsymbol{k}^{1/2},r,\nu,M}(Q_N, P)$ under the null hypothesis, where we treat as fixed the features $Z_1, \ldots, Z_M \overset{\text{i.i.d.}}{\sim} \nu$ used to construct the fGMMD. We would also like to verify that, at a minimum, the power of the test approaches 1 as $N \to \infty$. For $r \in [1, 2]$, let $\xi_{r,mj}(x) := \frac{(\mathcal{F}^{-1}(\delta_x - P)k_j^{1/2})(Z_m)}{(M\nu(Z_m))^{1/r}}$, where

$Z_m \overset{\text{i.i.d.}}{\sim} \nu$. Hence $\xi_r(x) \in \mathbb{R}^{MJ}$. The following result, proved in Appendix D.11, furnishes the necessary information.

Theorem 5.4.11 (Asymptotic distribution of fGMMD)**.** *Let $\Sigma_r := \mathrm{Cov}_P(\xi_r)$, which we assume to be finite. Then, under any realization of $Z_m \overset{i.i.d.}{\sim} \nu$, the following holds.*

1. *Under $H_0 : P = Q$,*

$$NF_{r,N} \overset{\mathcal{D}}{\Longrightarrow} \sum_{j=1}^{J} (\sum_{m=1}^{M} |\zeta_{mj}|^r)^{2/r}$$

   as $N \to \infty$, where $\zeta \sim \mathcal{N}(\mathbf{0}, \Sigma_r)$.

2. *Under $H_1 : P \neq Q$, $NF_{r,N} \to \infty$ as $N \to \infty$.*

The next result provides a roadmap for using fGMMD for hypothesis testing and is similar in spirit to [75, Theorem 3].

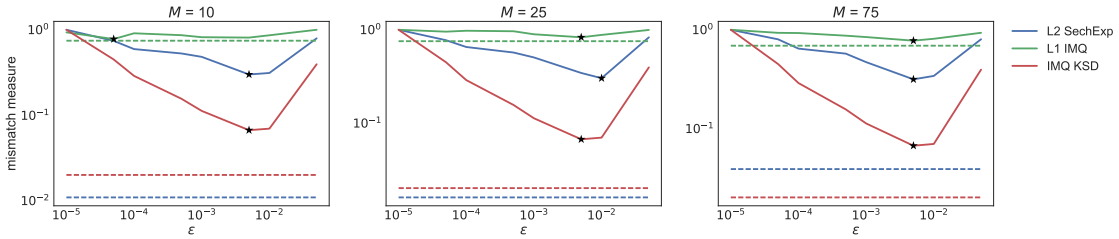
Theorem 5.4.12 (Goodness of fit testing with fGMMD)**.** *Let $\hat{\mu} := N^{-1} \sum_{n=1}^{N} \xi_r(X'_n)$ and $\hat{\Sigma} := N^{-1} \sum_{n=1}^{N} \xi_r(X'_n)\xi_r(X'_n)^\top - \hat{\mu}\hat{\mu}^\top$ with either $X'_n = X_n$ or $X'_n \overset{i.i.d.}{\sim} P$. Suppose for the test $NF_{r,N}$, the test threshold τ_α is set to the $(1-\alpha)$-quantile of the distribution of $\sum_{j=1}^{J} \Big(\sum_{m=1}^{M} |\zeta_{jm}|^r \Big)^{2/r}$, where $\zeta \sim \mathcal{N}(0, \hat{\Sigma})$. Then, under $H_0 : P = Q$, asymptotically the false positive rate is α. Under $H_1 : P \neq Q$, the test power $\mathbb{P}_{H_1}(NT_{p,N} > \tau_\alpha) \to 1$ as $N \to \infty$.*

5.5 Experiments

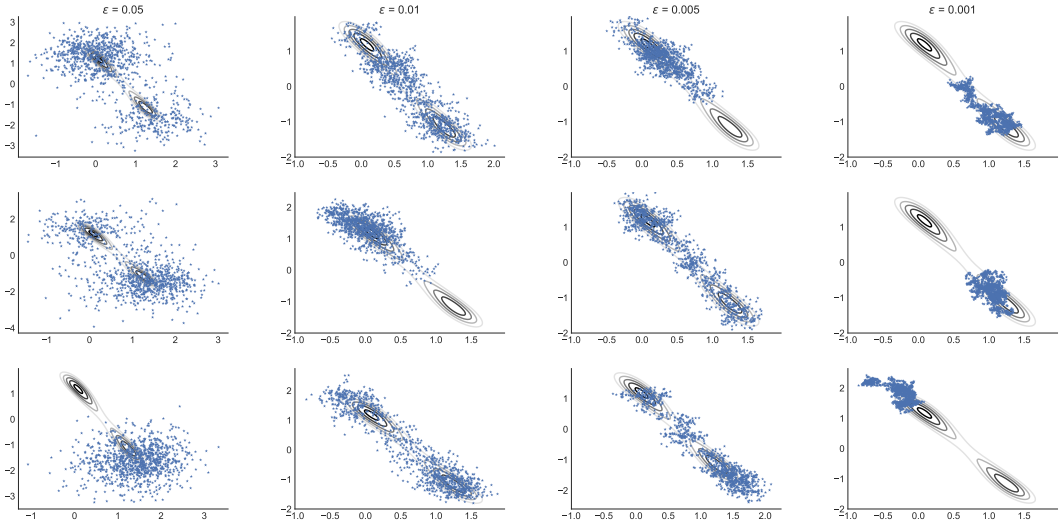
We now investigate the importance-sample and computational efficiency of our proposed fGMMDs and evaluate their benefits in MCMC hyperparameter selection and goodness-of-fit testing. We consider the fGMMDs described in Examples 5.4.1 and 5.4.2: the tilted sech kernel using $r = 2$ and $A(x) = \prod_{d=1}^{D} e^{a'\sqrt{1+x_d^2}}$ (L2 SechExp) and the inverse multiquadric kernel using $r = 1$ (L1 IMQ). We select kernel parameters as follows. First we choose γ and then select $\overline{\lambda}$, α, and ξ so that according to the theory from Section 5.4.2, $(\boldsymbol{k}^{1/2}, r, \nu)$ yields (C, γ) second moments. In particular, we choose $\overline{\lambda} = \gamma/4$, $\alpha = \gamma/3$, and $\xi = \gamma/2$. Except for the importance sample efficiency experiments, where we vary γ explicitly, all experiments use $\gamma = 1/4$. Let $\widehat{\mathrm{med}}_u$ denote the estimated median of the distance between data points under the u-norm, where the estimate is based on using a small subsample of the full dataset. For L2 SechExp, we take $a^{-1} = \sqrt{2\pi}\,\widehat{\mathrm{med}}_1$, except in the sample quality experiments where $a^{-1} = \sqrt{2\pi}$. For L1 IMQ, we take $\beta = -1/2$ and $c = \sqrt{2D}\,\widehat{\mathrm{med}}_2$, except in the sample quality experiments where $c = 1$.

5.5.1 Import sample-efficiency experiments

To validate the importance sample-efficiency theory from Sections 5.4.2 and 5.4.3, we calculated $\mathbb{P}[\text{fGMMD} > \text{GMMD}/4]$ as the importance sample size M was increased.



(a) Step size selection using KSDs and fGMMDs. Dotted lines are divergence measures of the high-quality samples. Both fGMMDs quickly converge to select step sizes consistent with the IMQ KSD.



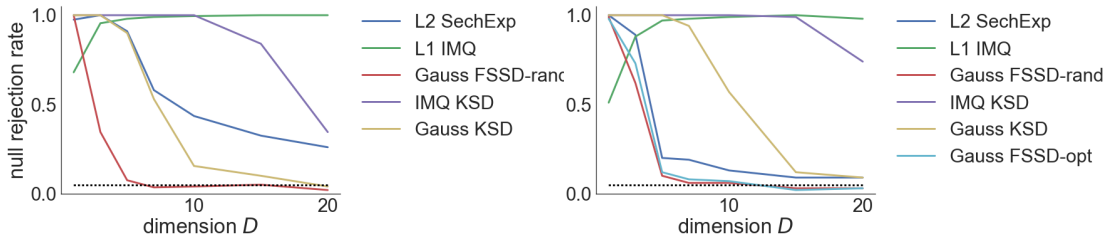
(b) SGLD sample points with equidensity contours of p overlaid. All quality measure selected a step size of $\varepsilon = .01$ or $.005$. The samples produced by SGLD with these step sizes are noticeably better than those produced using smaller or large step sizes.

Figure 5-3: Using fGMMDs for measuring sample quality

We considered choices of the parameters for L2 SechExp and L1 IMQ that produced (C_γ, γ) second moments for varying choices of γ. The results, shown in Figs. 5-1a and 5-1b, indicate greater sample efficiency for L1 IMQ than L2 SechExp. L1 IMQ is also more robust to the choice of γ. Fig. 5-1c, which plots the values of M necessary to for $\frac{\text{stdev(fGMMD)}}{\text{GMMD}} < \frac{1}{2}$, corroborates the greater sample efficiency of L1 IMQ.

5.5.2 Computational complexity experiment

We compared the computational complexity of the fGMMDs (with $M = 10$) to that of the IMQ KSD. We generated datasets of dimension $D = 10$ with the sample size N ranging from 500 to 5000. Even for moderate dataset sizes, the fGMMDs could be computed orders of magnitude faster than the KSD.



(a) Gaussian vs. Laplace. $n = 1000$, variance $= 1$

(b) Gaussian vs. Student's t. $n = 2000$, Student's t has df $= 5$

Figure 5-4: Power of fGMMD, FSSD, and KSD goodness-of-fit tests. Both fGMMDs offer competitive performance.

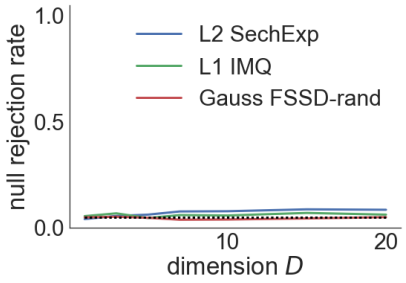


Figure 5-5: Size of fGMMD and FSSD goodness-of-fit tests for Gaussian null with $n = 1000$. All tests were close to calibrated.

5.5.3 Approximate MCMC hyperparameter selection

We follow the stochastic gradient Langevin dynamics [SGLD, 142] hyperparameter selection setup from Gorham and Mackey [60, Section 5.3]. SGLD with constant step size ε is a biased MCMC algorithm that approximates the overdamped Langevin diffusion. No Metropolis-Hastings correction is used and an unbiased estimate of the score function from a data subsample is calculated at each iteration. There is a bias-variance tradeoff in the choice of step size parameter: the stationary distribution of SGLD deviates more from its target as ε grows, but as ε gets smaller the mixing speed of SGLD decreases. Hence, an appropriate choice of ε is critical for accurate posterior inference. We target the bimodal Gaussian mixture model (GMM) posterior of Welling and Teh [142] and compare the step size selection made by the two fGMMDs to that of IMQ KSD [61]. Fig. 5-3a shows that L2 SechExp agrees with IMQ KSD even with just $M = 10$ importance samples while L1 IMQ is in agreement once $M = 25$, with all three measures settling on $\varepsilon = .005$. Fig. 5-3b compares the choice of $\varepsilon = .005$ to smaller and larger values of ε.

5.5.4 Goodness-of-fit testing

Finally, we investigate the performance fGMMDs for goodness-of-fit testing. In all of our experiments we used a standard Gaussian $p(x) = \mathcal{N}(x \mid 0, I)$ as the null distribution while varying the dimension of the data. We explored the power of fGMMD-based tests compared to FSSD [75] and KSD-based tests [35, 61, 86]. There are two types of FSSD tests: FSSD-rand used random sample locations and fixed hyperparameters while FSSD-opt uses a small subset of the data to optimize sample locations and hyperparameters for a power criterion. Our first experiment used $q(x) = \prod_{d=1}^{D} \mathsf{Lap}(x_d|0, 1/\sqrt{2})$, a product of Laplace distributions (see Fig. 5-4a). Our second experiment used $q(x) = \mathcal{T}(x|0, 5)$, a standard multivariate Student's t distribution with 5 degrees of freedom (see Fig. 5-4b). The L1 IMQ test performed better as the dimension increased, with power near 1 once the dimension reached 10 in both experiments. This is an intriguing empirical finding that we cannot fully explain. The L2 SechExp test outperformed FSSD (both randomized and optimized versions in the Student's t case). Only the quadratic-time IMQ KSD and L1 IMQ outperformed L2 SechExp on the Laplace experiment while the two KSDs and L1 IMQ were superior on the Student's t experiment. We also verified the size of the FSSD and fGMMD-based tests (see Fig. 5-5). All tests were close to calibrated despite using asymptotic null distributions.

5.6 Discussion and related work

We have introduced a new family of kernel-based discrepancy measures – GMMDs – designed to upper bound a target MMD. Using importance sampling, we estimate the GMMD and call the resulting estimator an fGMMD. The multiplicative error bounds we develop for fGMMDs only require the number of importance samples M to grow sublinearly in N, which implies that the computational complexity of fGMMDs is subquadratic in N. We validated our approach on two applications where kernel Stein discrepancies have shown excellent performance: measuring sample quality and goodness-of-fit testing. Empirically, the L1 IMQ fGMMD performed particularly well: it was superior to existing "linear-time" KSD approximations and typically performed as well or better than the state-of-the-art quadratic-time KSDs.

While we focused on Stein kernels with the Langevin Stein operator developed in [60, 105], our analyses extend readily to KSDs based on the diffusion Stein operators of Gorham et al. [62]. fGMMDs can also be used as drop-in replacements in other applications featuring expensive MMD computations including two-sample testing [34, 63], Monte Carlo variance reduction with control functionals [105], and probabilistic inference using Stein variational gradient descent [85]. fGMMDs could also be useful in the context of kernel quadrature [see, e.g., 9, 25], where one aims to approximate the expectation of target functions (in a weighted L^p norm sense) instead of approximating an MMD. For example, Bach [9] studies approximations to target functions in reproducing kernel Hilbert spaces but does not consider the function classes associated with more general GMMDs. An interesting direction for

future work would be the development of asymptotically consistent power estimates for fGMMD-based goodness-of-fit tests.

Appendix A

Chapter 2 Proofs

A.1 Marginal Likelihood Approximation

Proof By the assumption that $\mathcal{L}$ and $\tilde{\mathcal{L}}$ are non-positive, the multiplicative error assumption, and Jensen's inequality,

$$\tilde{\mathcal{E}} = \int e^{\tilde{\mathcal{L}}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \geq \int e^{(1+\varepsilon)\mathcal{L}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \geq \left(\int e^{\mathcal{L}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\right)^{1+\varepsilon} = \mathcal{E}^{1+\varepsilon}$$

and

$$\tilde{\mathcal{E}} = \int e^{\tilde{\mathcal{L}}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \leq \int e^{(1-\varepsilon)\mathcal{L}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \leq \left(\int e^{\mathcal{L}(\boldsymbol{\theta})}\pi_0(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\right)^{1-\varepsilon} = \mathcal{E}^{1-\varepsilon}.$$

□

A.2 Main Results

In order to construct coresets for logistic regression, we will use the framework developed by Feldman and Langberg [50]. For $n \in [N] := \{1, \ldots, N\}$, let $f_n : \mathcal{S} \to \mathbb{R}_+$ be a non-negative function from some set $\mathcal{S}$ and let $\bar{f} = \frac{1}{N}\sum_{n=1}^{N} f_n$ be the average of the functions. Define the *sensitivity* of $n \in [N]$ with respect to $\mathcal{S}$ by

$$\sigma_n(\mathcal{S}) := \sup_{s \in S} \frac{f_n(s)}{\bar{f}(s)},$$

and note that $\sigma_n(\mathcal{S}) \leq N$. Also, for the set $\mathcal{F} := \{f_n \,|\, n \in [N]\}$, define the dimension $\dim(\mathcal{F})$ of $\mathcal{F}$ to be the minimum integer h such that

$$\forall F \subseteq \mathcal{F},\ |\{F \cap R \,|\, R \in \mathbf{ranges}(\mathcal{F})\}| \leq (|F|+1)^h,$$

where $\mathbf{ranges}(\mathcal{F}) := \{\mathbf{range}(s,a) | s \in \mathcal{S}, a \geq 0\}$ and $\mathbf{range}(s,a) := \{f \in \mathcal{F} \,|\, f(s) \leq a\}$.

We make use of the following:

Theorem A.2.1 (Bachem et al. [11], Braverman et al. [24], Feldman and Langberg [50]). *Fix $\varepsilon > 0$. For $n \in [N]$, let $m_n \in \mathbb{R}_+$ be chosen such that*

$$m_n \geq \sigma_n(\mathcal{S})$$

and let $\bar{m}_N := \frac{1}{N}\sum_{n=1}^{N} m_n$. There is a universal constant c such that if $\mathcal{C}$ is a sample from $\mathcal{F}$ of size

$$|\mathcal{C}| \geq \frac{c\,\bar{m}_N^2}{\varepsilon^2}(\dim(\mathcal{F}) + \ln(1/\delta)),$$

such that the probability that each element of $\mathcal{C}$ is selected independently from $\mathcal{F}$ with probability $\frac{m_n}{N\bar{m}_N}$ that f_n is chosen, then with probability at least $1-\delta$, for all $s \in \mathcal{S}$,

$$\left|\bar{f}(s) - \frac{\bar{m}_N}{|\mathcal{C}|}\sum_{f\in\mathcal{C}}\frac{f(s)}{m_n}\right| \leq \varepsilon\bar{f}(s).$$

The set $\mathcal{C}$ in the theorem is called a *coreset*. In our application to logistic regression, $\mathcal{S} = \Theta$ and $f_n(\boldsymbol{\theta}) = -\ln p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})$. The key is to determine $\dim(\mathcal{F})$ and to construct the values m_n efficiently. Furthermore, it is necessary for $\bar{m}_N = o(\sqrt{N})$ at a minimum and preferable for $\bar{m}_N = O(1)$.

Letting $\mathbf{z}_n = y_n\mathbf{x}_n$ and $\phi(s) = \ln(1+\exp(-s))$, we can rewrite $f_n(\boldsymbol{\theta}) = \phi(\mathbf{z}_n \cdot \boldsymbol{\theta})$. Hence, the goal is to find an upper bound

$$m_n \geq \sigma_n(\Theta) = \sup_{\boldsymbol{\theta}\in\Theta}\frac{N\,\phi(\mathbf{z}_n\cdot\boldsymbol{\theta})}{\sum_{n'=1}^{N}\phi(\mathbf{z}_{n'}\cdot\boldsymbol{\theta})}.$$

To obtain an upper bound on the sensitivity, we will take $\Theta = \mathbb{B}_R$ for some $R > 0$.

Lemma A.2.2. *For all $a, b \in \mathbb{R}$, $\phi(a)/\phi(b) \leq e^{|a-b|}$.*

Proof The lemma is trivial when $a = b$. Let $\Delta = b - a \neq 0$ and $\rho(a) = \phi(a)/\phi(a+\Delta)$. We have

$$\rho'(a) = \frac{(1+e^a)\log(1+e^{-a}) - (1+e^{a+\Delta})\log(1+e^{-a-\Delta})}{(1+e^a)(1+e^{a+\Delta})\log^2(1+e^{-a-\Delta})}.$$

Examining the previous display we see that $\operatorname{sgn}(\rho'(a)) = \operatorname{sgn}(\Delta)$. Hence if $\Delta > 0$,

$$\begin{aligned}
\sup_a \frac{\phi(a)}{\phi(a+\Delta)} &= \lim_{a\to\infty}\frac{\phi(a)}{\phi(a+\Delta)} \\
&= \lim_{a\to\infty}\frac{\phi'(a)}{\phi'(a+\Delta)} \\
&= \lim_{a\to\infty}\frac{e^{-a}}{1+e^{-a}}\frac{1+e^{-a-\Delta}}{e^{-a-\Delta}} \\
&= e^{\Delta} = e^{|b-a|},
\end{aligned}$$

where the second equality follows from L'Hospital's rule. Similarly, if $\Delta < 0$,

$$\begin{aligned}\sup_a \frac{\phi(a)}{\phi(a+\Delta)} &= \lim_{a\to-\infty} \frac{e^{-a}}{1+e^{-a}} \frac{1+e^{-a-\Delta}}{e^{-a-\Delta}} \\ &= \lim_{a\to-\infty} e^{\Delta} \frac{e^{-a}}{e^{-a-\Delta}} \\ &= 1 \le e^{|b-a|},\end{aligned}$$

where in this case we have used L'Hospital's rule twice. □

Lemma A.2.3. *The function $\phi(s)$ is convex.*

Proof A straightforward calculation shows that $\phi''(s) = \frac{e^s}{(1+e^s)^2} > 0$. □

Lemma A.2.4. *For a random vector $\mathbf{Z} \in \mathbb{R}^d$ with finite mean $\bar{\mathbf{Z}} = \mathbb{E}[\mathbf{Z}]$ and a fixed vectors $\mathbf{V}, \boldsymbol{\theta}^* \in \mathbb{R}^d$,*

$$\inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \mathbb{E}\left[\frac{\phi(\mathbf{Z}\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))}{\phi(\mathbf{V}\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))}\right] \ge e^{-R\|\bar{\mathbf{Z}}-\mathbf{V}\|_2 - |(\bar{\mathbf{Z}}-\mathbf{V})\cdot\boldsymbol{\theta}^*|}.$$

Proof Using Lemmas A.2.2 and A.2.3, Jensen's inequality, and the triangle inequality, we have

$$\begin{aligned}\inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \mathbb{E}\left[\frac{\phi(\mathbf{Z}\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))}{\phi(\mathbf{V}\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))}\right] &\ge \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \frac{\phi(\mathbb{E}[\mathbf{Z}]\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))}{\phi(\mathbf{V}\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*))} \\ &\ge \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} e^{-|(\bar{\mathbf{Z}}-\mathbf{V})\cdot(\boldsymbol{\theta}+\boldsymbol{\theta}^*)|} \\ &\ge \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} e^{-|(\bar{\mathbf{Z}}-\mathbf{V})\cdot\boldsymbol{\theta}| - |(\bar{\mathbf{Z}}-\mathbf{V})\cdot\boldsymbol{\theta}^*)|} \\ &= e^{-R\|\bar{\mathbf{Z}}-\mathbf{V}\|_2 - |(\bar{\mathbf{Z}}-\mathbf{V})\cdot\boldsymbol{\theta}^*|}.\end{aligned}$$

□

We now prove the following generalization of Lemma 2.2.1

Lemma A.2.5. *For any k-clustering $\mathcal{Q}$, $\boldsymbol{\theta}^* \in \mathbb{R}^d$, and $R > 0$,*

$$\sigma_n(\boldsymbol{\theta}^* + \mathbb{B}_R) \le m_n := \left\lceil \frac{N}{1 + \sum_{i=1}^k |G_i^{(-n)}| e^{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2 - |(\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n)\cdot\boldsymbol{\theta}^*|}} \right\rceil.$$

Furthermore, m_n can be calculated in $O(k)$ time.

Proof Straightforward manipulations followed by an application of Lemma A.2.4

yield

$$\begin{aligned}
\sigma_n(\boldsymbol{\theta}^* + \mathbb{B}_R)^{-1} &= \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \frac{1}{N} \sum_{n'=1}^{N} \frac{\phi(\mathbf{z}_{n'} \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))}{\phi(\mathbf{z}_n \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))} \\
&= \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \frac{1}{N} \left[1 + \sum_{i=1}^{k} \sum_{\mathbf{z}\in G_i^{(-n)}} \frac{\phi(\mathbf{z} \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))}{\phi(\mathbf{z}_n \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))} \right] \\
&= \inf_{\boldsymbol{\theta}\in\mathbb{B}_R} \frac{1}{N} \left[1 + \sum_{i=1}^{k} |G_i^{(-n)}| \, \mathbb{E}\left[\frac{\phi(\mathbf{Z}_{G,i}^{(-n)} \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))}{\phi(\mathbf{z}_n \cdot (\boldsymbol{\theta} + \boldsymbol{\theta}^*))} \right] \right] \\
&\geq \frac{1}{N} \left[1 + \sum_{i=1}^{k} |G_i^{(-n)}| e^{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2 - |(\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n)\cdot\boldsymbol{\theta}^*|} \right].
\end{aligned}$$

To see that the bound can be calculated in $O(k)$ time, first note that the cluster i_n to which $\mathbf{z}_n$ belongs can be found in $O(k)$ time while $\bar{\mathbf{Z}}_{G,i_n}^{(-n)}$ can be calculated in $O(1)$ time. For $i \neq i_n$, $G_i^{(-n)} = G_i$, so $\bar{\mathbf{Z}}_{G,i}^{(-n)}$ is just the mean of cluster i, and no extra computation is required. Finally, computing the sum takes $O(k)$ time. □

In order to obtain an algorithm for generating coresets for logistic regression, we require a bound on the dimension of the range space constructed from the examples and logistic regression likelihood.

Proposition A.2.6. *The set of functions* $\mathcal{F} = \{f_n(\boldsymbol{\theta}) = \phi(\mathbf{z}_n \cdot \boldsymbol{\theta}) \,|\, n \in [N]\}$ *satisfies* $\dim(\mathcal{F}) \leq d + 1$.

Proof For all $F \subseteq \mathcal{F}$,

$$|\{F \cap R \,|\, R \in \mathbf{ranges}(\mathcal{F})\}| = |\{\mathbf{range}(F, \boldsymbol{\theta}, a) \,|\, \boldsymbol{\theta} \in \Theta, a \geq 0\}|,$$

where $\mathbf{range}(F, \boldsymbol{\theta}, a) := \{f_n \in \mathcal{F} \,|\, f_n(\boldsymbol{\theta}) \leq a\}$. But, since ϕ is invertible and monotonic,

$$\begin{aligned}
\{f_n \in \mathcal{F} \,|\, f_n(\boldsymbol{\theta}) \leq a\} &= \{f_n \in \mathcal{F} \,|\, \phi(\mathbf{z}_n \cdot \boldsymbol{\theta}) \leq a\} \\
&= \{f_n \in \mathcal{F} \,|\, \mathbf{z}_n \cdot \boldsymbol{\theta} \leq \phi^{-1}(a)\},
\end{aligned}$$

which is exactly a set of points shattered by the hyperplane classifier $\mathbf{z} \mapsto \operatorname{sgn}(\mathbf{z}\cdot\boldsymbol{\theta} - b)$, with $b := \phi^{-1}(a)$. Since the VC dimension of the hyperplane concept class is $d+1$, it follows that [79, Lemmas 3.1 and 3.2]

$$\begin{aligned}
|\{\mathbf{range}(F, \boldsymbol{\theta}, a) \,|\, \boldsymbol{\theta} \in \Theta, a \geq 0\}| &\leq \sum_{j=0}^{d+1} \binom{|F|}{j} \leq \sum_{j=0}^{d+1} \frac{|F|^j}{j!} \\
&\leq \sum_{j=0}^{d+1} \binom{d+1}{j} |F|^j = (|F| + 1)^{d+1}.
\end{aligned}$$

□

Proof Combine Theorem A.2.1, Lemma 2.2.1, and Proposition A.2.6. The algorithm has overall complexity $O(Nk)$ since it requires $O(Nk)$ time to calculate the sensitivities by Lemma 2.2.1 and $O(N)$ time to sample the coreset. □

A.3 Sensitivity Lower Bounds

Lemma A.3.1. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^{d-1}$ be unit vectors such that for some $\epsilon > 0$, for all $k \neq k'$, $\mathbf{v}_k \cdot \mathbf{v}_{k'} \leq 1 - \epsilon$. Then for $0 < \delta < \sqrt{1/2}$, there exist unit vectors $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathbb{R}^d$ such that*

- *for $k \neq k'$, $\mathbf{z}_k \cdot \mathbf{z}_{k'} \geq 1 - 2\delta^2 > 0$*
- *for $k = 1, \ldots, K$ and $\alpha > 0$, there exists $\boldsymbol{\theta}_k \in \mathbb{R}^d$ such that $\|\boldsymbol{\theta}\|_2 \leq \sqrt{2}\delta\alpha$, $\boldsymbol{\theta}_k \cdot \mathbf{z}_k = -\frac{\alpha\epsilon\delta^2}{2}$ and for $k \neq k$, $\boldsymbol{\theta}_k \cdot \mathbf{z}_{k'} \geq \frac{\alpha\epsilon\delta^2}{2}$.*

Proof Let $\mathbf{z}_k$ be defined such that $z_{ki} = \delta v_{ki}$ for $i = 1, \ldots, d-1$ and $z_{kd} = \sqrt{1 - \delta^2}$. Thus, $\|\mathbf{z}_k\|_2 = 1$ and for $k \neq k'$,

$$\mathbf{z}_k \cdot \mathbf{z}_{k'} = \delta^2 \mathbf{v}_k \cdot \mathbf{v}_{k'} + 1 - \delta^2 \geq 1 - 2\delta^2$$

since $\mathbf{v}_k \cdot \mathbf{v}_{k'} \geq -1$. Let $\boldsymbol{\theta}_k$ be such that $\theta_{ki} = -\alpha\delta v_{ki}$ for $i = 1, \ldots, d-1$ and $\theta_{kd} = \frac{\alpha\delta^2(1-\epsilon/2)}{\sqrt{1-\delta^2}}$. Hence,

$$\boldsymbol{\theta}_k \cdot \boldsymbol{\theta}_k = \alpha^2\delta^2 \left(\mathbf{v}_k \cdot \mathbf{v}_k + \frac{(1-\epsilon/2)^2\delta^2}{1-\delta^2} \right) \leq 2\alpha^2\delta^2$$

$$\boldsymbol{\theta}_k \cdot \mathbf{z}_k = \alpha(-\delta^2 \mathbf{v}_k \cdot \mathbf{v}_k + \delta^2(1-\epsilon/2)) = -\frac{\alpha\epsilon\delta^2}{2},$$

and for $k' \neq k$,

$$\boldsymbol{\theta}_k \cdot \mathbf{z}_{k'} = \alpha(-\delta^2 \mathbf{v}_k \cdot \mathbf{v}_{k'} + \delta^2(1-\epsilon/2)) \geq \alpha\delta^2(-1 + \epsilon + 1 - \epsilon/2) = \frac{\alpha\epsilon\delta^2}{2}.$$

□

Proposition A.3.2. *Let $\mathbf{v}_1, \ldots, \mathbf{v}_K \in \mathbb{R}^{d-1}$ be unit vectors such that for some $\epsilon > 0$, for all $k \neq k'$, $\mathbf{v}_k \cdot \mathbf{v}_{k'} \leq 1 - \epsilon$. Then for any $0 < \epsilon' < 1$, there exist unit vectors $\mathbf{z}_1, \ldots, \mathbf{z}_K \in \mathbb{R}^d$ such that for k, k', $\mathbf{z}_k \cdot \mathbf{z}_{k'} \geq 1 - \epsilon'$ but for any $R > 0$,*

$$\sigma_k(\mathbb{B}_R) \geq \frac{K}{1 + (K-1)e^{-R\epsilon\sqrt{\epsilon'}/4}},$$

and hence $\sigma_k(\mathbb{R}^d) = K$.

Proof Let $\mathbf{z}_1, \dots, \mathbf{z}_K \in \mathbb{R}^d$ be as in Lemma A.3.1 with δ such that $\delta^2 = \epsilon'/2$. Since for $s \geq 0$, $\phi(s)/\phi(-s) \leq e^{-s}$, conclude that, choosing α such that $\sqrt{2}\alpha\delta = R$, we have

$$\begin{aligned}
\sigma_n(\mathbb{B}_R) &= \sup_{\boldsymbol{\theta} \in \mathbb{B}_R} \frac{K\,\phi(\mathbf{z}_k \cdot \boldsymbol{\theta})}{\sum_{k'=1}^K \phi(\mathbf{z}_{k'} \cdot \boldsymbol{\theta})} \\
&\geq \frac{K\,\phi(-\alpha\epsilon\delta^2/2)}{\phi(-\alpha\epsilon\delta^2/2) + (K-1)\phi(\alpha\epsilon\delta^2/2)} \\
&\geq \frac{K}{1 + (K-1)e^{-\alpha\epsilon\delta^2/2}} \\
&= \frac{K}{1 + (K-1)e^{-R\epsilon\sqrt{\epsilon'}/4}}.
\end{aligned}$$

□

Proof Choose $\mathbf{v}_1, \dots, \mathbf{v}_N \in \mathbb{R}^{d-1}$ to be any N distinct unit vectors. Apply Proposition A.3.2 with $K = N$ and $\epsilon = 1 - \max_{n \neq n'} \mathbf{v}_n \cdot \mathbf{v}_{n'} > 0$. □

Proof First note that if $\mathbf{v}$ is uniformly distributed on $\mathbb{S}^d$, then the distribution of $\mathbf{v} \cdot \mathbf{v}'$ does not depend on the distribution of V' since $\mathbf{v} \cdot \mathbf{v}'$ and $\mathbf{v} \cdot \mathbf{v}''$ are equal in distribution for all $\mathbf{v}', \mathbf{v}'' \in \mathbb{S}^d$. Thus it suffices to take $v_1' = 1$ and $v_i' = 0$ for all $i = 2, \dots, d$. Hence the distribution of $\mathbf{v} \cdot \mathbf{v}'$ is equal to the distribution of v_1. The CDF of v_1 is easily seen to be proportional to the surface area (SA) of $C_s := \{\mathbf{v} \in \mathbb{S}^d \,|\, v_1 \leq s\}$. That is, $\mathbb{P}[v_1 \leq s] = \mathrm{SA}(C_s)/\mathrm{SA}(C_1)$. Let $U \sim \mathsf{Beta}(\frac{d-1}{2}, \frac{1}{2})$, and let $B(a,b)$ be the beta function. It follows from [84, Eq. 1], that by setting $s = 1 - \epsilon$ with $\epsilon \in [0, 1/2]$,

$$\begin{aligned}
\mathbb{P}[v_1 \geq 1 - \epsilon] &= \frac{1}{2}\mathbb{P}[-\sqrt{1-U} \leq \epsilon - 1] \\
&= \frac{1}{2}\mathbb{P}[U \leq 2\epsilon - \epsilon^2] \\
&= \frac{1}{2B(\frac{d-1}{2}, \frac{1}{2})} \int_0^{2\epsilon-\epsilon^2} t^{(d-3)/2}(1-t)^{-1/2}\,\mathrm{d}t \\
&\leq \frac{1}{2B(\frac{d-1}{2}, \frac{1}{2})}(1-\epsilon)^{-1} \int_0^{2\epsilon-\epsilon^2} t^{(d-3)/2}\,\mathrm{d}t \\
&= \frac{1}{(d-1)B(\frac{d-1}{2}, \frac{1}{2})} \frac{(2-\epsilon)^{(d-1)/2}}{1-\epsilon} \epsilon^{(d-1)/2} \\
&\leq \frac{2^{(d+1)/2}}{(d-1)B(\frac{d-1}{2}, \frac{1}{2})} \epsilon^{(d-1)/2}.
\end{aligned}$$

Applying a union bound over the $\binom{d}{2}$ distinct vector pairs completes the proof. □

Lemma A.3.3 (Hoeffding's inequality [23, Theorem 2.8])**.** *Let A_k be zero-mean,*

independent random variables with $A_k \in [-a, a]$. Then for any $t > 0$,

$$\mathbb{P}\left(\sum_{k=1}^{K} A_k \geq t\right) \leq e^{-\frac{t^2}{2a^2K}}.$$

Proof We say that unit vectors $\mathbf{v}$ and $\mathbf{v}'$ are $(1-\epsilon)$-*orthogonal* if $|\mathbf{v} \cdot \mathbf{v}'| \leq 1 - \epsilon$. Clearly $\|\mathbf{v}_n\|_2 = 1$. For $n \neq n'$, by Hoeffding's inequality $\mathbb{P}(|\mathbf{v}_n \cdot \mathbf{v}_{n'}| \geq 1 - \epsilon) \leq 2e^{-(1-\epsilon)^2 D/2}$. Applying a union bound to all $\binom{K}{2}$ pairs of vectors, the probability that any pair is not $(1-\epsilon)$-orthogonal is at most

$$2\binom{K}{2} e^{-(1-\epsilon)^2 D/2} \leq \frac{1}{2}.$$

Thus, with probability at least $\frac{1}{2}$, $\mathbf{v}_1, \ldots, \mathbf{v}_N$ are pairwise $(1-\epsilon)$-orthogonal. □

Proof The data from Theorem 2.2.4 satisfies $\mathbf{z}_n \cdot \mathbf{z}_{n'} \geq 1 - \epsilon'$, so for $n \neq n'$,

$$\|\mathbf{z}_n - \mathbf{z}_{n'}\|_2^2 = 2 - 2\mathbf{z}_n \cdot \mathbf{z}_{n'} \leq 2\epsilon'.$$

Applying Lemma 2.2.1 with the clustering $\mathcal{Q} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ and combining it with the lower bound in Theorem 2.2.4 yields the result. □

A.4 A Priori Expected Sensitivity Upper Bounds

Proof First, fix the number of datapoints $N \in \mathbb{N}$. Since $\mathbf{x}_n$ are generated from a mixture, let L_n denote the integer mixture component from which $\mathbf{x}_n$ was generated, let C_i be the set of integers $1 \leq j \leq N$ with $j \neq n$ and $L_j = i$, and let $C = (C_i)_{i=1}^{\infty}$. Note that with this definition, $|G_i^{(-n)}| = |C_i|$. Using Jensen's inequality and the upper bound from Lemma 2.2.1 with the clustering induced by the label sequence,

$$\begin{aligned}
\mathbb{E}[\sigma_n(\mathbb{B}_R)] \leq \mathbb{E}[m_n] &= N\mathbb{E}\left[\frac{1}{1 + \sum_i |C_i| e^{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2}}\right] \\
&= N\mathbb{E}\left[\mathbb{E}\left[\frac{1}{1 + \sum_i |C_i| e^{-R\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2}} \,|\, C\right]\right] \\
&\leq N\mathbb{E}\left[\frac{1}{1 + \sum_i |C_i| e^{-R\mathbb{E}\left[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2 \,|\, C\right]}}\right].
\end{aligned}$$

Using Jensen's inequality again and conditioning on the labels $\mathbf{y}$ and indicator L_n,

$$\mathbb{E}\left[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2 \,|\, C\right] \leq \sqrt{\mathbb{E}\left[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2^2 \,|\, C\right]}$$

$$= \sqrt{\mathbb{E}\Big[\mathbb{E}\Big[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2^2 \,|\, C, L_n, \mathbf{y}\Big] \,|\, C\Big]}.$$

For fixed labels $\mathbf{y}$ and clustering C, L_n, the linear combination in the expectation is multivariate normal with

$$\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n \sim \mathcal{N}\left(\frac{1}{|C_i|}\left(\sum_{j \in C_i} y_j\right)\mu_i - y_n\mu_n', \frac{1}{|C_i|}\Sigma_i + \Sigma_n'\right),$$

where μ_n', Σ_n' are the mean and covariance of the mixture component that generated $\mathbf{x}_n$. Further, for any multivariate normal random vector $\mathbf{W} \in \mathbb{R}^d$,

$$\mathbb{E}\big[\mathbf{W}^T\mathbf{W}\big] = \sum_{m=1}^{d} \mathbb{E}\big[W_m^2\big] = \sum_{m=1}^{d} \mathrm{Var}[W_m] + \mathbb{E}[W_m]^2,$$

so

$$\begin{aligned}
&\mathbb{E}\Big[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2^2 \,|\, L_n, C, \mathbf{y}\Big] \\
&= \mathrm{Tr}\left[\frac{1}{|C_i|}\Sigma_i + \Sigma_n'\right] + \left(\frac{\sum_{j \in C_i} y_j}{|C_i|}\right)^2 \mu_i^T\mu_i - 2Y_n\left(\frac{\sum_{j \in C_i} y_j}{|C_i|}\right)\mu_i^T\mu_n' + {\mu_n'}^T\mu_n'.
\end{aligned}$$

Exploiting the i.i.d.-ness of y_j for $j \in C_i$ given C, defining $\bar{y}_j = \mathbb{E}[y_i | L_i = j]$, and noting that $\mathbf{x}_n$ is sampled from the mixture model,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}\Big[\|\bar{\mathbf{Z}}_{G,i}^{(-n)} - \mathbf{z}_n\|_2^2 \,|\, L_n, C, \mathbf{y}\Big] \,|\, C\Big] \\
=&\sum_j \pi_j\left(\mathrm{Tr}\left[\frac{1}{|C_i|}\Sigma_i + \Sigma_j\right] + \frac{|C_i|\bar{y}_i^2 + 1 - \bar{y}_i^2}{|C_i|}\mu_i^T\mu_i - 2\bar{y}_j\bar{y}_i\mu_i^T\mu_j + \mu_j^T\mu_j\right) \\
=&\sum_j \pi_j\left(\frac{\mathrm{Tr}[\Sigma_i] + (1-\bar{y}_i^2)\mu_i^T\mu_i}{|C_i|} + \mathrm{Tr}[\Sigma_j] + \bar{y}_i^2\mu_i^T\mu_i - 2\bar{y}_j\bar{y}_i\mu_i^T\mu_j + \mu_j^T\mu_j\right) \\
=&A_i|C_i|^{-1} + B_{in},
\end{aligned}$$

where A_i and B_i are positive constants

$$\begin{aligned}
A_i &= \mathrm{Tr}[\Sigma_i] + \big(1 - \bar{y}_i^2\big)\mu_i^T\mu_i \\
B_i &= \sum_j \pi_j\big(\mathrm{Tr}[\Sigma_j] + \bar{y}_i^2\mu_i^T\mu_i - 2\bar{y}_i\bar{y}_j\mu_i^T\mu_j + \mu_j^T\mu_j\big).
\end{aligned}$$

Therefore, with 0^{-1} defined to be $+\infty$,

$$\mathbb{E}[m_n] \le N\mathbb{E}\left[\frac{1}{1 + \sum_i |C_i| e^{-R\sqrt{A_i|C_i|^{-1} + B_i}}}\right].$$

As $N \to \infty$, we expect the values of $|C_i|/N$ to concentrate around π_i. To get a finite sample bound using this intuition, we split the expectation into two conditional expectations: one where all $|C_i|/N$ are not too far from π_i, and one where they may be. Define $g : \mathbb{R}_+^\infty \to \mathbb{R}_+$ as

$$g(x) = \frac{1}{1 + \sum_i x_i e^{-R\sqrt{A_i x_i^{-1} + B_i}}},$$

$\pi = (\pi_1, \pi_2, \dots)$, $\epsilon = (\epsilon_1, \epsilon_2, \dots)$ with $\epsilon_i > 0$, and $\eta_i = \max(\pi_i - \epsilon_i, 0)$. Then

$$\begin{aligned}
\mathbb{E}[m_n] &\le N\mathbb{P}\left(\forall i, \frac{|C_i|}{N} \ge \eta_i\right) g(N\eta) + N\mathbb{P}\left(\exists i : \frac{|C_i|}{N} < \eta_i\right) \\
&= Ng(N\eta) + N\mathbb{P}\left(\exists i : \frac{|C_i|}{N} < \eta_i\right)(1 - g(N\eta)).
\end{aligned}$$

Using the union bound, noting that $1 - g(N\eta) \le 1$, and then using Hoeffding's inequality yields

$$\begin{aligned}
\mathbb{E}[m_n] &\le Ng(N\eta) + N\sum_i \mathbb{P}\left(\frac{|C_i|}{N} < \eta_i\right) \\
&\le Ng(N\eta) + N\sum_{i:\pi_i > \epsilon_i} \mathbb{P}\left(\frac{|C_i|}{N} - \pi_i < -\epsilon_i\right) \\
&\le Ng(N\eta) + N\sum_{i:\pi_i > \epsilon_i} e^{-2N\epsilon_i^2} \\
&= \frac{1}{N^{-1} + \sum_i \eta_i e^{-R\sqrt{A_i N^{-1} \eta_i^{-1} + B_i}}} + \sum_{i:\pi_i > \epsilon_i} N e^{-2N\epsilon_i^2}.
\end{aligned}$$

We are free to pick ϵ as a function of π and N. Let $\epsilon = N^{-r}$ for any $0 < r < 1/2$. Note that this means $\eta_i = \max(\pi_i - N^{-r}, 0)$. Then

$$\mathbb{E}[m_n] = \frac{1}{N^{-1} + \sum_i \eta_i e^{-R\sqrt{A_i N^{-1} \eta_i^{-1} + B_i}}} + \sum_{i:\eta_i > 0} N e^{-2N^{1-2r}}.$$

It is easy to see that the first term converges to $\left(\sum_i \pi_i e^{-R\sqrt{B_i}}\right)^{-1}$ by a simple asymptotic analysis. To show the second term converges to 0, note that for all N,

$$\begin{aligned}
\sum_i \pi_i &= \sum_{i:\pi_i > N^{-r}} \pi_i + \sum_{i:\pi_i \le N^{-r}} \pi_i \\
&\ge \sum_{i:\pi_i > N^{-r}} \pi_i \\
&\ge \sum_{i:\pi_i > N^{-r}} N^{-r}
\end{aligned}$$

$$= \left|\left\{i : \pi_i > N^{-r}\right\}\right| N^{-r}.$$

Since $\sum_i \pi_i = 1 < \infty$, $|\{i : \pi_i > N^{-r}\}| = O(N^r)$. Therefore there exists constants $C, M < \infty$ such that

$$\left|\left\{i : \pi_i > N^{-r}\right\}\right| \leq M + CN^r,$$

and thus

$$\sum_{i:\pi_i > N^{-r}} Ne^{-2N^{1-2r}} \leq N(M + CN^r)e^{-2N^{1-2r}} \to 0, \qquad N \to \infty.$$

Finally, since $\bar{m}_N = \frac{1}{N}\sum_{n=1}^{N} m_n$, we have $\mathbb{E}[\bar{m}_N] = \mathbb{E}[m_n]$, and the result follows. □

Proof This is a direct result of Proposition 2.2.8 with $\pi_1 = 1$, $\pi_i = 0$ for $i \geq 2$. □

Appendix B

Chapter 3 Proofs

B.1 Chebyshev Approximation Results

We begin by summarizing some standard results on the approximation accuracy of Chebyshev polynomials. Let $\phi : [-1, 1] \to \mathbb{R}$ be a continuous function, and let ϕ_M be the M-th order Chebyshev approximation to ϕ. Let $\|f\|_\infty := \sup_s |f(s)|$ be the L^∞ norm of a function f; let $\mathbb{C}$ denote the set of complex numbers; and let $|z|$ be the absolute value of $z \in \mathbb{C}$.

Theorem B.1.1 (Mason and Handscomb [93, Theorem 5.14])**.** *If ϕ has $k + 1$ continuous derivatives, then $\|\phi - \phi_M\|_\infty = O(M^{-k})$.*

Theorem B.1.2 (Mason and Handscomb [93, Theorem 5.16])**.** *If ϕ can be extended to an analytic function on $E_r := \{z \in \mathbb{C} : |z + \sqrt{z^2 - 1}| = r\}$ for $r > 1$ and $C := \sup_{z \in E_r} |\phi(z)|$, then*

$$\|\phi - \phi_M\|_\infty \leq \frac{C}{r - 1} r^{-M}.$$

Chebyshev polynomials also provide a uniformly good approximation of the derivative of the function they are used to approximate.

Theorem B.1.3. *If ϕ can be extended to an analytic function on E_r for $r > 1$ and $C := \sup_{z \in E_r} |\phi(z)|$, then*

$$\|\phi' - \phi'_M\|_\infty \leq C r^{-M} \frac{r + 1}{(r - 1)^4} \big[M^2 r(r + 1) + M(2r^2 + r + 1) + r(r + 1)\big] =: B(C, r, M)$$

Proof The proof follows the same structure as that for Theorem 5.16 in Mason and Handscomb [93]. For Chebyshev polynomials, $\varsigma(\mathrm{d}s) = \frac{2}{\pi}(1 - s^2)^{-1/2}\mathrm{d}s$. Note that $\phi(s) = \sum_{m=0}^\infty (\int \phi \psi_m \mathrm{d}\varsigma)\psi_m(s)$ and hence $\phi'(s) = \sum_{m=0}^\infty (\int \phi \psi_m \mathrm{d}\varsigma)\psi'_m(s)$. Since $\psi'_m = m U_{m-1}$, where $\{U_m\}_{m \geq 0}$ are the Chebyshev polynomials of the second kind,

$$\phi'(s) - \phi'_M(s) = \sum_{m=M+1}^\infty \frac{2m}{\pi} \int_{-1}^1 (1 - v^2)^{-1/2} \phi(v) \psi_m(v) U_{m-1}(s) \mathrm{d}v.$$

Define the conformal mappings $s = \frac{1}{2}(\xi+\xi^{-1})$ and $v = \frac{1}{2}(\zeta+\zeta^{-1})$, and $\phi(v) =: \tilde{\phi}(\zeta) = \tilde{\phi}(\zeta^{-1})$. By assumption, $|\tilde{\phi}(\zeta)| \le C$. Let $\mathcal{C}_1$ denote the complex unit circle and for $r \in \mathbb{R}_+$, let $\mathcal{C}_r := r\mathcal{C}_1$. Using the conformal mappings, we have

$$
\begin{aligned}
&\phi'(s) - \phi'_M(s) \\
&= \sum_{m=M+1}^{\infty} \frac{m}{4\mathrm{i}\pi} \oint_{\mathcal{C}_1} \tilde{\phi}(\zeta)(\zeta^m + \zeta^{-m}) \frac{\xi^m - \xi^{-m}}{\xi - \xi^{-1}} \frac{\mathrm{d}\zeta}{\zeta} \\
&= \sum_{m=M+1}^{\infty} \frac{m}{2\mathrm{i}\pi} \oint_{\mathcal{C}_r} \tilde{\phi}(\zeta)\zeta^{-m} \frac{\xi^m - \xi^{-m}}{\xi - \xi^{-1}} \frac{\mathrm{d}\zeta}{\zeta} \\
&= \frac{1}{2\mathrm{i}\pi} \oint_{\mathcal{C}_r} \frac{\tilde{\phi}(\zeta)}{\xi - \xi^{-1}} \left(\frac{\xi^{M+1}\zeta^{-M-1}(1 + M + \xi\zeta^{-1})}{(\xi\zeta^{-1} - 1)^2} - \frac{\xi^{-M-1}\zeta^{-M-1}(1 + M + \xi^{-1}\zeta^{-1})}{(\zeta^{-1}\xi^{-1} - 1)^2} \right) \frac{\mathrm{d}\zeta}{\zeta} \\
&\le \frac{C}{2\mathrm{i}\pi} \oint_{\mathcal{C}_r} \frac{\xi\zeta^{-M-1}\xi^{-M-1}}{\xi^2 - 1} \left(\frac{\xi^{2M+2}(1 + M + \xi\zeta^{-1})}{(\xi\zeta^{-1} - 1)^2} - \frac{(1 + M + \xi^{-1}\zeta^{-1})}{(\zeta^{-1}\xi^{-1} - 1)^2} \right) \frac{\mathrm{d}\zeta}{\zeta}.
\end{aligned}
$$

Letting $\eta := \xi^2$ and $\psi := \xi^{-1}\zeta^{-1}$, the absolute value of the integrand is

$$
\begin{aligned}
&\frac{|\psi|^{M+1}}{|\eta - 1|} \left| \frac{\eta^{M+1}(1 + M - \eta\psi)}{(\eta\psi - 1)^2} - \frac{1 + M - \psi}{(\psi - 1)^2} \right| \\
&= r^{-M-1} \frac{|\eta\psi - 1|^{-2}|\psi - 1|^{-2}}{|\eta - 1|} \left| \eta^{M+1}(1 + M - \eta\psi)(\psi - 1)^2 - (1 + M - \psi)(\eta\psi - 1)^2 \right| \\
&\le r^{-M-1} \frac{(r^{-1} - 1)^{-4}}{|\eta - 1|} \Big[|\psi||\eta^{M+2} - 1| + (M + 1)|\eta^{M+1} - 1| + 2|\psi|^2|\eta^{M+1} - 1| \\
&\qquad\qquad + 2(M + 1)|\psi||\eta^M - 1| + |\psi|^3|\eta^M - 1| + (M + 1)|\phi|^2|\eta^{M-1} - 1| \Big] \\
&\le \frac{r^{-M+3}}{(r - 1)^4} \left[\frac{M + 2}{r} + (M + 1)^2 + \frac{2(M + 1)}{r^2} + \frac{2M(M + 1)}{r} + \frac{M}{r^3} + \frac{M^2 - 1}{r^2} \right] \\
&= r^{-M} \frac{r + 1}{(r - 1)^4} \big[M^2 r(r + 1) + M(2r^2 + r + 1) + r(r + 1) \big].
\end{aligned}
$$

The final inequality follows from the fact that for $k \in \mathbb{N}$,

$$
|\eta^k - 1|/|\eta - 1| = |\sin(k \arg(\eta))/\sin(\arg(\eta)| \le k.
$$

The result now follows. □

Since ϕ_{logit} is smooth, we can apply Theorems B.1.2 and B.1.3 to obtain exponential convergence rates of the (derivative of the) Chebyshev approximation. The same is true in the Poisson and smoothed Huber regression cases.

Corollary B.1.4. *Fix $R > 0$. If $\phi(s) = \log(1 + e^{-Rs})$, $s \in [-1, 1]$, then for any $r \in (1, \pi/R + \sqrt{\pi^2/R^2 + 1})$,*

$$
\|\phi - \phi_M\|_\infty \le \frac{C(r, R)}{(r - 1)r^M} \qquad \textit{and} \qquad \|\phi' - \phi'_M\|_\infty \le B(C(r, R), r, M),
$$

where $C(r,R) := \left|\log\left(1+e^{-\frac{1}{2}R(r-r^{-1})i}\right)\right|$.

Proof The function e^{-Rs} is entire while log is analytic except at 0. Thus, we must determine the minimum value of r such that there exists $z \in E_r$ such that $1+e^{-Rz}=0$. Taking $z=a+b\mathrm{i}$, it must hold that $b \in \{k\pi/R : k \in \mathbb{Z}\}$ since otherwise e^{-Rz} would contain an imaginary component. If $b = 2k\pi/R$ then $e^{-Rz} = e^{-Ra} > 0$, so this cannot be a solution to $1+e^{-Rz}=0$. However, taking $b=(2k+1)\pi/R$ yields $1-e^{-Ra}=0 \implies a=0$. Hence, $z=(2k+1)\pi\mathrm{i}/R$ and thus

$$\begin{aligned}
|z+\sqrt{z^2-1}| &= |\pi\mathrm{i}/R + \sqrt{-(2k+1)^2\pi^2/R^2-1}| \\
&= |(\pi/R + \sqrt{(2k+1)^2\pi^2/R^2+1})\mathrm{i}| \\
&= \pi/R + \sqrt{(2k+1)^2\pi^2/R^2+1} \\
&\geq \pi/R + \sqrt{\pi^2/R^2+1}.
\end{aligned}$$

Thus we must choose $r < \pi/R + \sqrt{\pi^2/R^2+1}$. For any such r, $|\phi(z)|$ is maximized along E_r when $z=b\mathrm{i}$, which implies $b=\frac{1}{2}(r-r^{-1})$ and hence $C=C(r,R)$. The two inequalities now follow from, respectively, Theorems B.1.2 and B.1.3. □

Corollary B.1.5. *Fix* $R>0$. *If* $\phi(s)=e^{Rs}$, $s\in[-1,1]$, *then for any* $r>1$,

$$\begin{aligned}
\|\phi-\phi_M\|_\infty &\leq \frac{e^{\frac{1}{2}R(r+r^{-1})}}{(r-1)r^M} \\
\|\phi'-\phi'_M\|_\infty &\leq B(e^{\frac{1}{2}R(r+r^{-1})}, r, M).
\end{aligned}$$

Proof The proof is similar to that for Corollary B.1.4. The differences are as follows. The function e^{-Rs} is entire, so we may choose any $r>1$. For any such r, $|\phi(z)|$ is maximized along E_r when z is real, which implies $z=\frac{1}{2}(r+r^{-1})$ and hence $C=e^{\frac{1}{2}R(r+r^{-1})}$. □

Corollary B.1.6. *Fix* $R>0$. *If* $\phi(s)=b^2\left(\sqrt{1+\frac{R^2s^2}{b^2}}-1\right)$, $s\in[-1,1]$, *then for any* $r\in(1, b/R+\sqrt{b^2/R^2+1})$,

$$\begin{aligned}
\|\phi-\phi_M\|_\infty &\leq \frac{b^2\sqrt{1+\{(r^2+1)/(2rb)\}^2}-b^2}{r-1}r^{-M} \\
\|\phi'-\phi'_M\|_\infty &\leq B\left(b^2\sqrt{1+\{(r^2+1)/(2rb)\}^2}-b^2, r, M\right).
\end{aligned}$$

Proof The proof is similar to that for Corollary B.1.4. The differences are as follows. The square root function is analytic except at zero, so we must determine the minimum value of r such that there exists $z \in E_r$ such that $1+R^2z^2/b^2=0$.

Solving, we find that $z = \mathrm{i}b/R$. Thus, we have

$$|z + \sqrt{z^2 - 1}| = b/R + \sqrt{b^2/R^2 + 1}$$

and so must choose $1 < r < b/R + \sqrt{b^2/R^2 + 1}$. For any such r, $|\phi(z)|$ is maximized along E_r when z is real, which implies $z = \frac{r^2+1}{2r}$ and hence $C = b^2\left(\sqrt{1 + \left(\frac{r^2+1}{2rb}\right)^2} - 1\right)$. □

B.2 PASS-GLM Theorems and Proofs

Theorem B.2.1. *Let $\mathbb{B}_r(\boldsymbol{\theta}^*) := \{\boldsymbol{\theta} \in \Theta \mid \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq r\}$. Assume there exist parameters ε_N and ϱ_N such that for all $\boldsymbol{\theta} \in \mathbb{B}_{r_N}(\boldsymbol{\theta}_{MAP})$, where $r_N^2 := 4\varepsilon_N/\varrho_N$,*

(A) $|\mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) - \tilde{\mathcal{L}}_\mathcal{D}(\boldsymbol{\theta})| \leq \varepsilon_N$ and *(B) $-\log \pi_\mathcal{D}$ is ϱ_N-strongly convex.*[1]

Furthermore, assume that for all $\boldsymbol{\theta} \in \Theta$,

(C) $\log \pi_\mathcal{D}$ is strictly quasi-concave[2] *and* *(D) $\tilde{\mathcal{L}}_\mathcal{D}(\boldsymbol{\theta}) \leq \mathcal{L}_\mathcal{D}(\boldsymbol{\theta}) + \varepsilon_N$.*

Then $\|\boldsymbol{\theta}_{MAP} - \tilde{\boldsymbol{\theta}}_{MAP}\|_2^2 \leq \frac{4\varepsilon_N}{\varrho_N}$.

Remark (Assumptions). The error in the MAP estimate naturally depends on the error of the approximate log-likelihood (Assumption (A)) as well as the flatness of the posterior (Assumption (B)). In the latter case, if $\log \pi_\mathcal{D}$ is very flat, then even a small error from using $\tilde{\mathcal{L}}_\mathcal{D}$ in place of $\mathcal{L}_\mathcal{D}$ could lead to a large error in the approximate MAP solution. However, the stronger assumptions, (A) and (B), need hold only near the MAP solution.

Remark (Strict quasi-concavity). Requiring that $\log \pi_\mathcal{D}$ be only strictly quasi-concave (rather than strongly log-concave everywhere) substantially increases the applicability of the result. For instance, it allows heavy-tailed priors (e.g., Cauchy) as well as sparsity-inducing priors (e.g., Laplace/L_1 regularization).

Proof [Proof of Theorem B.2.1] An equivalent condition for f to be strictly quasi-convex is that if $f(\mathbf{v}) > f(\mathbf{w})$ then $\langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle > 0$ [126, Theorem 21.14]. We obtain the result by considering some $\boldsymbol{\theta}$ such that $\boldsymbol{\theta} \notin \mathbb{B}_{r_N}(\boldsymbol{\theta}_{\mathrm{MAP}})$. Since $\varpi := \log \pi_\mathcal{D}$ is strictly quasi-concave (by Assumption (C)), if it has a global maximum it is unique (if it had two global maxima, this would immediately yield a contradiction). By hypothesis $\boldsymbol{\theta}_{\mathrm{MAP}}$ is such a global maximum. Thus, $\varpi(\boldsymbol{\theta}_{\mathrm{MAP}}) > \varpi(\boldsymbol{\theta})$, which implies

$$\langle \nabla \varpi(\boldsymbol{\theta}), \boldsymbol{\theta}_{\mathrm{MAP}} - \boldsymbol{\theta} \rangle > 0. \tag{B.2.1}$$

[1] A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is ϱ-strongly convex if for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, $f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + (\varrho/2)\|\mathbf{v} - \mathbf{w}\|_2^2$.

[2] An arbitrary function $g : \mathbb{R}^d \to \mathbb{R}$ is strictly quasi-concave if for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, $\mathbf{v} \neq \mathbf{w}$, and $t \in (0, 1)$, $g(t\mathbf{v} + (1-t)\mathbf{w}) > \min\{g(\mathbf{v}), g(\mathbf{w})\}$.

Now, fix $\boldsymbol{\theta}'$ such that $\boldsymbol{\theta}' \notin \mathbb{B}_{r_N}(\boldsymbol{\theta}_{\mathrm{MAP}})$. Let $r'_N := \|\boldsymbol{\theta}' - \boldsymbol{\theta}_{\mathrm{MAP}}\|_2 > r_N$ and $\boldsymbol{\theta}'' := \frac{r_N}{r'_N}\boldsymbol{\theta}' + \frac{r'_N - r_N}{r'_N}\boldsymbol{\theta}_{\mathrm{MAP}}$, the projection of $\boldsymbol{\theta}'$ onto $\mathbb{B}_{r_N}(\boldsymbol{\theta}_{\mathrm{MAP}})$. Applying the fundamental theorem of calculus for line integrals on the linear path $\gamma[\boldsymbol{\theta}', \boldsymbol{\theta}'']$ from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}''$, parameterized as $\boldsymbol{\theta}(t) = t\boldsymbol{\theta}'' + (1-t)\boldsymbol{\theta}'$, we have

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}'') - \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}') &= \int_{\gamma[\boldsymbol{\theta}', \boldsymbol{\theta}'']} \nabla \varpi(\boldsymbol{\theta}) \cdot \mathrm{d}\boldsymbol{\theta} \\
&= \int_0^1 \nabla \varpi(\boldsymbol{\theta}(t)) \cdot (\boldsymbol{\theta}'' - \boldsymbol{\theta}') \, \mathrm{d}t \\
&= \frac{r'_N - r_N}{r'_N} \int_0^1 \nabla \varpi(\boldsymbol{\theta}(t)) \cdot (\boldsymbol{\theta}_{\mathrm{MAP}} - \boldsymbol{\theta}') \, \mathrm{d}t \\
&= \frac{r'_N - r_N}{r'_N} \int_0^1 C(t) \nabla \varpi(\boldsymbol{\theta}(t)) \cdot (\boldsymbol{\theta}_{\mathrm{MAP}} - \boldsymbol{\theta}(t)) \mathrm{d}t \\
&> 0,
\end{aligned}$$

where $C(t) := \frac{r'_N}{r'_N - t r'_N + t r_N}$ and the inequality follows from Eq. (B.2.1). Hence,

$$\varpi(\boldsymbol{\theta}') < \varpi(\boldsymbol{\theta}'') \tag{B.2.2}$$

and

$$\begin{aligned}
\log \pi_0(\boldsymbol{\theta}') + \tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}') &\leq \log \pi_0(\boldsymbol{\theta}') + \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}') + \varepsilon_N && \text{by Assumption (D)} \\
&< \log \pi_0(\boldsymbol{\theta}'') + \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}'') + \varepsilon_N && \text{by Eq. (B.2.2)} \\
&\leq \log \pi_0(\boldsymbol{\theta}_{\mathrm{MAP}}) + \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{\mathrm{MAP}}) + \varepsilon_N - \frac{\varrho_N r_N^2}{2} && \text{by Assumption (B)} \\
&= \log \pi_0(\boldsymbol{\theta}_{\mathrm{MAP}}) + \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{\mathrm{MAP}}) - \varepsilon_N && \text{by definition of } r_n \\
&\leq \log \pi_0(\boldsymbol{\theta}_{\mathrm{MAP}}) + \tilde{\mathcal{L}}_{\mathcal{D}}(\boldsymbol{\theta}_{\mathrm{MAP}}) && \text{by Assumption (A).}
\end{aligned}$$

So $\boldsymbol{\theta}'$ is not a global optimum of $\log \tilde{\pi}_{\mathcal{D}}$ and hence $\tilde{\boldsymbol{\theta}}_{\mathrm{MAP}} \in \mathbb{B}_{R_N}(\boldsymbol{\theta}_{\mathrm{MAP}})$. □

We present a generalization of Corollary 3.2.1. Let $\|\mathbf{T}\|_{op} := \sup_{\substack{\mathbf{v} \in \mathbb{R}^d \\ \|\mathbf{v}\|_2 = 1}} \|\mathbf{T}[\mathbf{v}]\|_{op}$ denote the operator norm of the tensor $\mathbf{T}$ (with $\|\mathbf{T}\|_{op} = \|\mathbf{T}\|_2$ if $\mathbf{T}$ is a matrix). Recall the Lipschitz operator bound property

$$\|\nabla h(x)\|_{op} = \sup_{y \neq x} \frac{\|h(x) - h(y)\|_{op}}{\|x - y\|_2}, \tag{B.2.3}$$

which holds for any sufficiently smooth $h : \mathbb{R}^d \to (\mathbb{R}^d)^{\otimes k}$. Recall also that for compatible operators T and T', $\|TT'\|_{op} \leq \|T\|_{op}\|T'\|_{op}$.

Corollary B.2.2. *Assume the tensor defined by* $T_{ijk} := \sum_{n=1}^N x_{ni} x_{nj} x_{nk}$ *satisfies*

$$\|\mathbf{T}\|_{op} \leq LN/d^2.$$

For the logistic regression model, assume that $\|\nabla^2 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{MAP})^{-1}\|_2 \leq cd/N$ *and that* $\|\mathbf{x}_n\|_2 \leq 1$ *for all* $n = 1, \ldots, N$*. Let* ϕ_M *be the order* M *Chebyshev approximation to* ϕ_{logit} *on* $[-R, R]$ *such that Eq.* (3.1) *holds. Let* $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ *denote the posterior approximation obtained by using* ϕ_M *with a strictly quasi-log concave prior. Let*

$$\varepsilon := \min_{r \in (1, \pi/R + \sqrt{\pi^2/R^2+1})} \left| \log\left(1 + e^{-\frac{1}{2}R(r - r^{-1})i}\right) \right| (r-1)^{-1} r^{-M}$$

and $\alpha^* := 1 + b - \sqrt{(b+1)^2 - 1}$*, where* $b := \frac{\varepsilon L^2 c^3}{54d}$*. If* $R - \|\boldsymbol{\theta}_{MAP}\|_2 \geq 2\sqrt{\frac{cd\varepsilon}{\alpha^*}}$*, then*

$$\|\boldsymbol{\theta}_{MAP} - \tilde{\boldsymbol{\theta}}_{MAP}\|_2^2 \leq \frac{4cd\varepsilon}{\alpha^*} \leq \frac{4}{27} c^4 L^2 \varepsilon^2 + 8cd\varepsilon$$

and Corollary 3.2.1 follows from the upper bound $\|\mathbf{T}\|_{op} \leq N$ *(using the assumption that* $\|\mathbf{x}_n\|_2 \leq 1$*).*

Proof By Corollary B.1.4, for all $s \in [-R, R]$, $|\phi_{\mathsf{logit}}(s) - \phi_M(s)| \leq \varepsilon N$. It is easy to verify that $\max_{s \in \mathbb{R}} |\phi'''_{\mathsf{logit}}(s)| = \frac{1}{6\sqrt{3}}$ and therefore $\|\nabla^3 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{op} \leq \frac{1}{6\sqrt{3}} \|\mathbf{T}\|_{op} \leq \frac{LN}{6\sqrt{3}d^2}$. Since by hypothesis $\|(\nabla^2 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{\mathrm{MAP}}))^{-1}\|_2 \leq cd/N$, $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}_{\mathrm{MAP}})$ is $N/(cd)$-strongly concave. We can write $\nabla(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1} = -(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1} \nabla^3 \mathcal{L}_{\mathcal{D}} (\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}$ if we treat the first $(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}$ as a matrix to matrix operator, $\nabla^3 \mathcal{L}_{\mathcal{D}}$ as a vector to matrix operator, and the second $(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}$ as a vector to vector operator. Thus

$$\|\nabla(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta})\|_{op} \leq \|(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta})\|_{op}^2 \|\nabla^3 \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})\|_{op} \leq \frac{c^2 d^2}{N^2} \frac{LN}{6\sqrt{3}d^2} = \frac{c^2 L}{6\sqrt{3}N}.$$

Using the triangle inequality and Eq. (B.2.3), we have

$$\begin{aligned}
\|(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta})\|_{op} &\leq \|(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}})\|_{op} + \|(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta}) - (\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}})\|_{op} \\
&\leq \|(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}})\|_{op} + \|\nabla(\nabla^2 \mathcal{L}_{\mathcal{D}})^{-1}(\boldsymbol{\theta})\|_{op} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}}\|_2 \\
&\leq \frac{cd}{N} + \frac{c^2 L}{6\sqrt{3}N} \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{MAP}}\|_2,
\end{aligned}$$

so $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ is $\alpha N/(cd)$-strongly concave for all $\boldsymbol{\theta} \in \mathbb{B}_{\Delta}(\boldsymbol{\theta}_{\mathrm{MAP}})$ if

$$\frac{cd}{N} + \frac{c^2 L \Delta}{6\sqrt{3}N} \leq \frac{cd}{N\alpha} \quad \Leftrightarrow \quad \Delta^2 \leq \frac{108 d^2 (1-\alpha)^2}{L^2 c^2 \alpha^2}.$$

To apply Theorem B.2.1, we require that $\Delta^2 \geq 4\varepsilon cd/\alpha$. Combining the two inequalities, we have

$$\frac{4\varepsilon cd}{\alpha} \leq \frac{108 d^2 (1-\alpha)^2}{L^2 c^2 \alpha^2} \quad \Leftrightarrow \quad \frac{\varepsilon c^3 L^2}{27d} \alpha \leq (1-\alpha)^2 \quad \Leftrightarrow \quad 0 \leq \alpha^2 - (2+b)\alpha + 1.$$

Solving the quadratic implies that the maximal viable α value is $\alpha^* = 1 + b - \sqrt{(b+1)^2 - 1} \geq \frac{1}{2(b+1)}$.

Requiring $R - \|\boldsymbol{\theta}_{\mathrm{MAP}}\|_2 \geq 2\sqrt{\frac{cd\varepsilon}{\alpha^*}}$ together with the hypothesis that $\|\mathbf{x}_n\| \leq 1$ ensures that we are considering only inner products $\mathbf{x}_n \cdot \boldsymbol{\theta} \in [-R, R]$. Since Eq. (3.1) holds by hypothesis, Assumption (D) holds. The result now follows from Theorem B.2.1. □

Proof [Proof sketch of Corollary 3.2.3] The proof is similar in spirit to Corollary B.2.2. The key differences are that we apply Corollary B.1.6 and use the condition that a constant fraction of the data satisfies $|\mathbf{x}_n \cdot \boldsymbol{\theta}_{\mathrm{MAP}} - y_n| \leq b/2$ to guarantee $\Theta(N)$-strong log-convexity of $-\log \pi_{\mathcal{D}}$ near the MAP. □

Recall that a centered random variable X is said to be σ^2-*subgaussian* [23, Section 2.3] if for all $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sX}] \leq e^{s^2\sigma^2/2}.$$

Theorem B.2.3. *Assume that*

(E) $-\log \tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ *is* $\tilde{\varrho}$*-strongly convex,*

(F) for all $n = 1, \ldots, N$, $\|\mathbf{x}_n\|_2 \leq 1$,

(G) there exist constants $a_n, b, R, \alpha \in \mathbb{R}_+$ *such that*

$$\|\nabla_{\boldsymbol{\theta}} \phi(\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle) - \nabla_{\boldsymbol{\theta}} \phi_M(\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle)\|_2 \leq a_n + b \max(0, |\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle| - R), \text{ and}$$

(H) $-\log \pi_{\mathcal{D}}(\boldsymbol{\theta})$ *is* ϱ*-strongly convex with mean* $\bar{\boldsymbol{\theta}}$*.*

Let σ_1, σ_2 *be the subgaussianity constants of, respectively, the random variables*

$$\langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}}\rangle - \delta_1 \quad \text{and} \quad \|y_n \mathbf{x}_n\|_2^2 - \delta_2,$$

where the randomness is over $n \sim \mathsf{Unif}\{1, \ldots, N\}$*. Let* $\delta_1 := \mathbb{E}[\langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}}\rangle]$, $\delta_2 := \mathbb{E}[\|y_n \mathbf{x}_n\|_2^2]$*, and* $\bar{a} := \sum_{n=1}^N a_n$*. Then there exists an explicit constant* ε *(equal to zero if* $b = 0$ *and depending on* R*,* ϱ*,* σ_1*,* σ_2*,* δ_1*, and* δ_2 *otherwise) such that*

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) \leq \tilde{\varrho}^{-1}(\bar{a} + Nb\varepsilon).$$

Remark (Value of ε). The definition of the constant ε is given in the proof of the theorem.

Remark (Assumptions). Our posterior approximation result primarily depends on the peakedness of the approximate posterior (Assumption (E)) and the error of the approximate gradients (Assumption (G)). If the gradients are poorly approximated then the error can be large while if the (approximate) posterior is flat then even small likelihood errors could lead to large shifts in expected values of the parameters and hence large Wasserstein error.

Remark (Verifying assumptions). In the corollaries we use Theorem B.1.3 to control the gradient error in the case of Chebyshev polynomial approximations, which allows us to satisfy Assumption (G). Whether Assumption (E) holds will depend on the choices of M, ϕ, and π_0. For example, if $M = 2$ and $-\log \pi_0$ is convex, then the assumption holds. This assumption could be relaxed to only assume, e.g., a "bounded concavity" condition along with strong convexity in the tails. See Eberle [46], Gorham et al. [62, Section 4], and Appendix C.1 for full details. It is possible that Assumption (H) could also be weakened. The key is to have some control of the tails of $\pi_{\mathcal{D}}$. Both $\langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}} \rangle$ and $\|y_n \mathbf{x}_n\|_2^2$ are subgaussian since $y_n \mathbf{x}_n$ is bounded.

Proof [Proof of Theorem B.2.3] By Assumption (G), we have that

$$\begin{aligned}
\mathrm{err}(\boldsymbol{\theta}) &:= \|\nabla \log \pi_{\mathcal{D}}(\boldsymbol{\theta}) - \nabla \log \tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})\|_2 \\
&\leq \sum_{n=1}^{N} \|\nabla_{\boldsymbol{\theta}} \phi(\langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle) - \nabla_{\boldsymbol{\theta}} \phi_M(\langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle)\|_2 \\
&\leq \bar{a} + \sum_{n=1}^{N} b \max(0, |\langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle| - R).
\end{aligned}$$

By Lemma B.2.4, the random variable $W := \langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle - \delta_1$ is (λ, β)-subexponential. Hence for $t \geq 0$,

$$\mathbb{P}(W \geq t) \vee \mathbb{P}(W - \delta \leq -t) \leq \bar{p}(t, \lambda, \beta) := e^{-\left(\frac{t^2}{2\lambda^2} \wedge \frac{t}{2\beta}\right)}.$$

We can now bound $\pi_{\mathcal{D}}(\mathrm{err})$:

$$\begin{aligned}
\pi_{\mathcal{D}}(\mathrm{err}) &\leq aN + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\theta} \sim \pi_{\mathcal{D}}}[b \max(0, |\langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle| - R)]. \\
&= aN + bN \mathbb{E}_{n \sim \mathsf{Unif}\{1,\ldots,N\}} \mathbb{E}_{\boldsymbol{\theta} \sim \pi_{\mathcal{D}}}[\max(0, |\langle y_n \mathbf{x}_n, \boldsymbol{\theta} \rangle| - R)) \\
&= aN + bN \mathbb{E}[\max(0, |W + \delta_1| - R))] \\
&= aN + bN \mathbb{E}[(W + \delta_1 + R) \mathbb{1}(W + \delta_1 \leq -R) + (W + \delta_1 - R) \mathbb{1}(W + \delta_1 \geq R)].
\end{aligned} \tag{B.2.4}$$

For the second term in the expectation, we have

$$\begin{aligned}
&\mathbb{E}[(W + \delta_1 - R) \mathbb{1}(W \geq R - \delta_1)] \\
&= \int_{R - \delta_1}^{\infty} (w + \delta_1 - R) p(\mathrm{d}w) \\
&= \int_{R - \delta}^{\infty} \mathbb{P}(W \geq t) \, \mathrm{d}t \\
&\leq 0 \vee (\delta_1 - R) + \int_{0 \vee (R - \delta_1)}^{\infty} \bar{p}(t, \lambda, \beta) \mathrm{d}t =: B(R, \delta_1, \lambda, \beta),
\end{aligned}$$

By symmetry, the first term in the expectation in Eq. (B.2.4) is bounded by $B(R, -\delta_1, \lambda, \beta)$,

so

$$\pi_{\mathcal{D}}(\mathrm{err}) \leq \bar{a} + Nb(B(R, \delta_1, \lambda, \beta) + B(R, -\delta_1, \lambda, \beta)).$$

Assumption (E) implies that $\tilde{\pi}_{\mathcal{D}}$ satisfies Assumption 2.A of Huggins and Zou [73] with $C = 1$ and $\rho = e^{-\tilde{\varrho}}$. By Theorem 2 of Gorham et al. [62], it is not necessary for the Lipschitz conditions in Assumption 2.A of Huggins and Zou [73] to hold. Furthermore, it can easily be seen that 2.B(3) of Huggins and Zou [73] is not necessary if both $\pi_{\mathcal{D}}$ and $\tilde{\pi}_{\mathcal{D}}$ are strongly convex. The remaining portions of Assumption 2.B of Huggins and Zou [73] are satisfied, however. Thus we can apply Theorem 3.4 from Huggins and Zou [73], which yields

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) \leq \tilde{\varrho}^{-1}\pi_{\mathcal{D}}(\mathrm{err}) \leq \tilde{\varrho}^{-1}(\bar{a} + Nb\varepsilon),$$

where $\varepsilon := B(R, \delta_1, \lambda, \beta) + B(R, -\delta_1, \lambda, \beta)$. □

Lemma B.2.4. *Under the conditions of Theorem B.2.3, the random variable* $\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle - \delta_1$ *is* (λ, β)*-subexponential, where* $\lambda^2 := 4\left(\frac{1+\delta_2}{\varrho} \vee \sigma_1^2\right)$ *and* $\beta^2 := \frac{2\sigma_2^2}{\varrho}$.

Proof Let $\mathbf{z}_n = y_n \mathbf{x}_n$. For $|s| \leq 1/\beta$, we have

$$\begin{aligned}
\mathbb{E}[e^{s(\langle \mathbf{z}_n, \boldsymbol{\theta}\rangle - \delta_1)}] &= \mathbb{E}[\mathbb{E}[e^{s\langle \mathbf{z}, \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\rangle} \mid \mathbf{z}_n = \mathbf{z}] e^{s(\langle \bar{\boldsymbol{\theta}}, \mathbf{z}_n\rangle - \delta_1)}] \\
&\leq \mathbb{E}[e^{s^2 \|\mathbf{z}_n\|_2^2/\varrho'} e^{s(\langle \bar{\boldsymbol{\theta}}, \mathbf{z}_n\rangle - \delta_1)}] && \text{Assumption (H)} \\
&\leq 0.5\mathbb{E}[e^{2s^2 \|\mathbf{z}_n\|_2^2/\varrho'} + e^{2s(\langle \bar{\boldsymbol{\theta}}, \mathbf{z}_n\rangle - \delta_1)}] && \text{AM-GM inequality} \\
&\leq 0.5[e^{4s^4\sigma_2^2/\varrho^2 + 2s^2\delta_2/\varrho} + e^{2s^2\sigma_1^2}] && \text{subgaussianity} \\
&\leq 0.5[e^{2s^2(1+\delta_2)/\varrho} + e^{2s^2\sigma_1^2}] && \text{bound on } |s| \\
&\leq e^{s^2\lambda^2/2}.
\end{aligned}$$

□

Corollary B.2.5. *Let* ϕ_2 *be the second-order Chebyshev approximation to* ϕ_{logit} *on* $[-R, R]$ *and let* $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}}_{MAP}, \tilde{\boldsymbol{\Sigma}})$ *denote the posterior approximation obtained by using* ϕ_2 *with a Gaussian prior* $\pi_0(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$. *Let* $\bar{\boldsymbol{\theta}} := \int \boldsymbol{\theta}\pi_{\mathcal{D}}(\mathrm{d}\boldsymbol{\theta})$, *let* $\delta_1 := N^{-1}\sum_{n=1}^N \langle y_n \mathbf{x}_n, \bar{\boldsymbol{\theta}}\rangle$, *and let* σ_1 *be the subgaussanity constant of the random variable* $\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle - \delta_1$, *where* $n \sim \mathsf{Unif}\{1, \ldots, N\}$. *Assume that* $|\delta_1| \leq R$, *that* $\|\tilde{\boldsymbol{\Sigma}}\|_2 \leq cd/N$, *and that* $\|\mathbf{x}_n\|_2 \leq 1$ *for all* $n = 1, \ldots, N$. *Then with* $\sigma_0^2 := \|\boldsymbol{\Sigma}_0\|_2$, *we have*

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) \leq cd\left(a(R) + \sqrt{2}\,\sigma_0 e^{8(2+\sigma_1^2\sigma_0^{-2}) - \sqrt{2}\frac{R-|\delta_1|}{\sigma_0}}\right),$$

where $a(R)$ is bounded by

$$\min_{r\in(1,\pi/R+\sqrt{\pi^2/R^2+1})} \left|\log\left(1+e^{-\frac{1}{2}R(r-r^{-1})i}\right)\right| \frac{(r+1)(9r^2+7r+2)}{r^2(r-1)^4}.$$

Proof Assumption (E) holds by construction. The bound on

$$a(R) := \sup_{s\in[-R,R]} |\phi'_{\mathsf{logit}}(s) - \phi'_2(s)|$$

follows immediately from Corollary B.1.4 in the case of $M = 2$. Furthermore, since $\phi'_2(s) = b_{1,1} + b_{1,2}s$, for $|s| > R$, the additional error is at most $|b_{1,2}|(|s| - R)$. In the case of a Chebyshev approximation, it is easy to verify that $|b_{1,2}| \le 0.25$ for all R (since as $R \to 0$, $b_{1,2} \to \phi''_{\mathsf{logit}}(0) = -0.25$ and $-b_{1,2}$ is a decreasing function of R). In short, $|\phi'_{\mathsf{logit}}(s) - \phi'_2(s)| \le a(R) + 0.25\max(0, |s| - R)$ and therefore, using Assumption (F), we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}}\phi(\langle y_n\mathbf{x}_n, \boldsymbol{\theta}\rangle) - \nabla_{\boldsymbol{\theta}}\phi_M(\langle y_n\mathbf{x}_n, \boldsymbol{\theta}\rangle)\|_2 \\
&= \|\phi'(\langle y_n\mathbf{x}_n, \boldsymbol{\theta}\rangle)y_n\mathbf{x}_n - \phi'_M(\langle y_n\mathbf{x}_n, \boldsymbol{\theta}\rangle)y_n\mathbf{x}_n\|_2 \\
&\le a(R) + .25\max(0, |\langle y_n\mathbf{x}_n, \boldsymbol{\theta}\rangle| - R).
\end{aligned}$$

Hence Assumption (G) holds with $a_n = a(R)$ and $b = 0.25$.

Now, clearly $-\log\pi_{\mathcal{D}}$ is σ_0^{-2}-strongly convex. Since $\|\mathbf{x}_n\|_2 \le 1$, conclude that $\delta_2 \le 1$ and $\sigma_2 \le 1/2$. To upper bound ε, note that

$$B(R, \delta_1, \lambda, \beta) + B(R, -\delta_1, \lambda, \beta) \le 2B(R, |\delta_1|, \lambda, \beta)$$

and that $\bar{p}(t, \lambda, \beta) \le e^{\frac{\lambda^2}{4\beta^2}}e^{-t/\beta}$. Also, $\lambda^2 \le 4(2\sigma_0^2 + \sigma_1^2)$ and $\beta^2 = \sigma_0^2/2$. Using this upper bound in $B(R, |\delta_1|, \lambda, \beta)$ along with straightforward simplifications yields:

$$2B(R, |\delta_1 a|, \lambda, \beta) \le 2\beta e^{\frac{\lambda^2}{4\beta^2}}e^{-\frac{R-|\delta_1|}{\beta}} \le \sqrt{2}\sigma_0 e^{8(2+\sigma_1^2\sigma_0^{-2})}e^{-\sqrt{2}\frac{R-|\delta_1|}{\sigma_0}}.$$

The result now follows from Theorem B.2.3 since $-\log\tilde{\pi}_{\mathcal{D}}$ is $\|\tilde{\boldsymbol{\Sigma}}\|_2^{-1}$-strongly convex and hence by assumption $N/(cd)$-strongly convex. □

Corollary B.2.6. *Let $f_M(s)$ be the order-M Chebyshev approximation to e^t on the interval $[-R, R]$, and let $\tilde{\pi}_{\mathcal{D}}(\boldsymbol{\theta})$ denote the posterior approximation obtained by using the approximation $\log\tilde{p}(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) := y_n\mathbf{x}_n \cdot \boldsymbol{\theta} - f_M(\mathbf{x}_n \cdot \boldsymbol{\theta}) - \log y_n!$ with a log-concave prior on $\Theta = \mathbb{B}_R(\mathbf{0})$. If $\inf_{s\in[-R,R]} f''_M(s) \ge \tilde{\varrho} > 0$ and $\|\mathbf{x}_n\|_2 \le 1$ for all $n = 1, \ldots, N$, then with $\tau := \|\sum_{n=1}^N \mathbf{x}_n\mathbf{x}_n^\top\|_2$, we have*

$$d_{\mathcal{W}}(\pi_{\mathcal{D}}, \tilde{\pi}_{\mathcal{D}}) \le \frac{N}{\tilde{\varrho}\tau}\min_{r>1} e^{\frac{1}{2}R(r+r^{-1})}\frac{(r+1)[M^2r(r+1) + M(2r^2+r+1) + r(r+1)]}{r^M(r-1)^4}.$$

Note that $\inf_{s\in[-R,R]} f''_M(s) \ge \tilde{\varrho} > 0$ holds as long as M is even and sufficiently large.

Proof Since by hypothesis $\inf_{s\in[-R,R]} f_M''(s) \geq \tilde{\varrho} > 0$, the prior is log-concave, and $-\log\tilde{\pi}_{\mathcal{D}}$ is $\tilde{\varrho}\tau$-strongly convex (i.e., Assumption (E) holds). Using Assumption (F), we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\theta}} \log p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \log \tilde{p}(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})\|_2 \\
&= \|e^{\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle)} y_n \mathbf{x}_n - f_M'(\langle y_n \mathbf{x}_n, \boldsymbol{\theta}\rangle) y_n \mathbf{x}_n\|_2 \\
&\leq \sup_{s\in[-R,R]} |e^{-s} - f_M'(s)| =: a(R).
\end{aligned}$$

which is bounded according to Corollary B.1.5. Hence Assumption (G) holds with $a_n = a(R)$ and $b = 0$. The result now follows immediately from Theorem B.2.3. □

Appendix C

Chapter 4 Proofs

C.1 Exponential contractivity

A natural generalization of the strong concavity case is to assume that $\log \pi$ is strongly concave for x and x' far apart and that $\log \pi$ has "bounded convexity" when x and x' are close together. It turns out that in such cases Assumption 4.A still holds. More formally, the following assumption can be used even when the drift is not a gradient. For $f : \mathcal{X} \to \mathbb{R}^d$ and $r > 0$, let

$$\kappa(r) := \inf\left\{-2\frac{(f(x) - f(x')) \cdot (x - x')}{r^2} : x, x' \in \mathcal{X}, \|x - x'\|_2 = r\right\}.$$

Define the constant $R_0 = \inf\{R \geq 0 : \kappa(r) \geq 0 \ \forall r \geq R\}$.

Assumption A.9 (Strongly log-concave tails)**.** *For the function $f \in C^1(\mathcal{X}, \mathbb{R}^d)$, there exist constants $R, \ell \in [0, \infty)$ and $k \in (0, \infty)$ such that*

$$\kappa(r) \geq -\ell \text{ for all } r \leq R \text{ and } \kappa(r) \geq k \text{ for all } r > R.$$

Furthermore, $\kappa(r)$ is continuous and $\int_0^1 r\kappa(r)^- \mathrm{d}r < \infty$.

Theorem C.1.1 (Eberle [46], Wang [140])**.** *If Assumption A.9 holds for $f = b$ then Assumption 4.A holds for*

$$C = \exp\left(\frac{1}{4}\int_0^{R_0} r\kappa(r)\mathrm{d}s\right)$$

$$\frac{1}{\log(1/\rho)} \leq \begin{cases} \frac{3e}{2}\max(R^2, 8k^{-1}) & \text{if } \ell R_0^2 \leq 8 \\ 8\sqrt{2\pi}R^{-1}\ell^{-1/2}(\ell^{-1} + k^{-1})e^{\ell R^2/8} + 32R^{-2}k^{-2} & \text{otherwise.} \end{cases}$$

For detailed calculations for the case of a mixture of Gaussians model, see Gorham et al. [62].

C.2 Proofs of the main results in Section 4.2

We state all our results in the more general case of a diffusion on a convex space $\mathcal{X} \subseteq \mathbb{R}^d$. We begin with some additional definitions. Any set $\mathcal{G} \subseteq C(\mathcal{X})$ defines an *integral probability metric* (IPM)

$$d_{\mathcal{G}}(\mu, \nu) = \sup_{\phi \in \mathcal{G}} |\mu(\phi) - \nu(\phi)|,$$

where μ and ν are measures on $\mathcal{X}$. The *Wasserstein metric* $d_{\mathcal{W}}$ corresponds to $\mathcal{W} := \{\phi \in C(\mathcal{X}) \,|\, \|\phi\|_L \leq 1\}$. The set $\mathcal{H} := \{\phi \in C^1(\mathcal{X}) \,|\, \|h\|_L \leq 1\}$ will be used to define an IPM $d_{\mathcal{H}}$. For a set $\mathcal{Z} \subseteq \mathbb{R}^n$, we use $\partial\mathcal{Z}$ to denote the boundary of $\mathcal{Z}$.

Suppose $\|b - \tilde{b}\|_2 \leq \epsilon$. We first state several standard properties of the Wasserstein metric and invariant measures of diffusions. The proofs are included here for completeness.

Lemma C.2.1. *For any $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $d_{\mathcal{H}}(\mu, \nu) = d_{\mathcal{W}}(\mu, \nu)$.*

Proof [Proof sketch] The result follows since any Lipschitz function is continuous and a.e.-differentiable, and continuously differentiable functions are dense in the class of continuous and a.e.-differentiable functions. □

We use the notation $(X_t)_{t\geq 0} \sim \mathsf{Diff}(b, \Sigma)$ if X_t is a diffusion defined by

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \Sigma \mathrm{d}W_t - n_t L(\mathrm{d}t).$$

A diffusion X_t is said to be *strong Feller* if its semigroup operator $(\pi_t\phi)(x) := \mathbb{E}[\phi(X_{x,t})]$, $\phi \in C(\mathcal{X})$, satisfies the property that for all bounded ϕ, $\pi_t\phi$ is bounded and continuous.

Proposition C.2.2. *Assume Assumption 4.B(1) holds and let $(X_t)_{t\geq 0} \sim \mathsf{Diff}(b, I)$. Then for each $x \in \mathcal{X}$, $X_{x,t}$ has the invariant density π and is strong Feller.*

Proof The existence of the diffusions follows from Tanaka [134, Theorem 4.1], the strong Feller property follows from Ethier and Kurtz [48, Ch. 8, Theorems 1.5 & 1.6], and the fact that π is the unique stationary measure follows since $\mathcal{A}_b^*\pi = 0$. □

By the same proof as Proposition C.2.2, we have

Proposition C.2.3 (Diffusion properties)**.** *For $f \in C^0(\mathcal{X}, \mathbb{R}^d)$ with $\|f\|_L < \infty$, the diffusion $(X_t)_{t\geq 0} \sim \mathsf{Diff}(f, I)$ exists and has an invariant distribution π_f.*

Proposition C.2.4 (Expectation of the generator)**.** *For $f \in C^0(\mathcal{X}, \mathbb{R}^d)$, let the diffusion $(X_t)_{t\geq 0} \sim \mathsf{Diff}(f, I)$ have invariant density π_f and assume that linear functions are π_f-integrable. Then for all $\phi \in C^2(\mathcal{X})$ such that $\|\phi\|_L < \infty$ and $\mathcal{A}_f\phi$ is π_f-integrable, $\pi_f(\mathcal{A}_f\phi) = 0$.*

Proof Let P_t be the semigroup operator associated with $(X_t)_{t\geq 0}$:

$$(P_t\phi)(x) = \mathbb{E}[\phi(X_{x,t})].$$

Since by hypothesis linear functions are π_f-integrable and ϕ is Lipschitz, ϕ is π_f-integrable. Thus, $P_t\phi$ is π_f-integrable and by the definition of an invariant measure (see [12, Definition 1.2.1] and subsequent discussion),

$$\pi_f(P_t\phi) = \pi_f\phi. \tag{C.2.1}$$

Using the fact that $\partial_t P_t = P_t\mathcal{A}_f$ [12, Eq. (1.4.1)], differentiating both size of Eq. (C.2.1), applying dominated convergence, and using the hypothesis that $\mathcal{A}_f\phi$ is π_f-integrable yields

$$0 = \partial_t\pi_f(P_t\phi) = \pi_f(\partial_t P_t\phi) = \pi_f(P_t\mathcal{A}_f\phi) = \pi_f(\mathcal{A}_f\phi).$$

□

We next show that the solution to Eq. (4.7) is Lipschitz continuous with a Lipschitz constant depending on the mixing properties of the diffusion associated with the generator.

Proposition C.2.5 (Differential equation solution properties)**.** *If Assumptions 4.A and 4.B(1) hold, then for any $h \in C^1(\mathcal{X})$ with $\|h\|_L < \infty$, the function*

$$u_h(x) := \int_0^\infty (\pi(h) - \mathbb{E}[h(X_{x,t})])\,\mathrm{d}t$$

exists and satisfies

$$\|u_h\|_L \le \frac{C}{\log(1/\rho)}\|h\|_L \tag{C.2.2}$$

$$(\mathcal{A}_b u_h)(x) = h(x) - \pi(h). \tag{C.2.3}$$

Proof We follow the approach of Mackey and Gorham [88]. By Assumption 4.A and the definition of Wasserstein distance, we have that there is a coupling between $X_{x,t}$ and $X_{x',t}$ such that

$$\mathbb{E}[\|X_{x,t} - X_{x',t}\|_2] \le C\|x - x'\|_2\rho^t.$$

The function u_h is well-defined since for any $x \in \mathcal{X}$,

$$\begin{aligned}
\int_0^\infty |\pi(h) - \mathbb{E}[h(X_{x,t})]|\,\mathrm{d}t &= \int_0^\infty \left|\int_{\mathcal{X}} (\mathbb{E}[h(X_{x',t})] - \mathbb{E}[h(X_{x,t})])\pi(x')\,\mathrm{d}x'\right|\mathrm{d}t \\
&\le \sup_{z\in\mathcal{X}}\|\nabla h(z)\|_2 \int_0^\infty \int_{\mathcal{X}} \mathbb{E}[\|X_{x,t} - X_{x',t}\|_2]\pi(x')\,\mathrm{d}x'\,\mathrm{d}t \\
&= \sup_{z\in\mathcal{X}}\|\nabla h(z)\|_2 \int_0^\infty \int_{\mathcal{X}} \|x - x'\|_2 C\rho^t\pi(x')\,\mathrm{d}x'\,\mathrm{d}t \\
&\le \|h\|_L\,\mathbb{E}_{X\sim\pi}[\|x - X\|_2] \int_0^\infty C\rho^t\,\mathrm{d}t
\end{aligned}$$

$$< \infty,$$

where the first line uses the property that $\pi(h) = \int_{\mathcal{X}} \mathbb{E}[h(X_{x',t})]\pi(x')\mathrm{d}x'$ and the final inequality follows from Assumption 4.B(1) and the assumption that $0 < \rho < 1$. Furthermore, u_h has bounded Lipschitz constant since for any $x, x' \in \mathcal{X}$,

$$\begin{aligned}
|u_h(x) - u_h(x')| &= \left| \int_0^\infty \mathbb{E}[h(X_{x,t}) - h(X_{x',t})] \, \mathrm{d}t \right| \\
&\leq \sup_{z \in \mathcal{X}} \|\nabla h(z)\|_2 \int_0^\infty \mathbb{E}[\|X_{x,t} - X_{x',t}\|_2] \, \mathrm{d}t \\
&\leq \|h\|_L \|x - x'\|_2 \int_0^\infty C\rho^t \, \mathrm{d}t \\
&= \frac{C\|h\|_L}{\log(1/\rho)} \|x - x'\|_2.
\end{aligned}$$

Finally, we show that $(\mathcal{A}_b u_h)(x) = h(x) - \pi(h)$. Recall that for $h \in C(\mathcal{X})$, the semigroup operator is given by $(\pi_t h)(x) = \mathbb{E}[h(X_{x,t})]$. Since $X_{x,t}$ is strong Feller for all $x \in \mathcal{X}$ by Proposition C.2.2, for all $t \geq 0$, its generator satisfies [48, Ch. 1, Proposition 1.5]

$$h - \pi_t h = \mathcal{A}_b \int_0^t (\pi(h) - \pi_s h) \, \mathrm{d}s. \tag{C.2.4}$$

Hence,

$$\begin{aligned}
&|h(x) - \pi(h) - [h(x) - (\pi_t h)(x)]| \\
&= \left| \int_{\mathcal{X}} \mathbb{E}[h(X_{x,t})] - \mathbb{E}[h(X_{x',t})]\pi(x') \, \mathrm{d}x' \right| \\
&\leq \sup_{z \in \mathcal{X}} \|\nabla h(z)\|_2 \int_{\mathcal{X}} \mathbb{E}[\|X_{x',t} - X_{x,t}\|_2]\pi(x') \, \mathrm{d}x' \\
&\leq \|h\|_L \, \mathbb{E}_{X \sim \pi}[\|x - X\|_2] C\rho^t.
\end{aligned}$$

Thus, conclude that the left-hand side of Eq. (C.2.4) converges pointwise to $h(x) - \pi(h)$ as $t \to \infty$. Since $\mathcal{A}_b$ is closed [48, Ch. 1, Proposition 1.6], the right-hand side of Eq. (C.2.4) limits to $\mathcal{A}_b u_h$. Hence, u_h solves Eq. (C.2.3). □

We can now prove the main result bounding the Wasserstein distance between the invariant distributions of the original and perturbed diffusions.

Proof [Proof of Theorem 4.2.1] By Proposition C.2.3 and Assumption 4.B, the hypotheses of Proposition C.2.4 hold for $f = \tilde{b}$. Let $\mathcal{F} := \{u_h \,|\, h \in \mathcal{H}\}$. Then

$$\begin{aligned}
d_{\mathcal{W}}(\pi, \tilde{\pi}) &= \sup_{h \in \mathcal{H}} |\pi(h) - \tilde{\pi}(h)| \quad \text{by definition and Assumption 4.B} \\
&= \sup_{h \in \mathcal{H}} |\pi(\mathcal{A}_b u_h) - \tilde{\pi}(\mathcal{A}_b u_h)| \quad \text{by Eq. (C.2.3)}
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{h\in\mathcal{H}} |\tilde{\pi}(\mathcal{A}_b u_h)| \quad \text{by Proposition C.2.4} \\
&= \sup_{u\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b u)| \quad \text{by definition of } \mathcal{F} \\
&= \sup_{u\in\mathcal{F}} |\tilde{\pi}(\mathcal{A}_b u - \mathcal{A}_{\tilde{b}} u)| \quad \text{by Proposition C.2.4} \\
&= \sup_{u\in\mathcal{F}} |\tilde{\pi}(\nabla u \cdot b - \nabla u \cdot \tilde{b})| \qquad \text{by definition of } \mathcal{A}_b \\
&\leq \sup_{u\in\mathcal{F}} |\tilde{\pi}(\|\nabla u\|_2 \|b - \tilde{b})\|_2)| \\
&\leq \frac{C\epsilon}{\log(1/\rho)} \qquad \text{by Eq. (C.2.2) and } \|b - \tilde{b}\|_2 \leq \epsilon.
\end{aligned}
$$

□

A similar analysis can be used to bound the Wasserstein distance between π and $\tilde{\pi}$ when the approximate drift $\tilde{b}$ is itself stochastic.

Proof [Proof of Theorem 4.2.4] We will need to consider the joint diffusions $Z_t = (X_t, Y_t)$ and $\tilde{Z}_t = (\tilde{X}_t, \tilde{Y}_t)$ on $\mathcal{Z} := \mathcal{X} \times \mathbb{R}^d$, where

$$
\begin{aligned}
\mathrm{d}Z_t &= (b(X_t), b_{aux}(Y_t))\,\mathrm{d}t + (\sqrt{2}\,\mathrm{d}W_t^X, \Sigma\,\mathrm{d}W_t^Y) - n_t L(\mathrm{d}t) \\
\mathrm{d}\tilde{Z}_t &= (\tilde{b}(\tilde{X}_t, \tilde{Y}_t), b_{aux}(\tilde{Y}_t))\,\mathrm{d}t + (\sqrt{2}\,\mathrm{d}\tilde{W}_t^X, \Sigma\,\mathrm{d}\tilde{W}_t^Y) - n_t \tilde{L}(\mathrm{d}t).
\end{aligned}
$$

Notice that X_t and Y_t are independent and the invariant distribution of X_t is π. Let π_Z and $\tilde{\pi}_Z$ be the invariant distributions of Z_t and $\tilde{Z}_t$, respectively. Also note that the generators for Z_t and $\tilde{Z}_t$ are, respectively,

$$
\begin{aligned}
\mathcal{A}_Z \phi(z) &= \nabla\phi \cdot (b(x), b_{aux}(y)) + \Delta\phi_x(z) + \Sigma^\top\Sigma : \nabla^2\phi_y(z) \\
\mathcal{A}_{\tilde{Z}} \phi(z) &= \nabla\phi \cdot (\tilde{b}(x,y), b_{aux}(y)) + \Delta\phi_x(z) + \Sigma^\top\Sigma : \nabla^2\phi_y(z).
\end{aligned}
$$

where ∇^2 is the Hessian operator.

By Proposition C.2.3 and 4.B, the hypotheses of Proposition C.2.4 hold for $f(x,y) = (\tilde{b}(x,y), b_{aux}(y))$. Let $\mathcal{H}_Z := \{h \in C^1(\mathcal{Z}) \,|\, \|h\|_L \leq 1\}$ and $\mathcal{F}_Z := \{u_h \,|\, h \in \mathcal{H}_Z\}$. Also, for $z = (x,y) \in \mathcal{Z}$, let $\mathrm{id}_Y(z) = y$. Then, by reasoning analogous to that in the proof of Theorem 4.2.1,

$$
\begin{aligned}
d_{\mathcal{W}}(\pi, \tilde{\pi}) &\leq d_{\mathcal{W}}(\pi_Z, \tilde{\pi}_Z) \\
&= \sup_{h\in\mathcal{H}_Z} |\pi_Z(h) - \tilde{\pi}_Z(h)| \\
&= \sup_{u\in\mathcal{F}_Z} |\tilde{\pi}_Z(\mathcal{A}_Z u - \mathcal{A}_{\tilde{Z}} u)| \\
&= \sup_{u\in\mathcal{F}_Z} |\tilde{\pi}_Z(\nabla u \cdot b - \nabla u \cdot \tilde{b})| \\
&= \sup_{u\in\mathcal{F}_Z} |\mathbb{E}[\nabla u(\tilde{X}, \tilde{Y}) \cdot \mathbb{E}[b(\tilde{X}) - \tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X}]]| \\
&\leq \sup_{u\in\mathcal{F}_Z} |\mathbb{E}[\|\nabla u(\tilde{X}, \tilde{Y})\|_2 \|\mathbb{E}[b(\tilde{X}) - \tilde{b}(\tilde{X}, \tilde{Y}) \,|\, \tilde{X}]\|_2]|
\end{aligned}
$$

$$\leq \frac{C\,\tilde{\pi}(\epsilon)}{\log(1/\rho)}.$$

□

Proof [Proof of Theorem 4.2.5] The proof is very similar to that of Theorem 4.2.1, the only difference is in the Lipshitz coefficient of the differential equation solution $u_h(x)$ in C.2.5. Using polynomial contractivity, we have

$$\begin{aligned}
|u_h(x) - u_h(x')| &= \left| \int_0^\infty \mathbb{E}[h(X_{x,t}) - h(X_{x',t})]\,\mathrm{d}t \right| \\
&\leq \sup_{z \in \mathcal{X}} \|\nabla h(z)\|_2 \int_0^\infty \mathbb{E}[\|X_{x,t} - X_{x',t}\|_2]\,\mathrm{d}t \\
&\leq \|h\|_L \|x - x'\|_2 \int_0^\infty C(t+\beta)^{-\alpha}\,\mathrm{d}t \\
&= \frac{C\|h\|_L}{(\alpha-1)\beta^{\alpha-1}} \|x - x'\|_2.
\end{aligned}$$

Plugging in this Lipschitz constant, we have

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{(\alpha-1)\beta^{\alpha-1}}.$$

□

C.3 Checking the Integrability Condition

The following result gives checkable conditions under which Assumption 4.B(3) holds. Let $\mathbb{B}_R := \{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\}$.

Proposition C.3.1 (Ensuring integrability)**.** *Assumption 4.B(3) is satisfied if* $b = \nabla \log \pi$, $\tilde{b} = \nabla \log \tilde{\pi}$, $\|b - \tilde{b}\|_2 \leq \epsilon$, *and either*

1. *there exist constants* $R > 0, B > 0, \delta > 0$ *such that for all* $x \in \mathcal{X} \setminus \mathbb{B}_R$, $\|b(x) - \tilde{b}(x)\|_2 \leq B/\|x\|_2^{1+\delta}$; *or*

2. *there exists a constant* $R > 0$ *such that for all* $x \in \mathcal{X} \setminus \mathbb{B}_R$, $x \cdot (b(x) - \tilde{b}(x)) \geq 0$.

Proof For case (1), first we note that since $\int_{\mathcal{X}} (\pi(x) - \tilde{\pi}(x))\,\mathrm{d}x = 0$, by the (generalized) intermediate value theorem, there exists $x^* \in \mathcal{X}$ such that $\pi(x^*) - \tilde{\pi}(x^*) = 0$, and hence $\log \pi(x^*) - \log \tilde{\pi}(x^*) = 0$. Let $p[x^*, x]$ be any path from x^* to x. By the

fundamental theorem of calculus for line integrals,

$$\begin{aligned}
|\log \pi(x) - \log \tilde{\pi}(x)| &= \left| \log \tilde{\pi}(x^*) - \log \pi(x^*) + \int_{\gamma[x^*,x]} (b(r) - \tilde{b}(r)) \cdot \mathrm{d}r \right| \\
&= \left| \int_{\gamma[x^*,x]} (b(r) - \tilde{b}(r)) \cdot r'(t)\, \mathrm{d}t \right| \\
&\leq \int_{\gamma[x^*,x]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t.
\end{aligned}$$

First consider $x \in \mathcal{X} \cap \mathbb{B}_R$. Choosing $p[x^*, x]$ to be the linear path $\gamma[x^*, x]$, we have

$$\begin{aligned}
|\log \pi(x) - \log \tilde{\pi}(x)| &\leq \epsilon \int_{\gamma[x^*,x]} \|r'(t)\|_2 \, \mathrm{d}t \\
&= \epsilon \|x - x^*\|_2 \\
&\leq (R + \ell^*)\epsilon, \qquad \text{(C.3.1)}
\end{aligned}$$

where $\ell^* := \|x^*\|_2$.

Next consider $x \in \mathcal{X} \setminus \mathbb{B}_R$. Let $\ell := \|x\|_2$ and $x' = \frac{R}{\ell} x$. Choose $p[x^*, x]$ to consist of the concatenation of the linear paths $\gamma[x^*, 0]$, $\gamma[0, x']$, and $\gamma[x', 0]$, so

$$\begin{aligned}
&\int_{p[x^*,x]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t \\
&= \int_{\gamma[x^*,0]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t + \int_{\gamma[0,x']} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t \\
&\quad + \int_{\gamma[x',x]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t.
\end{aligned}$$

Now, we bound each term:

$$\begin{aligned}
\int_{\gamma[x^*,0]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t &\leq \ell^* \epsilon \\
\int_{\gamma[0,x']} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t &\leq R\epsilon \\
\int_{\gamma[x',x]} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \, \mathrm{d}t &\leq (\ell - R) B \int_0^1 \frac{1}{(R + (\ell - R)t)^{1+\delta}} \\
&= (\ell - R) B \left[\frac{1}{(\ell - R) R^\delta} - \frac{1}{(\ell - R)\ell^\delta} \right] \\
&\leq \frac{B}{R^\delta}.
\end{aligned}$$

It follows that there exists a constant $\tilde{B} > 0$ such that for all $x \in \mathcal{X}$, $|\log \pi(x) - \log \tilde{\pi}(x)| < \tilde{B}$. Hence $\tilde{B}^{-1}\pi < \tilde{\pi} < \tilde{B}\pi$, so ϕ is π-integrable if and only if it is $\tilde{\pi}$-integrable.

Case (2) requires a slightly more delicate argument. Let x^* and ℓ^* be the same as in case (1). For $x \in \mathcal{X} \cap \mathbb{B}_R$, it follows from Eq. (C.3.1) that

$$\log \pi(x) - \log \tilde{\pi}(x) \geq -(R + \ell^*)\epsilon.$$

For $x \in \mathcal{X} \setminus \mathbb{B}_R$, arguing as above yields

$$\begin{aligned}
\log \pi(x) - \log \tilde{\pi}(x) &= \int_{p[x^*,x]} (b(r) - \tilde{b}(r)) \cdot \mathrm{d}r \\
&\geq -\int_{p[x^*,r']} \|b(r) - \tilde{b}(r)\|_2 \|r'(t)\|_2 \,\mathrm{d}t \\
&\quad + \int_{\gamma[x',x]} (b(r) - \tilde{b}(r)) \cdot r'(t) \,\mathrm{d}t \\
&\geq -(R + \ell^*)\epsilon + \int_{\gamma[x',x]} (b(q(t)x) - \tilde{b}(q(t)x)) \cdot ax \,\mathrm{d}t \\
&\geq -(R + \ell^*)\epsilon,
\end{aligned}$$

where we have used the fact that $r(t) = q(t)x$ for some linear function $q(t)$ with slope $a > 0$. Combining the previous two displays, conclude that for all $x \in \mathcal{X}$, $\tilde{\pi}(x) \leq e^{(R+\ell^*)\epsilon}\pi(x)$, hence Assumption 4.B(3) holds. □

We suspect Proposition C.3.1 continues to hold even when $b \neq \nabla \log \pi$ and $\tilde{b} \neq \nabla \log \tilde{\pi}$. Note that condition (1) always holds if $\mathcal{X}$ is compact, but also holds for unbounded $\mathcal{X}$ as long as the error in the gradients decays sufficiently quickly as $\|x\|_2$ grows large. Given an approximate distribution for which $\|b - \tilde{b}\|_2 \leq \epsilon/2$, it is easy to construct a new distribution that satisfies condition (2):

Proposition C.3.2. *Assume that $\tilde{\pi}$ satisfies $\|b - \tilde{b}\|_2 \leq \epsilon/2$ and let*

$$f_R(x) := -\frac{\epsilon x}{2\|x\|_2}\{(2\|x\|_2/R - 1)\mathbb{1}[R/2 \leq \|x\|_2 < R] + \mathbb{1}[\|x\|_2 \geq R]\}.$$

Then the distribution

$$\tilde{\pi}_R(x) \propto \tilde{\pi}(x)e^{f_R(x)}$$

satisfies condition (2) of Proposition C.3.1.

Proof Let $\tilde{b}_R := \nabla \log \tilde{\pi}_R$. First we verify that $\|b - \tilde{b}_R\|_2 \leq \epsilon$. For $x \in \mathcal{X} \cap \mathbb{B}_{R/2}$, $\tilde{\pi}_R(x) = \tilde{\pi}(x)$, so $\|b(x) - \tilde{b}_R(x)\|_2 \leq \epsilon/2$. Otherwise $x \in \mathcal{X} \setminus \mathbb{B}_{R/2}$, in which case since $\|f_R(x)\| \leq \epsilon/2$ it follows that $\|b(x) - \tilde{b}_R(x)\|_2 \leq \epsilon$. To verify condition (2), calculate that for $x \in \mathcal{X} \setminus \mathbb{B}_R$,

$$x \cdot (b(x) - \tilde{b}_R(x)) = x \cdot \left(b(x) - \tilde{b}(x) - \frac{\epsilon x}{2\|x\|_2}\right) \geq \frac{\epsilon\|x\|_2}{2} - \frac{x \cdot \epsilon x}{2\|x\|_2} = 0.$$

□

By taking R very large in Proposition C.3.2, we can ensure the integrability condition holds without having any practical effect on the approximating drift since $\tilde{b}_R(x) = \tilde{b}(x)$ for all $x \in \mathbb{B}_{R/2}$. Thus, it is safe to view Assumption 4.B(3) as a mild regularity condition.

C.4 Approximation Results for Piecewise Deterministic Markov Processes

In the section we obtain results for a broader class of PDMPs which includes the ZZP a special case [16]. The class of PDMPs we consider are defined on the space $E := \mathbb{R}^d \times \mathcal{B}$, where $\mathcal{B}$ is a finite set. Let $A \in C^0(E, \mathbb{R}_+^{\mathcal{B}})$ and let $F \in C^0(E, \mathbb{R}^d)$ be such that for each $\theta \in \mathcal{B}$, $F(\cdot, \theta)$ is a smooth vector field for which the differential equation $\partial_t x_t = F(x_t, \theta)$ with initial condition $x_0 = x$ has a unique global solution. For $\phi \in C(E)$, the standard differential operator $\nabla_x \phi(x, \theta) \in \mathbb{R}^d$ is given by $(\nabla_x \phi(x, \theta))_i := \frac{\partial \phi}{\partial x_i}(x, \theta)$ for $i \in [d]$ and the discrete differential operator $\nabla_\theta \phi(x, \theta) \in \mathbb{R}^{\mathcal{B}}$ is given by $(\nabla_\theta \phi(x, \theta))_{\theta'} := \phi(x, \theta') - \phi(x, \theta)$. The PDMP $(X_t, \Theta_t)_{t \geq 0}$ determined by the pair (F, A) has infinitesimal generator

$$\mathcal{A}_{F,A} \phi = F \cdot \nabla_x \phi + A \cdot \nabla_\theta \phi.$$

We consider the cases when either or both of A and F are approximated (in the case of ZZP, only A is approximated while F is exact). The details of the polynomial contractivity condition depend on which parts of (F, A) are approximated. We use the same notation for the true and approximating PDMPs with, respectively, infinitesimal generators $\mathcal{A}_{F,A}$ and $\mathcal{A}_{\tilde{F},\tilde{A}}$, as we did for the ZZPs in Section 4.5.

Assumption D.10 (**PDMP error and polynomial contractivity**).

1. *There exist $\epsilon_F, \epsilon_A \geq 0$ such that $\|F - \tilde{F}\|_2 \leq \epsilon_F$ and $\|A - \tilde{A}\|_1 \leq \epsilon_A$.*

2. *For each $(x, \theta) \in E$, let $\mu_{x,\theta,t}$ denote the law of the PDMP $(X_{x,\theta,t}, \Theta_{x,\theta,t})$ with generator $\mathcal{A}_{F,A}$. There exist constants $\alpha > 1$ and $\beta > 0$ and a function $B \in C(E \times E, \mathbb{R}_+)$ such that for all $x, x' \in \mathbb{R}^d$ and $\theta, \theta' \in \mathcal{B}$,*

$$d_{\mathcal{W}}(\mu_{x,\theta,t}, \mu_{x',\theta',t}) \leq B(x, \theta, x', \theta')(t + \beta)^{-\alpha}.$$

Furthermore, if $\epsilon_F > 0$, then there exists $C_F > 0$ such that $B(x, \theta, x', \theta) \leq C_F \|x - x'\|_2$ and if $\epsilon_A > 0$, then there exists $C_A > 0$ such that $B(x, \theta, x, \theta') \leq C_A$. If $\epsilon_F = 0$ take $C_F = 0$ and if $\epsilon_A = 0$ take $C_A = 0$.

We also require some regularity conditions similar to those for diffusions.

Assumption D.11 (**PDMP regularity conditions**). *Let π and $\tilde{\pi}$ denote the stationary distributions of the PDMPs with, respectively, infinitesimal generators $\mathcal{A}_{F,A}$ and $\mathcal{A}_{\tilde{F},\tilde{A}}$.*

1. *The stationary distributions π and $\tilde{\pi}$ exist.*

2. *The target density satisfies $\int_E x^2 \pi(\mathrm{d}x, \mathrm{d}\theta) < \infty$.*

3. *If a function $\phi \in C(E, \mathbb{R})$ is π-integrable then it is $\tilde{\pi}$-integrable.*

Theorem C.4.1 (PDMP error bounds)**.** *If Assumptions D.10 and D.11 hold, then*

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C_F \epsilon_F + C_A \epsilon_A}{(\alpha - 1)\beta^{\alpha - 1}}.$$

Proof [Proof sketch] For $h \in C_L(\mathbb{R}^d)$, we need to solve

$$h - \pi(h) = \mathcal{A}_{F,A} u.$$

Similarly to before, the solution is

$$u_h(x, \theta) := \int_0^\infty (\pi(h) - \mathbb{E}[h(X_{x,\theta,t})]) \, \mathrm{d}t,$$

which can be verified as in the diffusion case using Assumptions D.10(2) and D.11. Furthermore, for $x, x' \in \mathbb{R}^d$ and $\theta, \theta' \in \mathcal{B}$, by Assumption D.10(2),

$$\begin{aligned} |u_h(x, \theta) - u_h(x', \theta)| &\leq \|h\|_L \int_0^\infty C_F \|x - x'\|_2 (t + \beta)^{-\alpha} \, \mathrm{d}t \\ &= \frac{C_F \|h\|_L}{(\alpha - 1)\beta^{\alpha - 1}} \|x - x'\|_2 \end{aligned}$$

and

$$|u_h(x, \theta) - u_h(x, \theta')| \leq \|h\|_L \int_0^\infty C_A (t + \beta)^{-\alpha} \, \mathrm{d}t = \frac{C_A \|h\|_L}{(\alpha - 1)\beta^{\alpha - 1}}.$$

We bound $d_{\mathcal{W}}(\pi, \tilde{\pi})$ as in Theorem 4.2.4, but now using the fact that for $u = u_h$, $h \in C_L(\mathbb{R}^d)$, we have

$$\begin{aligned} \mathcal{A}_{F,A} u_h - \mathcal{A}_{\tilde{F},\tilde{A}} u_h &= (F - \tilde{F}) \cdot \nabla_x u_h + (A - \tilde{A}) \cdot \nabla_\theta u_h \\ &\leq \|F - \tilde{F}\|_2 \|\nabla_x u_h\|_2 + \|A - \tilde{A}\|_1 \|\nabla_\theta u_h\|_\infty \\ &\leq \frac{C_F \epsilon_F + C_A \epsilon_A}{(\alpha - 1)\beta^{\alpha - 1}}. \end{aligned}$$

□

C.4.1 Hamiltonian Monte Carlo

We can write an idealized form of Hamiltonian Monte Carlo (HMC) as a PDMP $(X_t, P_t)_{t \geq 0}$ by having the momentum vector $P_t \in \mathbb{R}^d$ refresh at a constant rate λ.

Let R_t be a compound Poisson process with rate $\lambda > 0$ and jump size distribution $\mathcal{N}(0, M)$, where $M \in \mathbb{R}^{d\times d}$ is a positive-definite mass matrix. That is, if Γ_t is a homogenous Poisson (counting) process with rate λ and $J_i \sim \mathcal{N}(0, M)$, then

$$R_t \sim \sum_{i=1}^{\Gamma_t} J_i.$$

We can then write the HMC dynamics as

$$\begin{aligned} \mathrm{d}X_t &= M^{-1}P_t\,\mathrm{d}t \\ \mathrm{d}P_t &= \nabla \log \pi(X_T)\,\mathrm{d}t + \mathrm{d}R_t. \end{aligned}$$

The infinitesimal generator for $(X_t, P_t)_{t\geq 0}$ is

$$\begin{aligned} &\mathcal{A}_{\lambda,M,\pi}\phi(x,p) \\ &= M^{-1}p \cdot \nabla_x \phi(x,p) + \nabla \log \pi(x) \cdot \nabla_p \phi(x,p) + \lambda\left(\int \phi(x,p')\nu_M(\mathrm{d}p') - \phi(x,p)\right), \end{aligned}$$

where ν_M is the density of $\mathcal{N}(0, M)$. Let $\mu_{x,p,t}$ denote the law of $(X_{x,p,t}, P_{x,p,t})$ with generator $\mathcal{A}_{\lambda,M,\pi}$. The proof of the following theorem is similar to that for Theorem C.4.1:

Theorem C.4.2 (HMC error bounds)**.** *Assume that:*

1. $\|\nabla \log \pi - \nabla \log \tilde{\pi}\|_2 \leq \epsilon$.
2. *There exist constants* $C > 0$ *and* $0 < \rho < 1$ *such that*

$$d_{\mathcal{W}}(\mu_{x,p,t}, \mu_{x,p',t}) \leq C\|p - p'\|_2\rho^t.$$

3. *The stationary distributions of the PDMPs with, respectively, infinitesimal generators* $\mathcal{A}_{\lambda,M,\pi}$ *and* $\mathcal{A}_{\lambda,M,\tilde{\pi}}$*, exist (they are, respectively,* $\pi \times \mu_M$ *and* $\tilde{\pi} \times \mu_M$*).*
4. *The target density satisfies* $\int_E x^2\pi(\mathrm{d}x) < \infty$.
5. *If a function* $\phi \in C(\mathbb{R}^d, \mathbb{R})$ *is* π*-integrable then it is* $\tilde{\pi}$*-integrable.*

Then

$$d_{\mathcal{W}}(\pi, \tilde{\pi}) \leq \frac{C\epsilon}{\log(1/\rho)}.$$

C.5 Analysis of computational–statistical trade-off

In this section we prove Theorem 4.4.1. In order to apply results on the approximation accuracy of ULA [27, 39, 43], we need the following property to hold for the exact and approximate drift functions.

Assumption E.12 (Strong log-concavity)**.** *There exists a positive constant $k_f > 0$ such that for all $x, x' \in \mathcal{X}$,*

$$(f(x) - f(x')) \cdot (x - x') \leq -k_f \|x - x'\|_2^2.$$

We restate the requirements given in Assumption 4.D with some additional notations.

Assumption E.13.

1. *The function $\log \pi_0 \in C^3(\mathbb{R}^d, \mathbb{R})$ is k_0-strongly concave, $L_0 := \|\nabla \log \pi_0\|_L < \infty$, and $\|\nabla^2[\partial_j \log \pi_0]\|_2 \leq M_0 < \infty$ for $j = 1, \ldots, d$.*

2. *There exist constants k_ϕ, L_ϕ, and M_ϕ such that for $i = 1, \ldots, N$, the function $\phi_i \in C^3(\mathbb{R}, \mathbb{R})$ is k_ϕ-strongly concave, $\|\phi_i'\|_L \leq L_\phi < \infty$, and $\|\phi_i'''\|_\infty \leq M_\phi < \infty$.*

3. *The matrix $A_N := \sum_{i=1}^N y_i y_i^\top$ satisfies $\|A_N\|_2 = \Theta(N)$.*

Note that under Assumption E.12, there is a unique $x^\star \in \mathbb{R}^d$ such that $f(x^\star) = 0$. Our results in this section on based on the following bound on the Wasserstein distance between the law of ULA Markov chain and π_f:

Theorem C.5.1 ([43, Theorem 3], [44, Corollary 3])**.** *Assume that E.12 holds and the $L_f := \|f\|_L < \infty$. Let $\kappa_f := 2k_f L_f/(k_f + L_f)$ and let $\mu_{x,T}$ denote the law of $X'_{x,T}$. Take $\gamma_i = \gamma_1 i^{-\alpha}$ with $\alpha \in (0, 1)$ and set*

$$\gamma_1 = 2(1 - \alpha)\kappa_f^{-1}(2/T)^{1-\alpha} \log\left(\frac{\kappa_f T}{2(1 - \alpha)}\right).$$

If $\gamma_1 < 1/(k_f + L_f)$, then

$$d_{\mathcal{W}}^2(\mu_{x,T}, \pi_f) \leq 16(1 - \alpha)L_f^2 \kappa_f^{-3} d T^{-1} \log\left(\frac{\kappa_f T}{2(1 - \alpha)}\right).$$

For simplicity we fix $\alpha = 1/2$, though the same results hold for all $\alpha \in (0, 1)$, just with different constants. Take $\{\gamma_i\}_{i=1}^\infty$ as defined in Theorem C.5.1. Let $x^\star := \arg\max_x \mathcal{L}(x)$ and let $S_k := \sum_{i=1}^N \|y_i\|_2^k$. The drift for this model is

$$b(x) := \nabla \mathcal{L}(x) = \nabla \log \pi_0(x) + \sum_{i=1}^N \phi_i'(x \cdot y_i) y_i.$$

By Taylor's theorem, the j-th component of $b(x)$ can be rewritten as

$$\begin{aligned} b_j(x) &= \partial_j \log \pi_0(x^\star) + \nabla \partial_j \log \pi_0(x^\star) \cdot (x - x^\star) + R(\partial_j \log \pi_0, x) \\ &\quad + \sum_{i=1}^N \phi_i'(x^\star \cdot y_i) y_{ij} + \phi_i''(x^\star \cdot y_i) y_{ij} y_i \cdot (x - x^\star) + R(\phi_i'(\cdot \cdot y_i) y_{ij}, x) \\ &= \nabla \partial_j \log \pi_0(x^\star) \cdot (x - x^\star) + R(\partial_j \log \pi_0, x) \\ &\quad + \sum_{i=1}^N \phi_i''(x^\star \cdot y_i) y_{ij} y_i \cdot (x - x^\star) + R(\phi_i'(\cdot \cdot y_i) y_{ij}, x), \end{aligned} \tag{C.5.1}$$

where

$$R(f, x) := \|x - x^\star\|_2^2 \int_0^1 (1 - t) \nabla^2 f(x^\star + t(x - x^\star)) \, \mathrm{d}t.$$

Hence we can approximate the drift with a first-order Taylor expansion around $x^\star$:

$$\tilde{b}(x) := (\nabla^2 \log \pi_0)(x^\star)(x - x^\star) + \sum_{i=1}^N \phi_i''(x^\star \cdot y_i) y_i y_i^\top (x - x^\star).$$

Observe that Assumption E.12 is satisfied for $f = b$ and $f = \tilde{b}$ with $k_f = k_N := k_0 + k_\phi \|A_N\|_2$. Furthermore, Assumption 4.B is satisfied with $\|\tilde{b}\|_L \le L_N := L_0 + L_\phi S_2$ and $\|b\|_L \le L_N$ as well since

$$\begin{aligned} \|\phi_i'(x_1 \cdot y_i) y_i - \phi_i'(x_2 \cdot y_i) y_i\|_2 &\le |\phi_i'(x_1 \cdot y_i) - \phi_i'(x_2 \cdot y_i)| \|y_i\|_2 \\ &\le L_\phi |x_1 \cdot y_i - x_2 \cdot y_i| \|y_i\|_2 \\ &\le L_\phi \|y_i\|_2^2 \|x_1 - x_2\|_2. \end{aligned}$$

Thus, b and $\tilde{b}$ satisfy the same regularity conditions.

We next show that they cannot deviate too much from each other. Using Eq. (C.5.1) and regularity assumptions we have

$$\begin{aligned} \|b(x) - \tilde{b}(x)\|_2^2 &= \sum_{j=1}^d \left(R(\partial_j \log \pi_0, x) + \sum_{i=1}^N R(\phi_i'(\cdot \cdot y_i) y_{ij}, x) \right)^2 \\ &\le \|x - x^\star\|_2^4 \sum_{j=1}^d \left(M_0 + \sum_{i=1}^N M_\phi \|y_i\|_2^2 y_{ij} \right)^2 \\ &\le d \|x - x^\star\|_2^4 \left(M_0 + M_\phi \sum_{i=1}^N \|y_i\|_2^3 \right)^2. \end{aligned}$$

It follows from [43, Theorem 1(ii)] that

$$\tilde{\pi}(\|b - \tilde{b}\|_2) \le d^{3/2} M_N k_N^{-1},$$

where $M_N := M_0 + M_\phi S_3$.

Putting these results together with Theorems 4.2.1 and C.5.1 and applying the triangle inequality, we conclude that

$$d^2_{\mathcal{W}}(\mu^\star_T, \pi) \leq \frac{(k_N + L_N)^3 d}{k_N^3 L_N} \frac{\log\left(\frac{2k_N L_N}{k_N + L_N} T\right)}{T}$$
$$d^2_{\mathcal{W}}(\tilde{\mu}^\star_{\tilde{T}}, \pi) \leq \frac{2(k_N + L_N)^3 d}{k_N^3 L_N} \frac{\log\left(\frac{2k_N L_N}{k_N + L_N} \tilde{T}\right)}{\tilde{T}} + \frac{2d^3 M_N^2}{k_N^4}.$$

In order to compare the bounds we must make the computational budgets of the two algorithms equal. Recall that we measure computational cost by the number of d-dimensional inner products performed, so ULA with b costs TN and ULA with $\tilde{b}$ costs $(\tilde{T} + N)d$. Equating the two yields $\tilde{T} = N(T/d - 1)$, so we must assume that $T > d$. For the purposes of asymptotic analysis, assume also that S_i/N is bounded from above and bounded away from zero. Under these assumptions, in the case of $k_\phi > 0$, we conclude that

$$d^2_{\mathcal{W}}(\mu^\star_T, \pi) = \tilde{O}\left(\frac{d}{TN}\right) \qquad \text{and} \qquad d^2_{\mathcal{W}}(\tilde{\mu}^\star_{\tilde{T}}, \pi) = \tilde{O}\left(\frac{d^2}{N^2 T} + \frac{d^3}{N^2}\right),$$

establishing the result of Theorem 4.4.1. For large N, the approximate ULA with $\tilde{b}$ is more accurate.

Appendix D

Chapter 5 Proofs

D.1 Proof of Theorem 5.4.1: Tilted KSDs detect non-convergence

The result will follow from the following theorem which provides an upper bound on the bounded Lipschitz metric $d_{BL_{\|\cdot\|_2}}(\mu, P)$ in terms of the KSD and properties of A and Φ.

Theorem D.1.1 (Tilted KSD lower bound)**.** *Suppose $P \in \mathcal{P}$ and $k(x, y) = A(x)\Phi(x-y)A(y)$ with $\Phi \in C^2$, $A \in C^1$ positive, $1/A \in L^2$, and $\nabla \log A$ bounded and Lipschitz continuous. Then there exists a constant $\mathcal{M}_P$ such that, for all $\epsilon, \delta > 0$ and all probability measures μ,*

$$d_{BL_{\|\cdot\|_2}}(\mu, P) \leq \epsilon + (2\pi)^{-d/4}\mathcal{C}(\epsilon)\,\mathrm{KSD}_k(\mu, P)$$

for

$$\mathcal{C}(\epsilon) := \|1/A\|_{L^2}\mathcal{M}_P F\big(\mathbb{E}[\|G\|_2 B(G)](1 + M_1(\log A) + \mathcal{M}_P M_1(b + \nabla \log A))\epsilon^{-1}\big)^{1/2},$$

where $F(t) \triangleq \sup_{\omega\in\mathbb{R}^d} e^{-\|\omega\|_2^2/(2t^2)}/\hat{\Phi}(\omega)$, G is a standard Gaussian vector, and $B(y) \triangleq \sup_{x\in\mathbb{R}^d, u\in[0,1]} A(x)/A(x+uy)$.

Remarks By bounding F and optimizing over ϵ, one can derive rates of convergence in $d_{BL_{\|\cdot\|_2}}$. Thm. 5 and Sec. 4.2 of Gorham et al. [62] provide an explicit value for the *Stein factor* $\mathcal{M}_P$.

Since $\log A$ is Lipschitz, $B(y) \leq e^{\|y\|_2}$ so $\mathbb{E}[\|G\|_2 B(G)]$ is finite. Now suppose $\mathrm{KSD}_k(\mu_n, P) \to 0$ for a sequence of probability measures $(\mu_n)_{n\geq 1}$. For any $\epsilon > 0$, $\limsup_n d_{BL_{\|\cdot\|_2}}(\mu_n, P) \leq \epsilon$, since $F(t)$ is finite for all $t > 0$. Hence, $d_{BL_{\|\cdot\|_2}}(\mu_n, P) \to 0$, and, as $d_{BL_{\|\cdot\|_2}}$ metrizes weak convergence, $\mu_n \Rightarrow P$.

D.1.1 Proof of Theorem D.1.1: Tilted KSD lower bound

Our proof parallels that of [61, Thm. 13]. Fix any $h \in BL_{\|\cdot\|_2}$. Since $A \in C^1$ is positive, Thm. 5 and Sec. 4.2 of Gorham et al. [62] imply that there exists a $g \in C^1$

which solves the Stein equation $\mathcal{T}_P(Ag) = h - \mathbb{E}_P[h(Z)]$ and satisfies $M_0(Ag) \le \mathcal{M}_P$ for $\mathcal{M}_P$ a constant independent of A, h, and g. Since $1/A \in L^2$, we have $\|g\|_{L^2} \le \mathcal{M}_P \|1/A\|_{L^2}$.

Since $\nabla \log A$ is bounded, $A(x) \le \exp \gamma\|x\|$ for some γ. Moreover, any measure in $\mathcal{P}$ is sub-Gaussian, so P has finite exponential moments. Hence, since A is also positive, we may define the tilted probability measure P_A with density proportional to Ap. The identity $\mathcal{T}_P(Ag) = A\mathcal{T}_{P_A} g$ implies that

$$M_0(A\nabla\mathcal{T}_{P_A} g) = M_0(\nabla\mathcal{T}_P(Ag) - \mathcal{T}_P(Ag)\nabla \log A) \le 1 + M_1(\log A).$$

Since b and $\nabla \log A$ are Lipschitz, we may apply the following lemma, proved in Appendix D.1.2 to deduce that there is a function $g_\epsilon \in \mathcal{K}^d_{k_1}$ for $k_1(x, y) \triangleq \Phi(x - y)$ such that $|(\mathcal{T}_P(Ag_\epsilon))(x) - (\mathcal{T}_P(Ag))(x)| = A(x)|(\mathcal{T}_{P_A} g_\epsilon)(x) - (\mathcal{T}_{P_A} g)(x)| \le \epsilon$ for all x with norm

$$\begin{aligned}&\|g_\epsilon\|_{\mathcal{K}^d_{k_1}} \\ &\le (2\pi)^{-d/4} F\big(\mathbb{E}[\|G\|_2 B(G)](1 + M_1(\log A) + \mathcal{M}_P M_1(b + \nabla \log A))\epsilon^{-1}\big)^{1/2} \|1/A\|_{L^2} \mathcal{M}_P.\end{aligned} \tag{D.1.1}$$

Lemma D.1.2 (Stein approximations with finite RKHS norm)**.** *Consider a function $A : \mathbb{R}^d \to \mathbb{R}$ satisfying $B(y) \triangleq \sup_{x\in\mathbb{R}^d, u\in[0,1]} A(x)/A(x + uy)$. Suppose $g : \mathbb{R}^d \to \mathbb{R}^d$ is in $L^2 \cap C^1$. If P has Lipschitz log density, and $k(x, y) = \Phi(x - y)$ for $\Phi \in C^2$ with generalized Fourier transform $\hat{\Phi}$, then for every $\epsilon \in (0, 1]$, there is a function $g_\epsilon : \mathbb{R}^d \to \mathbb{R}^d$ such that $|(\mathcal{T}_P g_\epsilon)(x) - (\mathcal{T}_P g)(x)| \le \epsilon/A(x)$ for all $x \in \mathbb{R}^d$ and*

$$\|g_\epsilon\|_{\mathcal{K}^d_k} \le (2\pi)^{-d/4} F(\mathbb{E}[\|G\|_2 B(G)](M_0(A\nabla\mathcal{T}_P g) + M_0(Ag)M_1(b))\epsilon^{-1})^{1/2} \|g\|_{L^2},$$

where $F(t) \triangleq \sup_{\omega\in\mathbb{R}^d} e^{-\|\omega\|_2^2/(2t^2)}/\hat{\Phi}(\omega)$ and G is a standard Gaussian vector.

Since $\|Ag_\epsilon\|_{\mathcal{K}^d_k} = \|g_\epsilon\|_{\mathcal{K}^d_{k_1}}$, the triangle inequality and the definition of the KSD now yield

$$\begin{aligned}&|\mathbb{E}_\mu[h(X)] - \mathbb{E}_P[h(Z)]| \\ &\quad= |\mathbb{E}_\mu[(\mathcal{T}_P(Ag))(X)]| \le |\mathbb{E}[(\mathcal{T}_P(Ag))(X) - (\mathcal{T}_P(Ag_\epsilon))(X)]| + |\mathbb{E}_\mu[(\mathcal{T}_P(Ag_\epsilon))(X)]| \\ &\quad\le \epsilon + \|g_\epsilon\|_{\mathcal{K}^d_{k_1}} \ \mathrm{KSD}_k(\mu, P).\end{aligned}$$

The advertised conclusion follows by applying the bound Eq. (D.1.1) and taking the supremum over all $h \in BL_{\|\cdot\|}$.

D.1.2 Proof of Lemma D.1.2: Stein approximations with finite RKHS norm

Our proof parallels that of Gorham and Mackey [61, Lem. 12]. Let Y denote a standard Gaussian vector with density ρ. For each $\delta \in (0, 1]$, we define $\rho_\delta(x) = \delta^{-d}\rho(x/\delta)$,

and for any function f we write $f_\delta(x) \triangleq \mathbb{E}[f(x+\delta Y)]$. Under our assumptions on $h = \mathcal{T}_P g$ and B, the mean value theorem and Cauchy-Schwarz imply that for each $x \in \mathbb{R}^d$ there exists $u \in [0,1]$ such that

$$\begin{aligned}|h_\delta(x) - h(x)| &= |\mathbb{E}_\rho[h(x+\delta Y) - h(x)]| = |\mathbb{E}_\rho[\langle \delta Y, \nabla h(x+\delta Y u)\rangle]| \\ &\le \delta\, \mathbb{E}_\rho[\|Y\|_2 / A(x+\delta Y u)] \le \delta M_0(A \nabla \mathcal{T}_P g)\, \mathbb{E}_\rho[\|Y\|_2 B(Y)]/A(x).\end{aligned}$$

Now, for each $x \in \mathbb{R}^d$ and $\delta > 0$,

$$\begin{aligned}h_\delta(x) &= \mathbb{E}_\rho[\langle b(x+\delta Y), g(x+\delta Y)\rangle] + \mathbb{E}[\langle \nabla, g(x+\delta Y)\rangle] \quad \text{and} \\ (\mathcal{T}_P g_\delta)(x) &= \mathbb{E}_\rho[\langle b(x), g(x+\delta Y)\rangle] + \mathbb{E}[\langle \nabla, g(x+\delta Y)\rangle],\end{aligned}$$

so, by Cauchy-Schwarz, the Lipschitzness of b, and our assumptions on g and B,

$$\begin{aligned}|(\mathcal{T}_P g_\delta)(x) - h_\delta(x)| &= |\mathbb{E}_\rho[\langle b(x) - b(x+\delta Y), g(x+\delta Y)\rangle]| \\ &\le \mathbb{E}_\rho[\|b(x) - b(x+\delta Y)\|_2 \|g(x+\delta Y)\|_2] \\ &\le M_0(Ag) M_1(b)\, \delta\, \mathbb{E}_\rho[\|Y\|_2 / A(x+\delta Y)] \\ &\le M_0(Ag) M_1(b)\, \delta\, \mathbb{E}_\rho[\|Y\|_2 B(Y)]/A(x).\end{aligned}$$

Thus, if we fix $\epsilon > 0$ and define $\tilde{\epsilon} = \epsilon / (\mathbb{E}_\rho[\|Y\|_2 B(Y)](M_0(A\nabla \mathcal{T}_P g) + M_0(Ag) M_1(b)))$, the triangle inequality implies

$$|(\mathcal{T}_P g_{\tilde{\epsilon}})(x) - (\mathcal{T}_P g)(x)| \le |(\mathcal{T}_P g_{\tilde{\epsilon}})(x) - h_{\tilde{\epsilon}}(x)| + |h_{\tilde{\epsilon}}(x) - h(x)| \le \epsilon / A(x).$$

To conclude, we will bound $\|g_\delta\|_{\mathcal{K}_k^d}$. By Wendland [143, Thm. 10.21],

$$\begin{aligned}\|g_\delta\|_{\mathcal{K}_k^d}^2 &= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{|\hat{g}_\delta(\omega)|^2}{\hat{\Phi}(\omega)}\, d\omega \\ &= (2\pi)^{d/2} \int_{\mathbb{R}^d} \frac{|\hat{g}(\omega)|^2 \hat{\rho}_\delta(\omega)^2}{\hat{\Phi}(\omega)}\, d\omega \\ &\le (2\pi)^{-d/2} \left\{ \sup_{\omega \in \mathbb{R}^d} \frac{e^{-\|\omega\|_2^2 \delta^2/2}}{\hat{\Phi}(\omega)} \right\} \int_{\mathbb{R}^d} |\hat{g}(\omega)|^2\, d\omega,\end{aligned}$$

where we have used the Convolution Theorem [143, Thm. 5.16] and the identity $\hat{\rho}_\delta(\omega) = \hat{\rho}(\delta\omega)$. Finally, an application of Plancherel's theorem [70, Thm. 1.1] gives $\|g_\delta\|_{\mathcal{K}_k^d} \le (2\pi)^{-d/4} F(\delta^{-1})^{1/2} \|g\|_{L^2}$.

D.2 Proof of Proposition 5.4.3

For each $j \in [J]$, apply Corollary D.13.2 with δ/J in place of δ to the random variable

$$\frac{1}{M} \sum_{m=1}^{M} \frac{|(\mathcal{F}^{-1} \Delta_N k_j^{1/2})(Z_m)|^r}{\nu(Z_m)}.$$

The result follows by plugging in the high probability lower bounds from Corollary D.13.2 into $\mathrm{fGMMD}^2_{\boldsymbol{k}^{1/2},r,\nu,M}(Q_N, P)$ and using the union bound.

D.3 Proof of Proposition 5.4.4

Let $w_j(z) := |(\mathcal{F}^{-1}\Delta_N k_j^{1/2})(z)|^r/\nu(z)$. It follows from the definition of $\mathcal{Q}(\boldsymbol{k}^{1/2}, \nu, c)$ and the finiteness condition on P that for some $c' > 0$

$$\sup_{Q_N \in \mathcal{Q}(\boldsymbol{k}^{1/2},\nu,c)} \sup_{j,z} |(\mathcal{F}^{-1}\Delta_N k_j^{1/2})(z)|^r/\nu(z) \leq c'.$$

Hence for any $Q_N \in \mathcal{Q}(\boldsymbol{k}^{1/2}, \nu, c)$ and $j \in [J]$,

$$\mathbb{E}[Y_j^2] \leq c'\mathbb{E}[Y_j].$$

D.4 Proof of Proposition 5.4.5

We have

$$\begin{aligned}
k_j(x,y) &:= \int k_j^{1/2}(x,\omega)\overline{k_j^{1/2}(y,\omega)}\rho(\omega)\,\mathrm{d}\omega \\
&= \int \mathcal{F}(\mathcal{O}_{j,x}f_j(x-\cdot))(\omega)\overline{\mathcal{F}(\mathcal{O}_{j,y}f_j(y-\cdot))(\omega)}\rho(\omega) \\
&= \int (\mathcal{O}_{j,x}f_j(x-\cdot) * \breve{\rho}^{1/2})(z)\,(\mathcal{O}_{j,y}f_j(\cdot - y) * \breve{\rho}^{1/2})(z)\,\mathrm{d}z \\
&= \mathcal{O}_{j,x}\mathcal{O}_{j,y}(f_j * \breve{\rho}^{1/2} * f_j * \breve{\rho}^{1/2})(x-y) \\
&= \mathcal{O}_{j,x}\mathcal{O}_{j,y}\mathcal{F}(\widehat{f_j}^2\rho)(x-y) \\
&= \mathcal{O}_{j,x}\mathcal{O}_{j,y}\Phi_j(x-y).
\end{aligned}$$

It now follows that we can rewrite the MMD under k_j as

$$\begin{aligned}
(\Delta_N \times \Delta_N)k_j(x,y) &= (\Delta_N\mathcal{O}_{j,x} \times \Delta_N\mathcal{O}_{j,y})\Phi_j(x-y) \\
&= (\Delta_N\mathcal{O}_{j,x} \times \Delta_N\mathcal{O}_{j,y})\langle\Phi_j(x-\cdot), \Phi_j(y-\cdot)\rangle_{\Phi_j} \\
&= \|\Delta_N\mathcal{O}_{j,x}\Phi_j(x-\cdot)\|^2_{\Phi_j}.
\end{aligned}$$

D.5 Proof of Theorem 5.4.6: (c, α) second moment bounds for fGMMD

Assumption E.14. *For each $j \in [J]$, $\omega_j^2\hat{\Phi}_j^{1/2}(\omega)$ is integrable.*

Assumption E.15. *For each $j \in [J]$, there exists a function B_j such that $|\mathcal{O}_{j,x}f_j(x-z)| \leq B_j(x,z)f_j(z)$.*

Assumption E.16. *For some norm $\|\cdot\|$ and for each $j \in [J]$, there exists a continuous nonincreasing function $\overline{f_j}$ and a constant $C_{f,j} > 0$ such that for all z, $\overline{f_j}(\|z\|) \leq f_j(z) \leq C_{f,j}\overline{f_j}(\|z\|)$ and $\lim_{R\to\infty} \overline{f_j}(R) = 0$.*

Assumption E.17. *There exists $b \in [0, 1-\xi)$ such that for each $j \in [J]$, there exists $C_{B,j} > 0$ such that $(|\Delta_N|B_j)(z) \leq C_{B,j} f_j(z)^{-b}$*

Assumption E.18. *For each $j \in [J]$, there exists $C_{\mathcal{O},j} > 0$ such that $\sup_\omega (1+\omega_j)^{-1} |\int \mathcal{O}_{j,x} e^{-i\omega\cdot x} \Delta_N(\mathrm{d}x)| \leq C_{\mathcal{O},j}$.*

Lemma D.5.1. *Let $\psi_j(\omega) := (1+\omega_j)^{-1} \Delta_N \mathcal{O}_{j,x} e^{-i\omega\cdot x}$. If Assumptions 5.F, 5.G, E.14 and E.18 hold, then for any $\lambda \in (1/2, \overline{\lambda})$,*

$$\begin{aligned}|\Delta_N \mathcal{O}_{j,x} f_j(x-z)| &\leq \|f_j\|_{\Phi_j^{(\lambda)}} \left(\|\psi_j\|_{L^\infty} \left\| (1+\partial_{x_j}) \Phi_j^{(1/4)} \right\|_{L^2} \right)^{2-2\lambda} \|\Delta_N \mathcal{O}_{j,x} \Phi_j(\hat{x} - \cdot)\|_{\Phi_j}^{2\lambda-1} \\ &= C_{j,\lambda,d}\, \mathrm{MMD}_{k_j}^{2\lambda-1}.\end{aligned}$$

Proof Apply Proposition D.6.1 with $\mathcal{D} = \Delta_N \mathcal{O}_{j,x}$, $f = f_j$, $h(\omega) = 1 + \omega_j$, and $t = 1/2$. The equality follows from Proposition 5.4.5. □

Lemma D.5.2. *If Assumptions 5.F, E.15 and E.17 hold, then*

$$|\Delta_N \mathcal{O}_{j,x} f_j(x-z)| \leq C_{B,j} f_j(z)^{1-b}.$$

Proof We have

$$|\Delta_N \mathcal{O}_{j,x} f_j(x-z)| \leq |\Delta_N||\mathcal{O}_{j,x} f_j(x-z)| \leq |\Delta_N| B_j(\cdot, z) f_j(z) \leq C_{B,j} f_j(z)^{1-b}.$$

□

Note that

$$w_j(z) := |(\mathcal{F}^{-1} \Delta_N k_j^{1/2})(z)|^r / \nu(z) = |\Delta_N \mathcal{O}_{j,x} f_j(x-z)|^r / \nu(z).$$

For a set S let $\nu_S(S') := \int_{S\cap S'} \nu(\mathrm{d}z)$. Let $K := \mathbb{B}_{\|\cdot\|}(R)$. Let $Z \sim \nu$ and $Y_j = w_j(Z)$. We have

$$\begin{aligned}\mathbb{E}[Y_j^2] &= \mathbb{E}[w_j(Z)^2] = \mathbb{E}[w_j(Z)^2 \mathbb{1}(Z \in K)] + \mathbb{E}[w_j(Z)^2 \mathbb{1}(Z \notin K)] \\ &\leq \|w_j\|_{L^1(\nu)} \|w_j \mathbb{1}(\cdot \in K)\|_{L^\infty(\nu)} + \|\mathbb{1}(\cdot \notin K)\|_{L^1(\nu)} \left\| w_j^2 \mathbb{1}(\cdot \notin K) \right\|_{L^\infty(\nu)} \\ &= \left\| \mathcal{F}^{-1} \Delta_N k_j^{1/2} \right\|_{L^r}^r \sup_{z\in K} w_j(z) + \nu(K^\complement) \sup_{z \in K^\complement} w_j(z)^2 \\ &= \mathbb{E}[Y_j] \sup_{z\in K} w_j(z) + \nu(K^\complement) \sup_{z \in K^\complement} w_j(z)^2\end{aligned}$$

Applying Lemma D.5.1 and Assumption 5.H we have

$$\begin{aligned}\sup_{z\in K} w_j(z) &\le C_{j,\lambda,d}^r \operatorname{MMD}_{k_j}^{r(2\lambda-1)} \sup_{z\in K} \nu(z)^{-1} \\ &\le C_{j,\lambda,d}^r C_{\nu,j} \sup_{z\in K} f_j(z)^{-\xi r} \operatorname{MMD}_{k_j}^{r(2\lambda-1)} \\ &\le C_{j,\lambda,d}^r C_{\nu,j} c_{r,d}^{r(2\lambda-1)} \|\rho\|_{L^t}^{r(\lambda-1/2)} \overline{f_j}(R)^{-\xi r} \operatorname{GMMD}_{k_j^{1/2},r}^{r(2\lambda-1)} \\ &= C_{j,\lambda,d}^r C_{\nu,j} c_{r,d}^{r(2\lambda-1)} \|\rho\|_{L^t}^{r(\lambda-1/2)} \overline{f_j}(R)^{-\xi r} \mathbb{E}[Y_j]^{2\lambda-1}\end{aligned}$$

Applying Lemma D.5.2 and Assumption 5.H we have

$$\begin{aligned}\sup_{z\in K^{\complement}} w_j(z)^2 &\le \sup_{z\in K^{\complement}} C_{B,j}^{2r} f_j(z)^{2(1-b)r} / \nu(z)^2 \\ &\le \sup_{z\in K^{\complement}} C_{B,j}^{2r} C_{\nu,j}^2 f_j(z)^{2(1-b-\xi)r} \\ &\le C_{B,j}^{2r} C_{\nu,j}^2 C_{f,j}^{2(1-b-\xi)r} \overline{f_j}(R)^{2(1-b-\xi)r}.\end{aligned}$$

Thus, we have that

$$\mathbb{E}[Y_j^2] \le C_{j,\lambda,d,r,\rho} \mathbb{E}[Y_j]^{2\lambda} \overline{f_j}(R)^{-\xi r} + C_{j,B,\xi} \overline{f_j}(R)^{2(1-b-\xi)r}.$$

As long as $\mathbb{E}[Y_j]^{2\lambda} \le C_{j,B,\xi} \overline{f_j}(0)^{2(1-b-\xi/2)r} / C_{j,\lambda,d,r,\rho}$, since $\overline{f_j}$ is continuous and non-increasing to zero we can choose R such that $\overline{f_j}(R)^{2(1-b-\xi)r} = C_{j,\lambda,d,r,\rho} \mathbb{E}[Y_j]^{2\lambda} / C_{j,B,\xi}$ and the result follows for $\mathbb{E}[Y_j]^{2\lambda} \le C_{j,B,\xi} \overline{f_j}(0)^{2(1-b-\xi/2)r} / C_{j,\lambda,d,r,\rho}$. Otherwise, we can guarantee that $\mathbb{E}[Y_j^2] \le C_\alpha \mathbb{E}[Y_j]^{2-\gamma_\alpha}$ be choosing C_α sufficiently large, since by assumption $\mathbb{E}[Y_j]$ is uniformly bounded over $Q_N \in \mathcal{Q}(\boldsymbol{k}^{1/2}, \nu, c, b, C_{B,j}, C_{\mathcal{O},j})$.

D.6 A uniform MMD-type bound

Let $\mathcal{D}$ denote a tempered distribution and Φ a stationary kernel. Also, define $\hat{\mathcal{D}}(\omega) := \mathcal{D}_x e^{-i\langle \omega, \hat{x}\rangle}$.

Proposition D.6.1. *Let f be a symmetric function such that for some $s \in (0,1]$, $f \in \mathcal{K}_{\Phi^{(s)}}$ and $\mathcal{D}_x f(\hat{x} - \cdot) \in \mathcal{K}_{\Phi^{(s)}}$. Then*

$$|\mathcal{D}_x f(\hat{x} - z)| \le \|f\|_{\Phi^{(s)}} \left\|\mathcal{D}_x \Phi^{(s)}(\hat{x} - \cdot)\right\|_{\Phi^{(s)}}$$

and for any $t \in (0, s)$ any function $h(\omega)$,

$$\left\|\mathcal{D}_x \Phi^{(s)}(\hat{x} - \cdot)\right\|_{\Phi^{(s)}}^{1-t} \le \left(\left\|h^{-1}\hat{\mathcal{D}}\right\|_{L^\infty} \left\|h\hat{\Phi}^{t/2}\right\|_{L^2}\right)^{1-s} \|\mathcal{D}_x \Phi(\hat{x} - \cdot)\|_{\Phi}^{s-t}.$$

Furthermore, if for some $c > 0$ and $r \in (0, s/2)$, $\hat{f} \le c\,\hat{\Phi}^r$, then

$$\|f\|_{\Phi^{(s)}} \le \frac{c \left\|\Phi^{(r-s/2)}\right\|_{L^2}}{(2\pi)^{d/4}}.$$

Proof The first inequality follows from an application of Cauchy-Schwartz:

$$\begin{aligned}
|\mathcal{D}_x f(\hat{x}-z)| &= |\langle f(\cdot - z), \mathcal{D}_x \Phi^{(s)}(\hat{x}-\cdot)\rangle_{\Phi^{(s)}}| \\
&\leq \|f(\cdot - z)\|_{\Phi^{(s)}} \left\|\mathcal{D}_x \Phi^{(s)}(\hat{x}-\cdot)\right\|_{\Phi^{(s)}} \\
&= \|f\|_{\Phi^{(s)}} \left\|\mathcal{D}_x \Phi^{(s)}(\hat{x}-\cdot)\right\|_{\Phi^{(s)}}.
\end{aligned}$$

For the first norm, we have

$$\begin{aligned}
\|f\|^2_{\Phi^{(s)}} &= (2\pi)^{-d/2} \int \frac{\hat{f}^2(\omega)}{\hat{\Phi}^s(\omega)}\,\mathrm{d}\omega \\
&\leq c^2 (2\pi)^{-d/2} \int \hat{\Phi}^{2r-s}(\omega)\,\mathrm{d}\omega \\
&= c^2 (2\pi)^{-d/2} \left\|\Phi^{(r-s/2)}\right\|^2_{L^2}.
\end{aligned}$$

Note that by the convolution theorem $\mathcal{F}(\mathcal{D}_x \Phi^{(s)}(\hat{x}-\cdot))(\omega) = \hat{\mathcal{D}}(\omega)\hat{\Phi}^s(\omega)$. For the second norm, applying Jensen's inequality and Hölder's inequality yields

$$\begin{aligned}
\left\|\mathcal{D}_x \Phi^{(s)}(\hat{x}-\cdot)\right\|^2_{\Phi^{(s)}} &= (2\pi)^{-d/2} \int \frac{\hat{\Phi}(\omega)^{2s} |\hat{\mathcal{D}}(\omega)|^2}{\hat{\Phi}^s(\omega)}\,\mathrm{d}\omega \\
&= (2\pi)^{-d/2} \left(\int \hat{\Phi}^t |\hat{\mathcal{D}}|^2\right) \int \frac{\hat{\Phi}(\omega)^t |\hat{\mathcal{D}}(\omega)|^2}{\int \hat{\Phi}^t |\hat{\mathcal{D}}|^2} \hat{\Phi}(\omega)^{s-t}\,\mathrm{d}\omega \\
&\leq (2\pi)^{-d/2} \left(\int \hat{\Phi}^t |\hat{\mathcal{D}}|^2\right) \left(\int \frac{\hat{\Phi}(\omega)^t |\hat{\mathcal{D}}(\omega)|^2}{\int \hat{\Phi}^t |\hat{\mathcal{D}}|^2} \Phi(\omega)^{1-t}\,\mathrm{d}\omega\right)^{\frac{s-t}{1-t}} \\
&= \left(\int \hat{\Phi}^t |\hat{\mathcal{D}}|^2\right)^{\frac{1-s}{1-t}} \|\mathcal{D}_x \Phi(\hat{x}-\cdot)\|^{2\frac{s-t}{1-t}}_{\Phi} \\
&\leq \left(\left\||h^{-1}\hat{\mathcal{D}}|^2\right\|_{L^\infty} \int h^2 \hat{\Phi}^t\right)^{\frac{1-s}{1-t}} \|\mathcal{D}_x \Phi(\hat{x}-\cdot)\|^{2\frac{s-t}{1-t}}_{\Phi} \\
&= \left(\left\|h^{-1}\hat{\mathcal{D}}\right\|^2_{L^\infty} \left\|h\hat{\Phi}^{t/2}\right\|^2_{L^2}\right)^{\frac{1-s}{1-t}} \|\mathcal{D}_x \Phi(\hat{x}-\cdot)\|^{2\frac{s-t}{1-t}}_{\Phi}.
\end{aligned}$$

□

We have

$$\begin{aligned}
\left\|\Delta_N \mathcal{O}_{j,x} \Phi^{(1/2)}(\hat{x}-\cdot)\right\|^2_{\Phi^{(1/2)}} &= (\Delta_N \mathcal{O}_{j,x} \times \Delta_N \mathcal{O}_{j,y}) h_j(\hat{x}-\hat{y}) \\
&= \int \Delta_N \mathcal{O}_{j,x} f_j(x-z) \Delta_N \mathcal{O}_{j,y} \hat{\rho}(z-y)^{1/2}\,\mathrm{d}z \\
&\leq \|\Delta_N \mathcal{O}_{j,x} f_j(x-\cdot)\|_{L^r} \left\|\Delta_N \mathcal{O}_{j,y} \hat{\rho}(\cdot - y)^{1/2}\right\|_{L^s}
\end{aligned}$$

D.7 Proof of Theorem 5.4.9

Let $K := \mathbb{B}_{\|\cdot\|}(R)$. Applying Lemmas D.5.1 and D.5.2 and Assumptions E.17 and E.18, we have

$$\begin{aligned}
\|\mathcal{F}^{-1}\Delta_N k_j^{1/2}\|_{L^r}^r &= \int |\Delta_N \mathcal{O}_{j,x} f_j(x-z)|^r \mathrm{d}z \\
&\leq \mathrm{vol}(K) \sup_{z\in K} |\Delta_N \mathcal{O}_{j,x} f_j(x-z)|^r + \int_{K^{\complement}} |\Delta_N \mathcal{O}_{j,x} f_j(x-z)|^r \mathrm{d}z \\
&\leq \mathrm{vol}(K) C_{j,\lambda,d}^r \,\mathrm{MMD}_{k_j}^{r(2\lambda-1)} + C_{B,j}^r \int_{K^{\complement}} f_j(z)^{r(1-b)} \mathrm{d}z \\
&\leq \mathrm{vol}(K) C_{j,\lambda,d}^r \,\mathrm{MMD}_{k_j}^{r(2\lambda-1)} + C_{B,j}^r C_{f,j}^r G_j(R).
\end{aligned}$$

Setting the terms equal and solving for R gives the first result.

D.8 Proof of Theorem 5.4.10

Applying Lemma D.5.1, we have

$$\begin{aligned}
\frac{1}{M}\sum_{m=1}^{M} \frac{|(\mathcal{F}^{-1}\Delta_N k_j^{1/2})(Z_m)|^r}{\nu(Z_m)} &\leq C_{j,\lambda,d}^r \frac{1}{M}\sum_{m=1}^{M} \frac{\mathrm{MMD}_{k_j}^{r(2\lambda-1)}}{\nu(Z_m)} \\
&= C_{j,\lambda,d}^r \left(\frac{1}{M}\sum_{m=1}^{M} \nu(Z_m)^{-1}\right) \mathrm{MMD}_{k_j}^{r(2\lambda-1)}.
\end{aligned}$$

The result now follows from the definition of fGMMD.

D.9 Proof of Theorem 5.4.7: Tilted hyperbolic secant fGMMD properties

We verify each of the assumptions in turn. By construction Assumption 5.F is satisfied with $\mathcal{O}_{j,x}(f)(x) = \mathcal{T}_{j,x} A(x) f(x)$. Note that since $\rho = 1$, $\hat{\Phi}_j = \hat{\Phi}_a^{\mathrm{sech}}$. Assumption 5.G holds since for any $\lambda \in (0,1)$, if follow from Proposition D.12.2 that

$$\widehat{f}_j / \hat{\Phi}_j^{\lambda/2} = \hat{\Phi}_{2a}^{\mathrm{sech}} / (\hat{\Phi}_a^{\mathrm{sech}})^{\lambda/2} \leq 2^{d/2} (\hat{\Phi}_{2a}^{\mathrm{sech}})^{1-\lambda} \in L^2.$$

Clearly Assumption E.14 holds as well.

Let $b := \nabla \log p$. To show Assumption E.15 holds, we apply Proposition D.12.3:

$$\begin{aligned}
&\mathcal{O}_{j,x} f_j(x-z) \\
&= \mathcal{T}_{j,x} A(x) \Phi_{2a}^{\mathrm{sech}}(x-z) \\
&= (\partial_{x_j} \log p(x) + \partial_{x_j} \log A(x) + \partial_{x_j} \log \Phi_{2a}^{\mathrm{sech}}(x-z)) A(x) \Phi_{2a}^{\mathrm{sech}}(x-z) \\
&\leq (\partial_{x_j} \log p(x) + \partial_{x_j} \log A(x) + \partial_{x_j} \log \Phi_{2a}^{\mathrm{sech}}(x-z)) A(x) e^{\sqrt{\frac{\pi}{2}} a \|x\|_1} \Phi_a^{\mathrm{sech}}(z). \quad \text{(D.9.1)}
\end{aligned}$$

Assumption 5.H follows directly from the definition of ν. Assumption E.16 holds for $\|\cdot\| = \|\cdot\|_1$, $\overline{f_j}(R) = e^{-\sqrt{2\pi}aR}$, and $C_{f,j} = 2$ since

$$e^{-a|x_d|} \leq \operatorname{sech}(ax_d) \leq 2e^{-a|x_d|}.$$

Next, we verify that $\mathcal{Q}^{\operatorname{sech}}(C) \subseteq \mathcal{Q}(b, C_{B,j}, C_{\mathcal{O},j})$. Take any $b > 0$ as fixed. It follows from Eq. (D.9.1) that

$$B_j(x,z) = (\partial_{x_j} \log p(x) + \partial_{x_j} \log A(x) + \partial_{x_j} \log \Phi_{2a}^{\operatorname{sech}}(x-z))A(x)e^{\sqrt{\frac{\pi}{2}}a\|x\|_1}.$$

We see that

$$\begin{aligned}
\partial_{x_j} \log \Phi_{2a}^{\operatorname{sech}}(x-z) &= \sqrt{2\pi}\, a \tanh(\sqrt{2\pi}\, a(x_j - z_j)) + \textstyle\sum_{d\neq j}^{D} \log \operatorname{sech}(\sqrt{2\pi}\, a(x_d - z_d)) \\
&\leq (\sqrt{2\pi}\, a)(1 + \textstyle\sum_{d\neq j}^{D} x_d + z_d) \\
&\leq (\sqrt{2\pi}\, a)(1 + \|x\|_1 + \|z\|_1). \qquad \text{(D.9.2)}
\end{aligned}$$

The hypotheses concerning $\nabla \log p$ and A, together with Eq. (D.9.2), imply for some constant $c' > 0$,

$$B_j(x,z) \leq c'(1 + \|x\|_1 + \|z\|_1)A(x)e^{\sqrt{\frac{\pi}{2}}a\|x\|_1}.$$

Hence from some $c'', c''' \geq 0$, $(Q_N B_j)(z) \leq c''(1 + \|z\|_1) \leq c''' f_j(z)^{-b}$, verifying Assumption E.17. In addition, for some constant c'''',

$$\begin{aligned}
(1+\omega_j)^{-1}|Q_N(\mathcal{O}_{j,x}e^{i\omega\cdot x})| &= (1+\omega_j)^{-1}|Q_N((\partial_{x_j}\log p(x) + \partial_{x_j}\log A(x) - i\omega_j)A(x)e^{i\omega\cdot x})| \\
&\leq Q_N(c''''(1+\|x\|_1)A(x)) \\
&\leq c''''C,
\end{aligned}$$

verifying Assumption E.18. Since $\rho = 1$, the equality of MMD and GMMD follows from Eq. (5.3).

D.10 Proof of Theorem 5.4.8: IMQ fGMMD properties

We verify each of the assumptions in turn. By construction Assumption 5.F is satisfied with $\mathcal{O}_{j,x} = \mathcal{T}_{j,x}$.

By Wendland [143, Theorem 8.15], $\Phi_{c,\beta}$ has generalized Fourier transform

$$\widehat{\Phi_{c,\beta}}(\omega) = \frac{2^{1+\beta}}{\Gamma(-\beta)}\left(\frac{\|\omega\|_2}{c}\right)^{-\beta-D/2} K_{\beta+D/2}(c\|\omega\|_2),$$

where $K_v(z)$ is the modified Bessel function of the third kind. We write $a(\ell) \dot{\sim} b(\ell)$ to denote asymptotic equivalence up to a constant: $\lim_\ell a(\ell)/b(\ell) = c$ for some $c \in$

$(0, \infty)$. Asymptotically [1, eq. 10.25.3],

$$\begin{aligned}
&\hat{\Psi}^{\mathrm{IMQ}}_{c,\beta}(\omega) \dot{\sim} \|\omega\|_2^{-\beta-D/2-1/2} e^{-c\|\omega\|_2}, && \|\omega\|_2 \to \infty \quad \text{and} \\
&\hat{\Psi}^{\mathrm{IMQ}}_{c,\beta}(\omega) \dot{\sim} \|\omega\|_2^{-(\beta+D/2)-|\beta+D/2|} = \|\omega\|_2^{-(2\beta+D)_+} && \|\omega\|_2 \to 0.
\end{aligned}$$

Note that we have $\hat{\Phi}_j = \hat{\Psi}^{\mathrm{IMQ}}_{c,\beta}$. Clearly Assumption E.14 holds. To verify Assumption 5.G, note that

$$\begin{aligned}
\hat{\Psi}^{\mathrm{IMQ}}_{c',\beta'}/(\hat{\Psi}^{\mathrm{IMQ}}_{c,\beta})^{\lambda/2} &\sim \|\omega\|_2^{-(\beta'+D/2-1/2)+(\beta+D/2-1/2)\lambda/2} e^{(-c'+c\lambda/2)\|\omega\|_2}, && \|\omega\|_2 \to \infty \quad \text{and} \\
&\sim \|\omega\|_2^{\lambda(2\beta+D)_+/2-(2\beta'+D)_+} = \|\omega\|_2^{\lambda(2\beta+D)/2} && \|\omega\|_2 \to 0,
\end{aligned}$$

so $\hat{\Psi}^{\mathrm{IMQ}}_{c',\beta'}/(\hat{\Psi}^{\mathrm{IMQ}}_{c,\beta})^{\lambda/2} \in L^2$ as long as $c' = c\overline{\lambda}/2 > c\lambda/2$ and $\lambda(2\beta + D) > -D$. The first condition holds by construction and second condition is always satisfied, since $2\beta + D \geq 0 > -D$.

Assumption E.15 holds since

$$\begin{aligned}
\mathcal{O}_{j,x} f_j(x - z) &= \mathcal{T}_{j,x} \Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z) \\
&= ((\partial_{x_j} \log p(x) + \partial_{x_j} \log \Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z)) \Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z) \\
&\leq ((\partial_{x_j} \log p(x) + \partial_{x_j} \log \Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z)) \frac{\Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z)}{\Psi^{\mathrm{IMQ}}_{c',\beta'}(z)} \Psi^{\mathrm{IMQ}}_{c',\beta'}(z). \quad \text{(D.10.1)}
\end{aligned}$$

Assumption 5.H follows directly from the definition of ν. Assumption E.16 trivially holds for $\|\cdot\| = \|\cdot\|_2$, $\overline{f_j}(R) = ((c')^2 + R^2)^{\beta'}$ and $C_{f,j} = 1$.

Next, we verify that $\mathcal{Q}^{\mathrm{IMQ}}(C) \subseteq \mathcal{Q}(b, C_{B,j}, C_{\mathcal{O},j})$. It follows from Eq. (D.10.1) that

$$\begin{aligned}
B_j(x, z) &= (\partial_{x_j} \log p(x) + \partial_{x_j} \log \Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z)) \frac{\Psi^{\mathrm{IMQ}}_{c',\beta'}(x - z)}{\Psi^{\mathrm{IMQ}}_{c',\beta'}(z)} \\
&\leq \left(C_1 + C_2 \|x\|_2 - \frac{2\beta' |x_j - z_j|}{(c')^2 + \|x - z\|_2^2} \right) \left(2 \frac{(c')^2 + \|x - z\|_2^2 + \|x\|_2^2}{(c')^2 + \|z\|_2^2} \right)^{-\beta} \\
&\leq 2(C_1 + C_2 \|x\|_2 - 2\beta')(1 + \|x\|_2 / c')^{-2\beta}.
\end{aligned}$$

Hence for some $c'' > 0$, $(Q_N B_j)(z) \leq c''$, verifying Assumption E.17 for $b = 0$. In addition,

$$\begin{aligned}
(1 + \omega_j)^{-1} |Q_N(\mathcal{O}_{j,x} e^{i\omega \cdot x})| &= (1 + \omega_j)^{-1} |Q_N((\partial_{x_j} \log p(x) - i\omega_j) e^{-i\omega \cdot x})| \\
&\leq Q_N(C_1 + C_2 \|x\|_2) + 1 \\
&\leq C_1 + C_2 C + 1,
\end{aligned}$$

verifying Assumption E.18.

Finally, we verify that $\rho = \hat{\Psi}^{\mathrm{IMQ}}_{c,\beta}/(\hat{\Psi}^{\mathrm{IMQ}}_{c',\beta'})^2 \in L^t$, where $t = r/(2 - r) = -D/(D +$

$4\beta'\underline{\xi}$). In fact,

$$\hat{\Psi}^{\mathrm{IMQ}}_{c,\beta}/(\hat{\Psi}^{\mathrm{IMQ}}_{c',\beta'})^2 \dot{\sim} \|\omega\|_2^{-2(\beta+D/2-1/2)/2+2(\beta'+D/2-1/2))} e^{2(-c/2+c')\|\omega\|_2}, \quad \|\omega\|_2 \to \infty \quad \text{and}$$
$$\dot{\sim} \|\omega\|_2^{2(2\beta'+D)_+-(2\beta+D)_+} = \|\omega\|_2^{-(2\beta+D)} \qquad \|\omega\|_2 \to 0,$$

so $\rho \in L^t$ whenever $c/2 > c'$ and

$$\frac{D}{(D+4\beta'\underline{\xi})}(2\beta+D) > -D \Leftrightarrow -\beta/(2\underline{\xi}) - D/(2\underline{\xi}) > \beta'.$$

Both these conditions hold by construction. Thus, the MMD lower bound on the GMMD follows from Eq. (5.3).

D.11 Proofs of Theorems 5.4.11 and 5.4.12: Asymptotics of fGMMD

The proofs of Theorems 5.4.11 and 5.4.12 rely on the following asymptotic result.

Theorem D.11.1. *Let $\xi_{ij} : \mathbb{R}^d \to \mathbb{R}, i = 1, \ldots, I, j = 1, \ldots, J$, be a collection of functions and let $X_n \overset{i.i.d.}{\sim} Q$, where Q is absolutely continuous with respect to Lebesgue measure. Define the random variables $\xi_{nij} := \xi_{ij}(X_n)$ and, for $r, s \geq 1$, the random variable*

$$T_{r,s,N} := \left(\sum_{i=1}^{I} \left(\sum_{j=1}^{J} \left| N^{-1} \sum_{n=1}^{N} \xi_{nij} \right|^r \right)^{s/r} \right)^{2/s}. \tag{D.11.1}$$

Assume that for all $i \in [I]$ and $j \in [J]$, ξ_{1ij} has a finite second moment. Then the following statements hold.

1. *If $\mu_{ij} := Q(\xi_{sj}) = 0$ for all i and j, then*

$$NT_{r,s,N} \overset{\mathcal{D}}{\Longrightarrow} \left(\sum_{i=1}^{I} \left(\sum_{j=1}^{J} |\zeta_{ij}|^r \right)^{s/r} \right)^{2/s} \text{ as } N \to \infty,$$

   where $\zeta \sim \mathcal{N}(0, \Sigma)$ and $\Sigma_{ij,i'j'} := \mathrm{Cov}(\xi_{1ij}, \xi_{1i'j'})$.

2. *If $\mu_{ij} \neq 0$ for some i and j, then*

$$NT_{r,s,N} \overset{a.s.}{\to} \infty \text{ as } N \to \infty.$$

Proof Let $V_{N,ij} = N^{-1/2} \sum_{n=1}^{N} \xi_{nij}$. Since the ξ_{1ij} have finite second moments, $\|\Sigma\| < \infty$. Hence, by the multivariate CLT,

$$V_N - N^{1/2}\mu \overset{\mathcal{D}}{\Longrightarrow} \mathcal{N}(0, \Sigma).$$

Observe that $NT_{r,s,N} = (\sum_{i=1}^{I}(\sum_{j=1}^{J}|V_{N,ij}|^r)^{s/r})^{2/s}$. Hence if $\mu = 0$, Eq. (D.11.1) follows from the continuous mapping theorem.

Assume $\mu_{ij} \neq 0$ for some i and j. By the strong law of large numbers, $N^{-1/2}V_N \xrightarrow{a.s.} \mu$. Together with the continuous mapping theorem conclude that $T_{p,N} \xrightarrow{a.s.} c$ for some $c > 0$. Hence $NT_{p,N} \xrightarrow{a.s.} \infty$. □

When $r = s = 2$, the fGMMD is a degenerate V-statistic, and we recover its well-known distribution [125, Sec. 6.4, Thm. B] as a corollary. A similar result was used in Jitkrittum et al. [75] to construct the asymptotic null for the FSSD, which is degenerate U-statistic.

Corollary D.11.2. *Under the hypotheses of Theorem D.11.1,*

$$NT_{2,2,N} \xRightarrow{\mathcal{D}} \sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{ij}\omega_{ij}^2 \text{ as } N \to \infty,$$

where $\lambda = \text{eigs}(\Sigma)$ *and* $\omega_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

To apply these results to fGMMDs, take $s = 2$ and apply Theorem D.11.1 with $\xi_{mj} = \xi_{r,mj}$. We can apply Theorem D.11.1. Since $P(\xi_{r,mj}) = 0$ for all $m \in [M]$ and $j \in [J]$, under $H_0 : Q = P$ part 1 of Theorem D.11.1 holds. Furthermore, when $Q \neq P$, there exists some m and j for which $Q(\xi_{r,mj}) \neq 0$. Thus, under $H_1 : Q \neq P$ part 2 of Theorem D.11.1 holds.

The proof of Theorem 5.4.12 is essentially identical to that of Jitkrittum et al. [75, Theorem 3].

D.12 Hyperbolic Secant Properties

Recall that the hyperbolic secant function is given by $\text{sech}(a) = \frac{2}{e^a + e^{-a}}$. For $x \in \mathbb{R}^d$, define the hyperbolic secant kernel

$$\Phi_a^{\text{sech}}(x) := \text{sech}\left(\sqrt{\frac{\pi}{2}}ax\right) := \prod_{i=1}^{d}\text{sech}\left(\sqrt{\frac{\pi}{2}}ax_i\right).$$

It is a standard result that

$$\hat{\Phi}_a^{\text{sech}}(\omega) = \Phi_{1/a}^{\text{sech}}(\omega). \tag{D.12.1}$$

We can relate $\Phi_a^{\text{sech}}(x)^\xi$ to $\Phi_{a\xi}^{\text{sech}}(x)$, but to do so we will need the following standard result:

Lemma D.12.1. *For* $a, b \geq 0$ *and* $\xi \in (0, 1]$,

$$\frac{a^\xi + b^\xi}{2^{1-\xi}} \leq (a+b)^\xi \leq a^\xi + b^\xi.$$

Proof The lower bound follows from an application of Jensen's inequality and the upper bound follows from the concavity of $a \mapsto a^\xi$. □

Proposition D.12.2. *For $\xi \in (0, 1]$,*

$$\Phi_a^{\text{sech}}(x)^\xi \le \Phi_a^{\text{sech}}(\xi x) = \Phi_{a\xi}^{\text{sech}}(x) \le 2^{d(1-\xi)} \Phi_a^{\text{sech}}(x)^\xi$$
$$2^{-d(1-\xi)} \hat{\Phi}_{a/\xi}^{\text{sech}}(x) \le \hat{\Phi}_a^{\text{sech}}(x)^\xi \le \hat{\Phi}_{a/\xi}^{\text{sech}}(x).$$

Thus, $\Phi_{a/\xi}^{\text{sech}}$ is equivalent to $(\Phi_a^{\text{sech}})^{(\xi)}$.

Proof Apply Lemma D.12.1 and Eq. (D.12.1). □

Proposition D.12.3. *For all $x, y \in \mathbb{R}^d$ and $a > 0$,*

$$\Phi_a^{\text{sech}}(x - z) \le e^{\sqrt{\frac{\pi}{2}} a \|x\|_1} \Phi_a^{\text{sech}}(z).$$

Proof Take $d = 1$ since the general case follows immediately. Without loss of generality assume that $x \ge 0$ and let $a' = \sqrt{\frac{\pi}{2}} a$. Then

$$\frac{\Phi_a^{\text{sech}}(x - z)}{\Phi_a^{\text{sech}}(z)} = \frac{e^{a'z} + e^{-a'z}}{e^{a'(x-z)} + e^{-a'(x-z)}} = \frac{e^{a'z} + e^{-a'z}}{e^{-a'z} + e^{2a'x} e^{a'z}} e^{a'x} \le e^{a'x}.$$

□

D.13 Concentration Inequalities

Theorem D.13.1 (Chung and Lu [33, Theorem 2.9])**.** *Let $X_1, \dots, X_m$ be independent random variables satisfying $X_i > -A$ for all $i = 1, \dots, m$. Let $X := \sum_{i=1}^m X_i$ and $\overline{X^2} := \sum_{i=1}^m \mathbb{E}[X_i^2]$. Then for all $t > 0$,*

$$\mathbb{P}(X \le \mathbb{E}[X] - t) \le e^{-\frac{1}{2} t^2 / (\overline{X^2} + At/3)}.$$

Let $\hat{X} := \frac{1}{m} \sum_{i=1}^m X_i$.

Corollary D.13.2. *Let $X_1, \dots, X_m$ be i.i.d. nonnegative random variables with mean $\bar{X} := \mathbb{E}[X_1]$. Assume there exists $c > 0$ and $\gamma \in [0, 2]$ such that $\mathbb{E}[X_1^2] \le c \bar{X}^{2-\gamma}$. If, for $\delta \in (0, 1)$ and $\varepsilon \in (0, 1)$,*

$$m \ge \frac{2c \log(1/\delta)}{\varepsilon^2} \bar{X}^{-\gamma},$$

then with probability at least $1 - \delta$, $\hat{X} \ge (1 - \varepsilon)\bar{X}$.

Proof Applying Theorem D.13.1 with $t = m\varepsilon\bar{X}$ and $A = 0$ yields

$$\mathbb{P}(\hat{X} \leq \varepsilon\bar{X}) \leq e^{-\frac{1}{2c}\varepsilon^2 m\bar{X}^\gamma}.$$

Upper bounding the right hand side by δ and solving for m yields the result. □

Corollary D.13.3. *Let $X_1, \ldots, X_m$ be i.i.d. nonnegative random variables with mean $\bar{X} := \mathbb{E}[X_1]$. Assume there exists $c > 0$ and $\gamma \in [0, 2]$ such that $\mathbb{E}[X_1^2] \leq c\bar{X}^{2-\gamma}$. Let $\epsilon' = |X^* - \bar{X}|$ and assume $\epsilon' \leq \eta X^*$ for some $\eta \in (0, 1)$. If, for $\delta \in (0, 1)$,*

$$m \geq \frac{2c\log(1/\delta)}{\varepsilon^2}\bar{X}^{-\gamma},$$

then with probability at least $1 - \delta$, $\hat{X} \geq (1 - \varepsilon)X^*$. In particular, if $\epsilon' \leq \frac{\sigma X^*}{\sqrt{n}}$ and $X^* \geq \frac{\sigma^2}{\eta^2 n}$, then with probability at least $1 - \delta$, $\hat{X} \geq (1 - \varepsilon)X^*$ as long as

$$m \geq \frac{2c(1-\eta)^2\eta^{2\gamma}}{\varepsilon^2\sigma^{2\gamma}\log(1/\delta)}n^\gamma.$$

Proof Apply Corollary D.13.2 with $\frac{\varepsilon X^*}{\bar{X}}$ in place of ε. □

Example D.13.1. If we take $\gamma = 1/4$ and $\eta = \varepsilon = 1/2$, then we need $X^* \geq \frac{4\sigma^2}{n}$ and $m \geq \frac{\sqrt{2}c\log(1/\delta)}{\sigma^{1/2}}n^{1/4}$ to guarantee that $\hat{X} \geq \frac{1}{2}X^*$ with probability at least $1 - \delta$.

Bibliography

[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover Publications, 1964. 124

[2] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005. 25

[3] S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning*, 2012. 18, 19

[4] P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv.org*, June 2017. 19

[5] P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26:29–47, 2016. 18, 61

[6] E. Angelino, M. J. Johnson, and R. P. Adams. Patterns of scalable Bayesian inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016. 18

[7] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. 28

[8] R. Azaïs, J.-B. Bardet, A. Génadot, N. Krell, and P.-A. Zitt. Piecewise deterministic Markov process — recent results. *ESAIM: Proceedings*, 44:276–290, Jan. 2014. 58, 59

[9] F. Bach. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, 18:1–38, May 2017. 63, 76

[10] O. Bachem, M. Lucic, and A. Krause. Coresets for Nonparametric Estimation—the Case of DP-Means. In *International Conference on Machine Learning*, 2015. 25

[11] O. Bachem, M. Lucic, S. H. Hassani, and A. Krause. Approximate k-means++ in sublinear time. In *AAAI Conference on Artificial Intelligence*, 2016. 25, 80

[12] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. 103

[13] A. D. Barbour. Stein's Method for Diffusion Approximations. *Probability theory and related fields*, 84:297–322, 1990. 55

[14] R. Bardenet, A. Doucet, and C. C. Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In *International Conference on Machine Learning*, pages 405–413, 2014. 18, 19, 33

[15] R. Bardenet, A. Doucet, and C. C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18:1–43, 2017. 18, 19, 33

[16] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt. Quantitative ergodicity for some switched dynamical systems. *Electronic Communications in Probability*, 17(0), 2012. 58, 59, 109

[17] M. J. Betancourt. The fundamental incompatibility of Hamiltonian Monte Carlo and data subsampling. In *International Conference on Machine Learning*, 2015. 18

[18] J. Bierkens and A. Duncan. Limit theorems for the zig-zag process. *Advances in Applied Probability*, 49:792–825, 2017. 58

[19] J. Bierkens and G. O. Roberts. A piecewise deterministic scaling limit of Lifted Metropolis-Hastings in the Curie-Weiss model. *The Annals of Applied Probability*, 2016.

[20] J. Bierkens, P. Fearnhead, and G. O. Roberts. The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *arXiv.org*, July 2016. 18, 58, 59

[21] F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations. *Journal of Functional Analysis*, 263(8): 2430–2457, Oct. 2012. 52

[22] A. Bouchard-Côté, S. J. Vollmer, and A. Doucet. The bouncy particle sampler: A non-reversible rejection-free Markov chain Monte Carlo method. *Journal of the American Statistical Association*, 2017. 18

[23] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013. 84, 95

[24] V. Braverman, D. Feldman, and H. Lang. New Framework for Online and Streaming Coreset Constructions. 2016. 80

[25] F.-X. Briol, C. J. Oates, J. Cockayne, W. Y. Chen, and M. A. Girolami. On the Sampling Problem for Kernel Quadrature. In *International Conference on Machine Learning*, 2017. 63, 76

[26] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, Dec. 2013. 18

[27] S. Bubeck, R. Eldan, and J. Lehec. Finite-Time Analysis of Projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*, July 2015. 56, 61, 111

[28] O. Butkovsky. Subgeometric rates of convergence of Markov processes in the Wasserstein metric. *The Annals of Applied Probability*, 24(2):526–552, Apr. 2014. 54

[29] T. Campbell, J. Straub, J. W. Fisher, III, and J. P. How. Streaming, distributed variational inference for Bayesian nonparametrics. In *Advances in Neural Information Processing Systems*, 2015. 18

[30] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010. 67

[31] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv.org*, July 2017. 19

[32] H. M. Choi and J. P. Hobert. The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064, 2013. 19

[33] F. Chung and L. Lu. *Complex Graphs and Networks*, volume 107. American Mathematical Society, Providence, Rhode Island, 2006. 127

[34] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems*, 2015. 63, 66, 76

[35] K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *International Conference on Machine Learning*, 2016. 63, 65, 76

[36] O. L. V. Costa and F. Dufour. Stability and Ergodicity of Piecewise Deterministic Markov Processes. *SIAM Journal on Control and Optimization*, 47(2): 1053–1077, Jan. 2008. 58

[37] M. F. Cusumano-Towner and V. K. Mansinghka. Quantifying the probable approximation error of probabilistic inference programs. *arXiv.org*, May 2016. 19

[38] B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. Balcan, and L. Song. Scalable Kernel Methods via Doubly Stochastic Gradients. *arXiv.org*, July 2014. 63

[39] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017. 56, 111

[40] K.-D. Dang, M. Quiroz, R. Kohn, M.-N. Tran, and M. Villani. Hamiltonian Monte Carlo with Energy Conserving Subsampling. *arXiv.org*, Aug. 2017. 18

[41] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1984. 58

[42] A. Durmus and E. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin Algorithm. *arXiv.org*, 2015. 19

[43] A. Durmus and E. Moulines. Sampling from a strongly log-concave distribution with the Unadjusted Langevin Algorithm. *HAL*, pages 1–25, Apr. 2016. 19, 56, 111, 112, 113

[44] A. Durmus and E. Moulines. Supplement to "Sampling from a strongly log-concave distribution with the Unadjusted Langevin Algorithm". *HAL*, pages 1–17, Apr. 2016. 112

[45] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! *arXiv.org*, page arXiv:1801.02309, Jan. 2018. 19

[46] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability theory and related fields*, pages 1–36, Oct. 2015. 52, 60, 67, 96, 101

[47] R. Entezari, R. V. Craiu, and J. S. Rosenthal. Likelihood inflating sampling algorithm. *arXiv.org*, May 2016. 18

[48] S. N. Ethier and T. G. Kurtz. Markov processes: characterization and convergence, volume 282. John Wiley & Sons, 2009. 55, 102, 104

[49] P. Fearnhead, J. Bierkens, M. Pollock, and G. O. Roberts. Piecewise Deterministic Markov Processes for Continuous-Time Monte Carlo. *arXiv.org*, Nov. 2016. 18

[50] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Symposium on Theory of Computing*, June 2011. 25, 27, 79, 80

[51] D. Feldman, M. Faulkner, and A. Krause. Scalable training of mixture models via coresets. In *Advances in Neural Information Processing Systems*, pages 2142–2150, 2011. 25, 31

[52] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Symposium on Discrete Algorithms*, pages 1434–1453. SIAM, 2013. 25

[53] J. Fontbona, H. Guérin, and F. Malrieu. Quantitative estimates for the long-time behavior of an ergodic variant of the telegraph process. *Advances in Applied Probability*, 44:977–994, 2012. 58, 59

[54] R. Ge, H. Lee, and A. Risteski. Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo. *arXiv.org*, Oct. 2017. 19

[55] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992. 19

[56] A. Gelman and K. Shirley. Inference from Simulations and Monitoring Convergence. In *Handbook of Markov Chain Monte Carlo*, pages 163–174. Chapman and Hall/CRC, 2011. 19

[57] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, Dec. 2008. 27, 28

[58] A. Gelman, A. Vehtari, P. Jylänki, T. Sivula, D. Tran, S. Sahai, P. Blomstedt, J. P. Cunningham, D. Schiminovich, and C. Robert. Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *arXiv.org*, Dec. 2014. 18

[59] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. 27

[60] J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In *Advances in Neural Information Processing Systems*, 2015. 20, 75, 76

[61] J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In *AISTATS*, 2017. 20, 63, 65, 67, 75, 76, 115, 116

[62] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *arXiv.org*, Nov. 2016. 52, 61, 67, 76, 96, 97, 101, 115

[63] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, Mar. 2012. 63, 64, 68, 76

[64] R. B. Grosse, S. Ancha, and D. M. Roy. Measuring the reliability of MCMC inference with bidirectional Monte Carlo. In *Advances in Neural Information Processing Systems*, 2016. 19

[65] F. Guo, X. Wang, K. Fan, T. Broderick, and D. Dunson. Boosting Variational Inference. In *NIPS Workshop on Advances in Approximate Bayesian Inference*, 2016. 18

[66] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, pages 223–242, 2001. 33, 35

[67] M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1-2):223–259, Oct. 2009. 52, 59

[68] L. Han, T. Yang, and T. Zhang. Local uncertainty sampling for large-scale multi-class logistic regression. *arXiv.org*, Apr. 2016. 25, 31

[69] L. Hasenclever, S. Webb, T. Lienart, S. Vollmer, B. Lakshminarayanan, C. Blundell, and Y. W. Teh. Distributed Bayesian learning with stochastic natural-gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18:1–37, 2017. 18

[70] R. Herb and P. Sally Jr. The Plancherel formula, the Plancherel theorem, and the Fourier transform of orbital integrals. In *Representation Theory and Mathematical Physics: Conference in Honor of Gregg Zuckerman's 60th Birthday, October 24–27, 2009, Yale University*, volume 557, page 1. American Mathematical Soc., 2011. 117

[71] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013. 18, 33

[72] J. Honorio and Y.-J. Li. The Error Probability of Random Fourier Features is Dimensionality Independent. *arXiv.org*, Oct. 2017. 63

[73] J. H. Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2017. 97

[74] T. Jaakkola and M. I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, volume 82, 1997. 37

[75] W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A Linear-Time Kernel Goodness-of-Fit Test. In *Advances in Neural Information Processing Systems*, 2017. 63, 67, 73, 76, 126

[76] J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. Approximations of Markov Chains and Bayesian Inference. *arXiv.org*, stat.CO:1–53, Jan. 2016. 61

[77] G. L. Jones and J. P. Hobert. Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, pages 312–334, 2001. 19

[78] G. L. Jones and J. P. Hobert. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics*, 32(2):784–817, 2004. 19

[79] M. J. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994. 82

[80] K. Khare and J. P. Hobert. Geometric ergodicity of the Bayesian lasso. *Electronic Journal of Statistics*, 7(0):2150–2163, 2013. 19

[81] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, 2014. 18, 19, 33

[82] A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic variational inference in Stan. In *Advances in Neural Information Processing Systems*, June 2015. 37

[83] P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *SIGKDD Conference on Knowledge Discovery and Data Mining*, 2006. 46, 47

[84] S. Li. Concise Formulas for the Area and Volume of a Hyperspherical Cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2011. 84

[85] Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, 2016. 63, 76

[86] Q. Liu, J. D. Lee, and M. I. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. In *International Conference on Machine Learning*, 2016. 63, 65, 76

[87] M. Lucic, O. Bachem, and A. Krause. Strong Coresets for Hard and Soft Bregman Clustering with Applications to Exponential Family Mixtures. In *International Conference on Artificial Intelligence and Statistics*, 2016. 25

[88] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21(56): 1–14, 2016. 103

[89] D. Maclaurin and R. P. Adams. Firefly Monte Carlo: Exact MCMC with subsets of data. In *Uncertainty in Artificial Intelligence*, 2014. 18

[90] D. Madigan, N. Raghavan, W. Dumouchel, M. Nason, C. Posse, and G. Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, 6(2):173–190, 2002. 25

[91] F. Maire, N. Friel, and P. Alquier. Informed Sub-Sampling MCMC: Approximate Bayesian Inference for Large Datasets. *arXiv.org*, June 2017. 18, 19

[92] O. Mangoubi and A. Smith. Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions. *arXiv.org*, Aug. 2017. 19

[93] J. C. Mason and D. C. Handscomb. *Chebyshev Polynomials*. Chapman and Hall/CRC, New York, 2003. 41, 89

[94] A. C. Miller, N. Foti, and R. P. Adams. Variational Boosting: Iteratively Refining Posterior Approximations. In *International Conference on Machine Learning*, 2017. 18

[95] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, 2001. 18

[96] S. Minsker, S. Srivastava, L. Lin, and D. Dunson. Robust and Scalable Bayes via a Median of Subset Posterior Measures. *Journal of Machine Learning Research*, 18(124):1–40, Nov. 2017. 18

[97] A. Mira and C. J. Geyer. On non-reversible Markov chains. *Monte Carlo Methods, Fields Institute/AMS*, pages 95–110, 2000. 58

[98] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. 27

[99] P. Monmarché. On $\mathcal{H}^1$ and entropic convergence for contractive PDMP. *Electronic Journal of Probability*, 20:1–30, 2015. 58, 59

[100] A. Müller. Integral probability metrics and their generating classes of functions. *Ann. Appl. Probab.*, 29(2):pp. 429–443, 1997. 64

[101] C. Musco and C. Musco. Recursive Sampling for the Nyström Method . In *Advances in Neural Information Processing Systems*, 2017. 63

[102] R. M. Neal. Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. Technical Report 0406, University of Toronto, 2004. 58

[103] R. M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, pages 113–162. Chapman and Hall/CRC, 2011. 49, 51

[104] R. Nishihara, P. Moritz, S. Wang, A. Tumanov, W. Paul, J. Schleier-Smith, R. Liaw, M. Niknami, M. I. Jordan, and I. Stoica. Real-time machine learning: The missing pieces. In *Workshop on Hot Topics in Operating Systems*, 2017. 47

[105] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, stat.ME, 2017. 76

[106] A. Pakman, D. Gilboa, D. Carlson, and L. Paninski. Stochastic bouncy particle sampler. In *International Conference on Machine Learning*, 2017. 18

[107] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008. 19

[108] N. S. Pillai and A. Smith. Ergodicity of approximate MCMC chains with applications to large data sets. *arXiv.org*, May 2014. 18, 61

[109] M. Pollock, P. Fearnhead, A. M. Johansen, and G. O. Roberts. The scalable Langevin exact algorithm: Bayesian inference for big data. *arXiv.org*, Sept. 2016. 18

[110] Q. Qin and J. P. Hobert. Asymptotically Stable Drift and Minorization for Markov Chains with Application to Albert and Chib's Algorithm. *arXiv.org*, Dec. 2017. 19

[111] M. Quiroz, M. Villani, and R. Kohn. Exact Subsampling MCMC. *arXiv.org*, Mar. 2016. 18

[112] M. Rabinovich, E. Angelino, and M. I. Jordan. Variational consensus Monte Carlo. *arXiv.org*, June 2015. 18

[113] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007. 63, 66

[114] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*, pages 1313–1320, 2009. 47

[115] D. J. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, 2015. 18

[116] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001. 35

[117] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004. 19

[118] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, Nov. 1996. 19, 33, 35, 49, 56

[119] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011. 55

[120] D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, Feb. 2017. 61

[121] T. Salimans and D. A. Knowles. Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression. *Bayesian Analysis*, 8(4):837–882, Dec. 2013. 18

[122] T. Salimans, D. P. Kingma, and M. Welling. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In *International Conference on Machine Learning*, 2015. 18

[123] S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. In *Bayes 250*, 2013. 18, 31

[124] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013. 68

[125] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980. 126

[126] C. Simon and L. E. Blume. *Mathematics for Economists*. W. W. Norton & Company, 1994. 92

[127] C.-J. Simon-Gabriel and B. Schölkopf. Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions. *arXiv.org*, Apr. 2016. 67

[128] B. K. Sriperumbudur and Z. Szabó. Optimal rates for random fourier features. In *Advances in Neural Information Processing Systems*, pages 1144–1152, Cambridge, MA, USA, 2015. MIT Press. 63

[129] S. Srivastava, V. Cevher, Q. Tran-Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *International Conference on Artificial Intelligence and Statistics*, 2015. 18

[130] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602, 1972. 55

[131] M. Stephanou, M. Varughese, and I. Macdonald. Sequential quantiles via Hermite series density estimation. *Electronic Journal of Statistics*, 11(1):570–607, 2017. 37

[132] D. J. Sutherland and J. Schneider. On the error of random fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871, 2015. 63

[133] G. Szegö. *Orthogonal Polynomials*. American Mathematical Society, 4th edition, 1975. 40

[134] H. Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Math. J.*, 9:163–177, 1979. 102

[135] Y. W. Teh, A. H. Thiery, and S. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17(7):1–33, Mar. 2016. 18, 54, 61

[136] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81 (393):82–86, 1986. 37

[137] A. W. van der Vaart. *Asymptotic Statistics*. University of Cambridge, 1998. 42

[138] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17(159):1–48, 2016. 61

[139] M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1 (1–2):1–305, 2008. 18

[140] J. Wang. L^p-Wasserstein distance for stochastic differential equations driven by Lévy processes. *Bernoulli*, 22(3):1598–1616, Aug. 2016. 101

[141] J. Wang, J. D. Lee, M. Mahdavi, M. Kolar, and N. Srebro. Sketching Meets Random Projection in the Dual: A Provable Recovery Algorithm for Big and High-dimensional Data. In *International Conference on Artificial Intelligence and Statistics*, 2017. 19

[142] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2011. 18, 19, 75

[143] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, New York, NY, 2005. 117, 123

[144] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000. 63

[145] D. B. Woodard, S. C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640, Apr. 2009. 19

[146] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou. Nyström Method vs Random Fourier Features - A Theoretical and Empirical Comparison. In *Advances in Neural Information Processing Systems*, 2012. 63

[147] G. Zanella and G. O. Roberts. Analysis of the Gibbs Sampler for Gaussian hierarchical models via multigrid decomposition. *arXiv.org*, page arXiv:1703.06098, Mar. 2017. 19

[148] F. Zhang and C. Gao. Convergence Rates of Variational Posterior Distributions. *arXiv.org*, page arXiv:1712.02519, Dec. 2017. 19

[149] J. Zhao and D. Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015. 63