Approximations and scaling limits of Markov chains with applications to MCMC and approximate inference

by

Jeffrey Negrea

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Statistical Sciences
University of Toronto

Approximations and scaling limits of Markov chains with applications to MCMC and approximate inference

Jeffrey Negrea
Doctor of Philosophy

Department of Statistical Sciences
University of Toronto
2022

Abstract

Markov chains are an essential tool in computational statistics because they form the basis for efficient exact and approximate inference methods, especially in Bayesian statistics. This dissertation offers insight into the viability of approximate inference methods based on both approximations to the transition kernels of a Markov chain for exact methods, and on Markov chains derived from unadjusted stochastic gradient methods. This dissertation also demonstrates how to tune computational methods based upon Markov chains in order to optimize efficiency and accuracy for approximate and exact inference. Results are obtained via two key theoretical methods: (1) an analysis of the perturbation sensitivity of Markov chains using operator theory, and (2) through scaling limits of Markov chains that facilitate a comparison to idealized continuous-time processes. The primary contributions of this dissertation are: (i) a perturbation analysis of reversible geometrically ergodic Markov chains, which characterizes the stability of the stationary distribution and rate of convergence under changes in the transition dynamics; (ii) results on the geometry of probability densities, generalized distributional integration-by-parts, and their consequences; (iii) a joint characterization of the optimal proposal scaling and shaping for the random-walk Metropolis algorithm; and (iv) a complete characterization of the statistical asymptotics of stochastic gradient algorithms as methods for approximate inference, with recommendations on how to tune them for accuracy and efficiency.

To my partner, Snow Murdock; and to my parents, Yolanda Barna and Horia Negrea.

Acknowledgements

My Ph.D. work was supported by an NSERC Vanier Canada Graduate Scholarship, an Ontario Graduate Scholarship, an NSERC Michael Smith Foreign Study Supplement, and by stipends from the University of Toronto and the Vector Institute. I thank Daniel Roy and Jeffrey Rosenthal for their guidance and mentorship throughout my Ph.D. I thank Daniel Rudolf for very helpful comments on the first version of the preprint that eventually became chapter 1. I also thank Gareth O. Roberts, Peter Rosenthal, and Don Hadwin for helpful discussions regarding that chapter. I thank Mufan (Bill) Li and Michaël Lalencette for helpful discussions regarding chapters 2 and 3. I thank Jonathan Huggins and Jun Yang for their input on chapter 4.

My time as a Ph.D. candidate was positively impacted by interactions I had with faculty members and peers. In no particular order, other faculty members at the University of Toronto that contributed to my terrific experience here, and whom I would like to acknowledge, include Jamie Stafford, Nancy Reid, Lei Sun, Radu Craiu, Stanislav Volgushev, Patrick Brown, Zhou Zhou, Alison Gibbs, Qiang Sun, Dehan Kong, David Duvenaud, Murat Erdogdu, Keith Knight, Sebastian Jaimungal, Ting-Kam Leonard Wong, Rohan Alexander, and Daniel Simpson. Other peers and former peers I would like to acknowledge include, in no particular order, Blair Bilodeau, Mahdi Haghifam, Yanbo Tang, Alex Stringer, Yuxiang (Alex) Gao, Victor Veitch, Xuancheng (Bill) Huang, Robert Zimmerman, Ekansh Sharma, Arvind Shrivats, Phillipe Casgraine, Zachary Naulet, Ali Al-Aradi, Alex Edmonds, and Yasaman Mahdaviyeh.

I would also like to thank my partner, Snow Murdock; my parents, Yolanda Barna and Horia Negrea; and my mother-in-law, Helena Wong, for their support during my graduate studies.

Contents

List of Tables

List of Figures

Attribution

This dissertation incorporates three separate projects as individual chapters. For all three included projects I was the main contributing author. In the case of collaborative projects, I briefly discuss an assignment of credit. Chapter 1 corresponds to [83], which was co-authored with Jeffrey S. Rosenthal. I was the primary contributor to both the proofs and writing of that work. Rosenthal's contributions were the original high level idea of the project, which he had suggested I work on, and his supervisory role where he provided advise for the direction of the project, discussions and refinements of proof techniques, and advise on framing of the results. Chapters 2 and 3 correspond to [82], of which I am the sole author. Chapter 4 corresponds to [84], a collaborative project between myself, Jonathan H. Huggins, Jun Yang, Daniel M. Roy, and Haoyue Feng. Huggins is responsible for the original empirical discovery that stochastic gradient methods can be used for misspecification-robust inference that that work is based on. I proposed and developed the scaling limit analysis and the overall direction of that work as a basis upon which to explain Huggins' empirical findings, and the scope of the project grew based upon my results. Huggins, Yang, and Roy contributed to the framing, direction, and literature review of that work, and to planning of experiments, though I am responsible for all final aspects. Feng, Huggins' Ph.D. student, ran some preliminary experiments in the early stages of the project, but these do not appear in [84] or in this dissertation.

Introduction

Markov chains are an essential tool in computational statistics because they form the basis for efficient exact and approximate inference methods, especially in Bayesian statistics. This dissertation offers insight into the viability of approximate inference methods based on both approximations to the transition kernels of a Markov chain for exact methods, and on Markov chains derived from unadjusted stochastic gradient methods. This dissertation also demonstrates how to tune computational methods based upon Markov chains in order to optimize efficiency and accuracy for approximate and exact inference. Results are obtained via two key theoretical methods: (1) an analysis of the perturbation sensitivity of Markov chains using operator theory, and (2) through scaling limits of Markov chains that facilitate a comparison to idealized continuous-time processes. The primary contributions of this dissertation are: (i) a perturbation analysis of reversible geometrically ergodic Markov chains, which characterizes the stability of the stationary distribution and rate of convergence under changes in the transition dynamics; (ii) results on the geometry of probability densities, generalized distributional integration-by-parts, and their consequences; (iii) a joint characterization of the optimal proposal scaling and shaping for the random-walk Metropolis algorithm; and (iv) a complete characterization of the statistical asymptotics of stochastic gradient algorithms as methods for approximate inference, with recommendations on how to tune them for accuracy and efficiency.

Chapter 1 establishes a number of results on the stability of reversible geometrically ergodic Markov chains under perturbations of the transition kernel and applications of these results to approximate Markov chain Monte Carlo methods. The tools used to prove our results are based on the operator-theoretic properties of reversible Markov chains. The results provide a theoretical justification to the intuition that small changes in the transition kernel should not affect the performance of MCMC methods too dramatically. Furthermore, we find that the Markov chains that mix faster are also more robust to perturbations. The results also include quantitative bounds on the mean-squared error for Monte Carlo estimates derived from perturbed or approximate Markov chains, and results that allow one to measure the approximation error due to perturbation under several different important metrics on the space of probability distributions. The results are applied to analyze several common approximations used in Bayesian inference.

Chapter 3 establishes a scaling limit for the random-walk Metropolis (RWM) algorithm that includes dependence between coordinates in order to characterize both the optimal RWM proposal scaling and the optimal RWM proposal covariance structure. This is done via a scaling limit for a block-independent target distribution, with in-block dependence, and between-block independence. We provide a variational characterization of the optimal proposal shaping matrix, and a formula for the optimal scaling. The optimal scaling, for any fixed shaping matrix, leads to an acceptance rate of ≈ 0.234 as in the seminal work of Roberts et al. [94]. We show that when the blocks are rotation-scalings of independent and identical components, that the optimal shaping problem can be solved explicitly, yielding the same recommendation as in Roberts and Rosenthal [97], to tune RWM so that the covariance of the proposals is proportional to the covariance of the target distribution. More generally, we show that their recommendation optimizes the instantaneous autocorrelation of linear functions in the the scaling limit, while a simple formula for true optimal solution is intractable. These results are informative to practitioners regarding how to tune the RWM algorithm to optimize sampling efficiency.

As a mathematical requirement to derive the results in Chapter 3, we developed several independently interesting results on the geometry of probability density functions and on their corresponding integration-by-parts formulae. As these results are of independent interest, they have been split off as their own chapter of this dissertation, Chapter 2. These results characterize regularity in the geometry of probability density functions that have bounded gradients on (almost) every level set, provide a general distributional integration by parts formula: $\mathbb{E}f(X)\nabla \log \pi(X) = -\mathbb{E}\nabla f(X)$ for $X \sim \pi$ that applies to as broad of a collection of densities π and test functions f as is possible, and is proven using geometric measure theory. We use the formula to derive several properties of densities such that $\nabla \log \pi$ is Lipschitz continuous, including the remarkable fact that if $\nabla \log \pi$ is L-Lipschitz and $X \sim \pi$ then $\nabla \log \pi(X)$ is sub-Gaussian with dimension-free sub-Gaussian constant, L. These results are of broader interest because these integration by parts formulae are widely useful, and the random variable $\nabla \log \pi(X)$ appears in the analysis of many algorithms used in computational statistics.

Chapter 4 establishes a scaling limit for stochastic gradient algorithms, and applies the scaling limit to understand how the tunings of these methods affect their utility for approximate Bayesian inference. We show how to tune these methods to match the statistical asymptotics of the posterior distribution, of the MLE, or other targets that adjust the posterior for model-misspecification such as the bagged posterior. Our scaling limit provides not only a limiting stationary distribution, but a full characterization of the paths of the process, making it stronger than existing results, and allowing us to provide a number of additional insights to auxiliary quantities such as mixing times and iterate averages. The theoretical results obtained via our scaling limit theory are supported by empirical results based on both real and simulated data. The results are useful to practitioners as they demonstrate how stochastic gradient methods can be implemented to achieve or interpreted in terms of various sampling desiderata, including very rapid approximate inference with tunable levels of adjustment for model misspecification.

Chapter 1

Perturbations of Geometrically Ergodic Reversible Markov Chains

1.1 Introduction

The use of Markov Chain Monte Carlo (MCMC) arises from the need to sample from probabilistic models when simple Monte Carlo is not possible. The procedure is to simulate a positive recurrent Markov process where the stationary distribution is the measure one intends to sample, so that the dynamics of the process converge to the distribution required. Temporally correlated samples may then be used to approximate various expectations; see e.g. Brooks et al. [18] and the many references therein. Examples of common applications may be found in hierarchical models, spatio-temporal models, random networks, finance, bioinformatics, etc.

Often, however, the transition dynamics of the Markov Chain required to run this process exactly are too computationally expensive due to prohibitively large datasets, intractable likelihoods, etc. In such cases it is tempting to instead *approximate* the transition dynamics of the Markov process in question, either deterministically as in the low-rank Gaussian approximation of Johndrow et al. [52], or stochastically as in the noisy Metropolis–Hastings procedure of Alquier et al. [2]. It is important then to understand whether these approximations will yield stable and reliable results. This chapter aims to provide quantitative tools for the analysis of these algorithms. Since the use of approximation for the transition dynamics may be interpreted as a *perturbation* of the transition kernel of the exact MCMC algorithm, we focus on bounds on the convergence of perturbations of Markov chains.

The primary purpose of this chapter is to extend existing quantitative bounds on the errors of approximate Markov chains from the uniformly ergodic case in [52] to the geometrically ergodic case (a weaker condition, for which multiple equivalent definitions may be found in Roberts and Rosenthal [96]). Our work will extend the theoretical results of [52] in the case that the exact chain is reversible, replacing the total variation metric with L_2 distances, and relaxing the uniform contraction condition to $L_2(\pi)$-geometric ergodicity.

1.1.1 Geometric Ergodicity

When analyzing the performance of exact MCMC algorithms, it is natural to decompose the error in approximation of expectations into a component for the transient phase error of the process and one for the Monte-Carlo approximation error. The former may be interpreted as the bias due to not having started the process in the stationary distribution. A Markov chain is *geometrically ergodic* if, from a suitable initial distribution ν, the marginal distribution of the n^{th} iterate of the chain converges to the stationary distribution, with an error that decays as $C(\nu)\rho^n$ for some $\rho \in (0,1)$ and some constant depending on the initial distribution $C(\nu)$, in some suitable metric on the space of probability measures. The geometric ergodicity condition essentially dictates that the transient phase error of the n^{th} sample decays exponentially quickly in n. The chain is *uniformly* (geometrically) ergodic if C can be chosen independently of the initial distribution. Geometric ergodicity is a desirable property as it ensures that cumulative transient phase error asymptotically does not dominate the Monte-Carlo error, while still being less restrictive than the uniform ergodicity condition, which often fails when the state space is not finite or compact (for example, an AR(1) process is geometrically ergodic but not uniformly ergodic).

When using approximate MCMC methods, one desires that the approximation preserves geometric ergodicity, so that convergence to stationarity is still fast and the transient phase error goes to zero quickly. This is an important issue, especially since Medina–Aguayo et al. [74] have shown that intuitive approximations such as Monte-Carlo within Metropolis may lead to transient approximating chains.

1.1.2 Outline of the Chapter

The outline of this chapter is as follows. Section 1.2 reviews related work. Then Section 1.3 contains our main theoretical results and their proofs. Theorem 1.1 therein provides bounds on the distance between stationary distributions, and gives a sufficient condition for the perturbed chain to be geometrically ergodic in $L_2(\pi)$, where π is the stationary distribution of the unperturbed chain. Theorem 1.2 and Theorem 1.3 give sufficient conditions for the perturbed chain to be geometrically ergodic according to several other variants of the definition of geometric ergodicity (for different metrics and families of initial distributions), and provide quantitative rates when possible. The remainder of Section 1.3 establishes bounds on autocorrelations, and mean-squared-error for Monte Carlo estimates of expected values computed with the perturbed chain.

Finally, Section 1.4 considers noisy and/or approximate Metropolis–Hastings algorithms. It provides sufficient conditions that one can check in order for our results from Section 1.3 to be applied. We use this to study Metropolis–Hastings with deterministic approximations to the target density, as well as the Monte Carlo within Metropolis algorithm, as in Medina–Aguayo et al. [73], and provide some examples of how these types of approximations might arise in practice.

1.2 Related Work

This section presents a brief review of related work, discussing convergence of perturbed Markov chains in the uniformly ergodic and geometrically ergodic cases with varying metrics and additional assumptions. The results in the literature have a wide range of assumptions required and a wide range of scopes for their various results. The results for uniformly ergodic chains have a simpler aesthetic, in line with what intuition for finite state space chains might inspire, as they do not require drift and minorization conditions to state. Our results cover the geometrically ergodic and reversible case, and use properties of reversibility to match the simpler aesthetic found in the literature for the uniformly ergodic case.

Close to the present work, Johndrow et al. [52] derive perturbation bounds to assess the robustness of approximate MCMC algorithms. The assumptions upon which their results rely are: the original chain is uniformly contractive in the total variation norm (this implies uniform ergodicity); and the perturbation is sufficiently small (in the operator norm induced by the total variation norm). The main results of their work are: the perturbed kernel is uniformly contractive in the total variation norm; the perturbed stationary distribution is close to the original stationary distribution in total variation; explicit bounds on the total variation distance between finite time approximate sampling distributions and the original stationary distribution; explicit bounds on total variation difference between the original stationary distribution and the mixture of finite time approximate sampling distributions; and explicit bounds on the MSE for integral approximation using approximate kernel and the true kernel. The results derived by [52] are applied within the same work to a wide variety of approximate MCMC problems including low rank approximation to Gaussian processes and sub-sampling approximations. In other work, Johndrow and Mattingly [50], use intuitive coupling arguments to establish similar results under the same uniform contractivity assumption.

Further results on perturbations for uniformly ergodic chains may be found in Mitrophanov [77]. This work is motivated in part by numerical rounding errors. Various applications of these results may be found in Alquier et al. [2]. The only assumption of [77] is that the original chain is uniformly ergodic. The work is unique in that it makes no assumption regarding the proximity of the original and perturbed kernel, though the level of approximation error does still scale linearly with the total variation distance of the original and perturbed kernels. The main results are: explicit bounds on the total variation distance between finite time sampling distributions; and explicit bounds on the total variation distance between stationary distributions.

The work of Roberts et al. [99] (see also Breyer et al. [17]) is also motivated by numerical rounding errors. The perturbed kernel is assumed to be derived from the original kernel by a *round-off function*, which e.g. maps the input to nearest multiple of 2^{-31}. In such cases, the new state space is at most countable while the old state space may have been uncountable and so the resulting chains have mutually singular marginal distributions at all finite times and mutually singular stationary distributions (if they have stationary distributions at all). The results of [99] require the analysis of Lyapunov drift conditions and drift functions (which we will avoid by working in an appropriate L_2 space). The key assumptions in [99] are: the original kernel is geometrically ergodic, and V is a Lyapunov drift function for the original kernel; the original and perturbed transition kernels are close in the V-norm; the perturbed kernel is defined via a round-off function with round-off error uniformly sufficiently small; and $\log V$ is uniformly continuous. The main results of the work include that: if the perturbed kernel is sufficiently close in the V-norm then geometric ergodicity is preserved; if the drift function, V, can be chosen so that $\log V$ is uniformly continuous and if the round-off errors can be made arbitrarily small then the kernels can be made arbitrarily close in the V-norm; explicit bounds on the total variation distance between the approximate finite-time sampling distribution and the true stationary distribution; and sufficient conditions for the approximating stationary distribution to be arbitrarily close in total variation to the true stationary distribution. They also prove results that do not require closeness in the V-norm, or even absolute continuity of the perturbed transitions; in such cases they show that a suitable drift condition on the original chain together with a uniformly small round-off error yields perturbed chains which are geometrically ergodic, and that the stationary measure varies continuously under such perturbations in the topology of weak convergence.

Pillai and Smith [89] provide bounds in terms of the Wasserstein topology (cf. Gibbs [36]). Their main focus is on approximate MCMC algorithms, especially approximation due to sub-sampling from a large dataset (e.g., when computing the posterior density). Their underlying assumptions are: the original and perturbed kernels satisfy a series of *drift-like conditions* with shared parameters; the original kernel has finite eccentricity for all states (where eccentricity of a state is defined as the expected distance between the state and a sample from the stationary distribution); the *Ricci curvature* of the original kernel has a non-trivial uniform lower bound on a positive measure subset of the state space; and the transition kernels are close in the Wasserstein metric, uniformly on the mentioned subset. Their main results under these assumptions are: explicit bounds on the Wasserstein distance between the approximate sampling distribution and the original stationary distribution; explicit bounds on the total variation distance of the original and perturbed stationary distributions and bounds on the mixing times of each chain; explicit bounds on the bias and L_1 error of Monte Carlo approximations; decomposition of the error from approximate MCMC estimation into components from *burn-in*, *asymptotic bias*, and *variance*; and rigorous discussion of the trade-off between the above error components.

Rudolf and Schweizer [104] also use the Wasserstein topology. They focus on approximate MCMC algorithms, with applications to auto-regressive processes and stochastic Langevin algorithms for Gibbs random fields. Their results use the following assumptions: the original kernel is Wasserstein ergodic; a Lyapunov drift condition for perturbed kernel is given, with drift function $\tilde{V}$; $\tilde{V}$ has finite expectation under the initial distribution; and the perturbation operator is uniformly bounded in a $\tilde{V}$-normalized Wasserstein norm. Their main results are: explicit bounds on the Wasserstein distance and weighted total variation distance between the original and perturbed finite time sampling distributions; and explicit bounds on the Wasserstein distance between stationary distributions.

Ferré et al. [35] build upon Keller and Liverani [56] to provide perturbation results for V-geometrically ergodic Markov chains using a simultaneous drift condition. They show that any perturbation to the transition kernel which shares its drift condition has a stationary distribution, is also V-geometrically ergodic, and that the perturbed stationary distributions is close to the original one. The assumption of a shared drift condition may be difficult to verify or not hold in some cases of interest related to approximate or noisy Markov chain Monte Carlo. Hervé and Ledoux [41] considers finite rank approximations to a transition kernel. That work gives sufficient conditions for approximations to inherit V-geometric ergodicity and provides a quantitative relationship between the rates of convergence and bounds the total variation distance between stationary measures. It also provides sufficient conditions for V-geometric ergodicity of a family of finite-rank approximations to a transition kernel to guarantee geometric ergodicity of the kernel, and provides a quantitative rates of convergence. In both of these results, as in [35], the results depend on a simultaneous drift condition for the approximations and the original kernel.

Each of the above works demonstrate bounds on various measures of error from using approximate finite-time sampling distributions and approximate ergodic distributions to calculate expectations of functions. On the other hand, the assumptions underlying the results vary dramatically. The results for uniformly ergodic chains are based on simpler and more intuitive assumptions than those for geometrically ergodic chains. Our work extends these results to geometrically ergodic chains and perturbations while preserving essentially the same level of simplicity in the assumptions. In particular we avoid the need to identify a Lyapunov drift condition, and our assumptions are expressed directly in terms of transition kernels, rather than a relationship between drift conditions which they satisfy.

1.3 Perturbation Bounds

This section extends the main results from Johndrow et al. [52] to the $L_2(\pi)$-geometrically ergodic case for, assuming the perturbation $P - P_\epsilon$ has bounded $L_2(\pi)$ operator norm.

1.3.1 Definitions and Notation

Let π be a probability measure on a measurable space $(\mathcal{X}, \Sigma)$. We make considerable use of the following norms on signed measures and their corresponding Banach spaces.

$$
\begin{aligned}
\|\lambda\|_{\mathrm{TV}} &= \sup_{A\in\Sigma} |\lambda(A)| & \mathcal{M}(\Sigma) &= \{\text{bounded signed measures on } (\mathcal{X},\Sigma)\} \\
\|\lambda\|_{L_2(\pi)} &= \left(\int \left(\frac{\mathrm{d}\lambda}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi\right)^{1/2} & L_2(\pi) &= \left\{\nu \ll \pi : \|\nu\|_{L_2(\pi)} < \infty\right\} \\
\|\cdot\|_{L_{2,0}(\pi)} &= \|\cdot\|_{L_2(\pi)} |_{L_{2,0}(\pi)} & L_{2,0}(\pi) &= \{\nu \in L_2(\pi) : \nu(\mathcal{X}) = 0\} \\
\|\lambda\|_{L_1(\pi)} &= \int \left|\frac{\mathrm{d}\lambda}{\mathrm{d}\pi}\right| \mathrm{d}\pi & L_1(\pi) &= \left\{\nu \ll \pi : \|\nu\|_{L_1(\pi)} < \infty\right\} \\
\|\lambda\|_{L_\infty(\pi)} &= \operatorname*{ess\,sup}_{X\sim\pi} \frac{\mathrm{d}\lambda}{\mathrm{d}\pi}(X) & L_\infty(\pi) &= \left\{\nu \ll \pi : (\exists b > 0)\left(\left|\frac{\mathrm{d}\nu}{\mathrm{d}\pi}\right| < b \quad \pi\text{-a.e.}\right)\right\}
\end{aligned}
$$

Note that $L_{2,0}(\pi)$ is a complete subspace of $L_2(\pi)$. Let

$$
\mathcal{M}_{+,1} = \{\lambda \in \mathcal{M} : [\forall A \in \Sigma \quad \lambda(A) \geq 0] \text{ and } [\lambda(\mathcal{X}) = 1]\}
$$

be the set of probability measures on $(\mathcal{X}, \Sigma)$. Note that for any probability measure, π, $L_\infty(\pi) \subset L_2(\pi) \subset L_1(\pi) \subset \mathcal{M}(\Sigma)$, though in general they are not complete subspaces of each other when their corresponding norms are not equivalent. For a norm, $\|\cdot\|$ on a vector space, we also write $\|\cdot\|$ the corresponding operator norm on the space of bounded linear operators from V to itself, $\mathcal{B}(V)$.

Definition 1.1 (Geometric Ergodicity)**.** *Let P be the kernel of a positive recurrent Markov chain with invariant measure π. Let λ be any measure with $\pi \ll \lambda$, and suppose that $\rho_{TV},\ \rho_1,\ \rho_2 \in (0,1)$. Then:*

*(i) P is π**-a.e.-TV geometrically ergodic with factor** ρ_{TV} if there exists $C_{TV} : \mathcal{X} \to \mathbb{R}_+$ such that for π-almost every $x \in \mathcal{X}$ and for all $n \in \mathbb{N}$:*

$$\|\delta_x P^n - \pi\|_{TV} \leq C_{TV}(x)\rho_{TV}^n \ .$$

The optimal rate *for π-a.e.-TV geometric ergodicity is the infimum over factors for which the above definition holds;*

$$\begin{aligned}\rho_{TV}^\star = \inf \{\rho > 0 \ s.t.\ \exists C : \mathcal{X} \to \mathbb{R}_+ \ with\ \pi(\{x : C(x) < \infty\}) = 1 \ and \\ \forall n \in \mathbb{N}, \pi\text{-}a.e.\ x \in \mathcal{X} \quad \|\delta_x P^n - \pi\|_{TV} \leq C(x)\rho^n\} \ .\end{aligned} \tag{1.1}$$

*(ii) P is $L_2(\lambda)$**-geometrically ergodic with factor** ρ_2 if $P : L_2(\lambda) \to L_2(\lambda)$ and there exists $C_2 : L_2(\lambda) \cap \mathcal{M}_{+,1} \to \mathbb{R}_+$ such that for every $\nu \in L_2(\lambda) \cap \mathcal{M}_{+,1}$ and for all $n \in \mathbb{N}$:*

$$\|\nu P^n - \pi\|_{L_2(\lambda)} \leq C_2(\nu)\rho_2^n \ .$$

The optimal rate *for $L_2(\lambda)$-geometric ergodicity is the infimum over factors for which the above definition holds;*

$$\begin{aligned}\rho_2^\star = \inf\Big\{\rho > 0 \ s.t.\ \exists C : L_2(\lambda) \cap \mathcal{M}_{+,1} \to \mathbb{R}_+ \ with \\ \forall n \in \mathbb{N}, \nu \in L_2(\lambda) \cap \mathcal{M}_{+,1} \quad \|\nu P^n - \pi\|_{L_2(\lambda)} \leq C(\nu)\rho^n\Big\} \ .\end{aligned}$$

Remark 1.1. *If P is π-reversible and aperiodic then P is $L_2(\pi)$-geometrically ergodic if and only if it is π-a.e. TV geometrically ergodic, as per Roberts and Rosenthal [96]. In this case the optimal rate of $L_2(\pi)$-geometric ergodicity, $\rho_2^\star$, is equal to the spectral radius of $P|_{L_{2,0}(\pi)}$,* *In this case, the spectrum of P is a subset of $[-\rho_2^\star, \rho_2^\star] \cup \{1\}$, and P is $L_2(\pi)$-geometrically ergodic with factor $\rho_2^\star$ and $C(\mu) = \|\mu - \pi\|_{L_2(\pi)}$. For more details see Proposition 1.2, and [96].* ◁

We abbreviate *geometric ergodicity* and *geometrically ergodic* as "GE" for brevity going forward.

1.3.2 Assumptions

We assume throughout that P is the transition kernel for a Markov chain on a countably generated state space $\mathcal{X}$ with σ-algebra Σ, which is reversible with respect to a stationary probability measure, π, and is π-irreducible and aperiodic. We call the Markov chain induced by P the "original" chain. The π-reversibility of P makes it natural to work in $L_2(\pi)$ since, in this case, P is a self-adjoint linear operator on a Hilbert space. This allows us access to the rich, elegant, and mature spectral theory of such operators. See for example [102, Chapter 12] and [28, Chapter 22]. We further assume that P is $L_2(\pi)$-geometrically ergodic with factor $0 < (1 - \alpha) < 1$. Equivalent definitions of $L_2(\pi)$-geometrically ergodic are given in Proposition 1.2. This assumption is weaker than the Doeblin condition used by [52], which implies uniform ergodicity.

Next, we assume that P_ϵ is a second, "perturbed" transition kernel, such that

$$\|P - P_\epsilon\|_{L_2(\pi)} \leq \epsilon$$

for some fixed $\epsilon > 0$, and that $P_\epsilon|_{L_2(\pi)} \in \mathcal{B}(L_2(\pi))$, i.e. that the perturbed transition kernel maps $L_2(\pi)$ measures to $L_2(\pi)$ measures. The norm condition quantifies the intuition that the perturbation is "small". We assume that P_ϵ is π-irreducible and aperiodic. We demonstrate (in Theorem 1.1) that under these assumptions P_ϵ has a unique stationary distribution, denoted by π_ϵ, with $\pi_\epsilon \in L_2(\pi)$.

Note that when $\mu \in L_1(\pi)$ we have $\|\mu - \pi\|_{\mathrm{TV}} = \frac{1}{2}\|\mu - \pi\|_{L_1(\pi)}$. On the other hand, $\|\cdot\|_{\mathrm{TV}}$ applies to all bounded measures, while $\|\cdot\|_{L_1(\pi)}$ applies only to the subspace of $L_1(\pi)$ measures. Note also that if $\pi \sim \pi_\epsilon$ (the two measures are mutually absolutely continuous), then $L_1(\pi)$ and $L_1(\pi_\epsilon)$ are equal as spaces and their norms are always equal, so in this case we need not distinguish between them.

To summarize, we assume that

Assumption 1.1 (Assumptions of Section 1.3.2)**.**

- *P is a Markov kernel that is*
  - *π-reversible for a prob. meas. π,*
  - *irreducible and aperiodic*
  - *$L_2(\pi)$-GE with factor $(1-\alpha)$,*
- *P_ϵ is a Markov kernel that is*
  - *irreducible and aperiodic,*
  - *$P_\epsilon : L_2(\pi) \to L_2(\pi)$, and*
  - *$\|P - P_\epsilon\|_{L_2(\pi)} < \epsilon$.*

The assumption that $P_\epsilon : L_2(\pi) \to L_2(\pi)$ and that $\|P_\epsilon\|_{L_2(\pi)} < \infty$ may seem difficult to verify. However, the following proposition shows us that it is satisfied for P_ϵ constructed based on the Metropolis–Hastings algorithm with suitable *jump kernels*. As long as the jump kernel, J, has $\|J\|_{L_2(\pi)} < \infty$ then it will be satisfied. Therefore, this assumption is not excessively restrictive for MCMC applications. The jump kernel, J, describes the conditional distribution of a new point in the chain proposed from x given that the proposal is accepted, and is related to the proposal kernel, Q, by $\alpha(x)J(x,A) = \int_A a(x,y)Q(x,dy)$ where $a(x,y)$ is the Metropolis–Hastings acceptance ratio and $\alpha(x) = \int_{\mathcal{X}} a(x,y)Q(x,dy)$ is the implied local jump-intensity.

Proposition 1.1. *If $P_\epsilon(x,\cdot) = (1-\alpha(x))\delta_x + \alpha(x)J(x,\cdot)$ with $\alpha : \mathcal{X} \to [0,1]$ measurable, and $J : L_2(\pi) \to L_2(\pi)$ and $\|J\|_{L_2(\pi)} < \infty$, then*

$$\|P_\epsilon\|_{L_2(\pi)} \leq 1 + \|J\|_{L_2(\pi)} . \tag{1.2}$$

Proof of Proposition 1.1. Consider the operator A on $L_2(\pi)$ given by the formula $[\nu A](C) = \int_C \alpha(x)\nu(dx)$ for all measurable sets C. Its adjoint, A', is given by the formula $[A'f](x) = \alpha(x)f(x)$ for all $x \in \mathcal{X}$ and $f \in L_2'(\pi)$. Since $\alpha : \mathcal{X} \to [0,1]$, then $A' : L_2'(\pi) \to L_2'(\pi)$ with $\|A'\|_{L_2'(\pi)} \leq 1$. Thus $A : L_2(\pi) \to L_2(\pi)$ with $\|A\|_{L_2(\pi)} \leq 1$. The same also holds for $I - A$. Now, $P_\epsilon = A + (I-A)J$, so $\|P_\epsilon\|_{L_2(\pi)} \leq 1 + \|J\|_{L_2(\pi)}$. □

Verifying that $\|P - P_\epsilon\|_{L_2(\pi)}$ is finite, and sufficiently small will be the main analytic burden faced when trying to apply our results to more general settings. The development of further tools to determine whether $\|P - P_\epsilon\|_{L_2(\pi)}$ is finite and to bound it quantitatively would be an interesting line of future research.

1.3.3 Convergence Rates and Closeness of Stationary Distributions

Theorem 1.1 (Geometric ergodicity of the perturbed chain and closeness of the stationary distributions in *original norm*, $L_2(\pi)$)**.** *Under the assumptions of Section 1.3.2, if in addition $\epsilon < \alpha$, then $\pi_\epsilon \in L_2(\pi)$,*

$$0 \leq \|\pi - \pi_\epsilon\|_{L_2(\pi)} \leq \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}} \ ,$$

P_ϵ is $L_2(\pi)$-geometrically ergodic with factor $1 - (\alpha - \epsilon)$, and for any initial probability measure $\mu \in L_2(\pi)$

$$\|\mu P_\epsilon^n - \pi\|_{L_2(\pi)} \leq (1 - (\alpha - \epsilon))^n \|\mu - \pi_\epsilon\|_{L_2(\pi)} + \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}} \ ,$$

The proof of this result is the content of Section 1.5.1. We follow the derivation in [52] with minimal structural modification, though the technicalities must be handled differently and additional theoretical machinery is required. We use the fact that the existence of a spectral gap for the restriction of P to $L_{2,0}(\pi)$ yields an inequality of the same form as uniform contractivity condition, but in the $L_2(\pi)$-norm as opposed to the total variation norm (cf. Theorem 2.1 of Roberts and Rosenthal [96]).

Remark 1.2. *Bounds on the differences between measures in $L_2(\pi)$-norm can be converted into bounds on the total variation distance since, by Cauchy-Schwarz, for any measure λ and any signed measure $\nu \in L_2(\lambda)$ we have $\|\nu\|_{TV} = \frac{1}{2}\|\nu\|_{L_1(\lambda)} \leq \frac{1}{2}\|\nu\|_{L_2(\lambda)}$. Thus, for example, under the assumptions of Theorem 1.1,*

$$\|\mu P_\epsilon^n - \pi\|_{TV} \leq \frac{1}{2}\left[(1 - (\alpha - \epsilon))^n \|\mu - \pi_\epsilon\|_{L_2(\pi)} + \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}}\right].$$

Similarly, under the assumptions of Theorem 1.1, we find that P_ϵ is $(L_2(\pi), \|\cdot\|_{TV})$-GE with factor $1 - (\alpha - \epsilon)$ (see Definition 1.2 below). ◁

In some situations, such as the computation of mean-squared errors in Theorem 1.5, it may be inconvenient or impossible to use to use the $L_2(\pi)$ norm when studying some aspects of P_ϵ. The next theorem will allow us to "switch" to other norms which may be more natural for a given task. First, however, we need to introduce one more notion of geometric ergodicity.

Definition 1.2 ($(V, \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \cdot \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert)$-Geometric Ergodicity)**.** *Let P be the kernel of a positive recurrent Markov chain with invariant measure π. Let V be a vector space of signed measures on $(\mathcal{X}, \Sigma)$ containing π, and let $\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \cdot \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert$ be a norm on V (for which V may not be complete).*

*P is $(V, \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \cdot \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert)$-**geometrically ergodic with factor** ρ if there exists $C : V \cap \mathcal{M}_{+,1} \to \mathbb{R}_+$ such that for every $\nu \in V \cap \mathcal{M}_{+,1}$ and for all $n \in \mathbb{N}$:*

$$\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \nu P^n - \pi \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \leq C(\nu)\rho^n \ .$$

The optimal rate *for $(V, \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \cdot \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert)$-geometric ergodicity is the infimum over factors for which the above definition holds;*

$$\rho^\star = \inf\left\{\rho > 0 : \exists C : V \cap \mathcal{M}_{+,1} \to \mathbb{R}_+ \ s.t. \ \forall n \in \mathbb{N}, \nu \in V \cap \mathcal{M}_{+,1} \quad \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \nu P^n - \pi \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \leq C(\nu)\rho^n\right\} \ .$$

We will be interested in this definition for the cases that $V = L_\infty(\pi)$ and $\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \cdot \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert$ is either $\|\cdot\|_{L_2(\pi)}$ or $\|\cdot\|_{L_1(\pi)}$.

Remark 1.3 (Relationships between $(L_\infty(\lambda), \|\cdot\|_{L_p(\lambda)})$-GE, a.e.-TV-GE, and $L_2(\lambda)$-GE)**.** *Clearly if P is $L_2(\lambda)$-GE with factor ρ_2 then it is also $(L_\infty(\lambda), \|\cdot\|_{L_2(\lambda)})$-GE with factor ρ_2. Conversely Roberts and Tweedie [100] show that if P is $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE with factor ρ_2 then it is also a.e.-TV-GE with some factor $\rho_{TV} \in (0, 1)$. However the factor for a.e.-TV-GE may in fact be worse than the factor of $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE or $(L_\infty(\pi), \|\cdot\|_{L_1(\pi)})$-GE. Baxendale [9] gives a detailed exposition on the barriers to the comparison of factors for geometric ergodicity given by different equivalent definitions.*

In Section 1.5.5 we give an example where the optimal rates for $L_2(\pi)$-GE and for $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE are distinct when P is not reversible. If P is π-reversible then the factors for $L_2(\pi)$-GE, $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE, and $(L_\infty(\pi), \|\cdot\|_{L_1(\pi)})$-GE must be the same.

This result combines a comment and Theorem 3 of [100], both stated but not proved. The formal statement of that result and its proof may be found in Section 1.5.6.

Finally, note that by definition $L_2(\pi)$-GE is equivalent to $(L_2(\pi), \|\cdot\|_{L_2(\pi)})$ with the same coefficient functions and factors, and that a.e.-TV-GE is equivalent to $(D, \|\cdot\|_{TV})$-GE where we can take $D = span(\{\pi\} \cup \{\delta_x : x \in \mathcal{X} \setminus N, r \in \mathbb{R}\})$ for some π-null set N. The null set, N, can be taken to be the same for all factors ρ by taking the union over the null sets for factors $\rho \in \mathbb{Q}$ (since a countable union of null sets is still null). $\triangleleft$

Lemma 1.1 (Characterization of optimal rates for $(V, |||\cdot|||)$-GE chains)**.** *If P is $(V, |||\cdot|||)$-GE with stationary measure π then the optimal rate for $(V, |||\cdot|||)$-GE is equal to*

$$\sup_{\mu \in V \cap \mathcal{M}_{+,1}} \limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n} \quad . \tag{1.3}$$

The proof of this result is found in Section 1.5.6.

Remark 1.4. *The quantity* $\limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n}$ *is the* local spectral radius *of $P - \Pi$ at μ with respect to $|||\cdot|||$, where Π is the rank-1 kernel defined by $\Pi(x, A) = \pi(A)$ for all $x \in \mathcal{X}$ and $A \in \Sigma$.* $\triangleleft$

Lemma 1.2 ($L_2(\pi)$-GE, $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE, and $(L_\infty(\pi), \|\cdot\|_{L_1(\pi)})$-GE are equivalent for π-reversible chains, with equal optimal rates.)**.** *Let $\rho \in [0, 1)$. The following are equivalent for a π-reversible Markov Chain P:*

(i) P is $(L_\infty(\pi), \|\cdot\|_{L_1(\pi)})$-geometrically ergodic with optimal rate ρ,

(ii) P is $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-geometrically ergodic with optimal rate ρ,

(iii) P is $L_2(\pi)$-geometrically ergodic with optimal rate ρ,

(iv) The spectral radius of $P|_{L_{2,0}(\pi)}$ is equal to ρ.

Remark 1.5. *Since either of (iii) or (iv) are equivalent to all the conditions listed in Roberts and Rosenthal [96, Theorem 2.1], indeed all of the items listed above are equivalent to all the items listed in their result. We only included (iii) and (iv) here for brevity, and since they are the ones most relevant to the present work. Moreover, all of these conditions are* *implied by any of the equivalent conditions for π-a.e.-TV-GE in Roberts and Rosenthal [96, Proposition 2.1] (though with possibly different optimal rates for each condition therein).* $\triangleleft$

The proof of this result is found in Section 1.5.6.

Theorem 1.1 controls the convergence of the perturbed chain P_ϵ in terms of the "original" norm (from $L_2(\pi)$). We also demonstrate that P_ϵ is geometrically ergodic in the $L_2(\pi_\epsilon)$ norm, as this would also allow us to use the equivalences in [96]. The following two results allow us to transfer the geometric ergodicity of P_ϵ in $L_2(\pi)$ to other notions of geometric ergodicity. Theorem 1.3 handles the case that the perturbed kernel is reversible, while Theorem 1.2 handles both that the perturbed kernel is reversible or non-reversible.

Theorem 1.2 (Geometric ergodicity of the perturbed chain in the *other norms*; $L_1(\pi_\epsilon)$, $L_2(\pi_\epsilon)$, total variation)**.** *Under the assumptions of Section 1.3.2, if $\epsilon < \alpha$, then:*

(i) *P_ϵ is a.e.-TV-geometrically ergodic with some factor $\rho_{TV} \in (0,1)$, and*

(ii) *P_ϵ is $(L_\infty(\pi_\epsilon), \|\cdot\|_{L_1(\pi_\epsilon)})$-GE with factor $\rho_1 = (1-(\alpha-\epsilon))$ and $C_1(\mu) = \|\mu - \pi_\epsilon\|_{L_2(\pi)}$, and*

(iii) *If $\pi \in L_\infty(\pi_\epsilon)$ then P_ϵ is $L_2(\pi_\epsilon)$-GE with factor $\rho_2 = (1-(\alpha-\epsilon))$ and with*

$$C_2(\mu) = \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \|\mu - \pi\|_{L_2(\pi)} \ .$$

The proof of this result is found in Section 1.5.2.

Example 1.1. *For example, consider perturbations of a Gaussian $AR(1)$ process. Let $Z_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$ and let $W_i \overset{iid}{\sim} \mu$. Take*

$$\begin{aligned} X_{t+1}|X_t &= (1-\alpha)X_t + Z_{t+1} \\ X^\epsilon_{t+1}|X^\epsilon_t &= (1-\alpha)X^\epsilon_t + W_{t+1}. \end{aligned} \tag{1.4}$$

Then the original chain, $\{X_t\}_{t\in\mathbb{N}}$ is not uniformly ergodic, but it is geometrically ergodic. Hence, the results of [2, 52] do not apply. The stationary measure of the exact chain is $\pi \equiv \mathcal{N}(0, \frac{\sigma^2}{\alpha(2-\alpha)})$, it is reversible, and the rate of geometric ergodicity is $(1-\alpha)$. Note *that the perturbed chain, which we will call a μ-$AR(1)$ process, may not be reversible and whether it is geometrically ergodic generally depends on the distribution μ.*

Now, letting ϕ_{σ^2} be the $\mathcal{N}(0,\sigma^2)$ density, for any μ with $\frac{\mathrm{d}\mu}{\mathrm{d}\phi_{\sigma^2}} \in [1-\epsilon, 1+\epsilon]$, ,

$$\begin{aligned}
\|P - P_\epsilon\|^2_{L_2(\pi)} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{\mu(y-(1-\alpha)x)}{\pi(y)} - \frac{\phi_{\sigma^2}(y-(1-\alpha)x)}{\pi(y)}\right)^2 \pi(y)dy\ \pi(x)dx \\
&\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \epsilon^2 \left(\frac{\phi_{\sigma^2}(y-(1-\alpha)x)}{\pi(y)}dy\right)^2 \pi(y)dy\ \pi(x)dx \\
&= \epsilon^2 \|P\|_{L_2(\pi)} \\
&= \epsilon^2
\end{aligned} \tag{1.5}$$

Therefore, when $\epsilon < \alpha$ we can extend the geometric ergodicity of the Gaussian AR process to the $\mu - AR(1)$ process using Theorem 1.2. We can also bound the discrepancy of the stationary measure of the perturbed chain from that $\mathcal{N}(0, \frac{\sigma^2}{\alpha(2-\alpha)})$ using Theorem 1.1. The subsequent results, Corollary 1.1 and Theorem 1.4 of this section may also be applied to this example to bound the discrepancy between the marginal distributions of the μ-$AR(1)$ from a $\mathcal{N}(0, \frac{\sigma^2}{\alpha(2-\alpha)})$ at any time, as well as the approximation error of the time-averaged law of the μ-$AR(1)$ from $\mathcal{N}(0, \frac{\sigma^2}{\alpha(2-\alpha)})$. $\triangleleft$

Theorem 1.3 ($L_2(\pi_\epsilon)$-Geometric ergodicity of the perturbed chain, reversible case)**.** *Under the assumptions of Section 1.3.2, if $\epsilon < \alpha$, and P_ϵ is π_ϵ-reversible, then P_ϵ is $L_2(\pi_\epsilon)$-GE with factor $\rho_2 = (1-\alpha+\epsilon)$ and coefficient function $C(\nu) = \|\nu\|_{L_2(\pi_\epsilon)}$.*

The proof of this result is found in Section 1.5.2.

Corollary 1.1 (Closeness of stationary distributions in $L_2(\pi_\epsilon)$)**.** *Suppose that $\epsilon < \alpha$, and that $\|P - P_\epsilon\|_{L_2(\pi_\epsilon)} \leq \varphi$. Then*

(i) if P_ϵ is π_ϵ reversible, and if $\varphi < \alpha - \epsilon$ then

$$\|\pi - \pi_\epsilon\|_{L_2(\pi_\epsilon)} \leq \frac{\varphi}{\sqrt{(\alpha-\epsilon)^2 - \varphi^2}}\ ,$$

and for any $\mu \in L_2(\pi_\epsilon)$

$$\|\mu P_\epsilon^n - \pi\|_{L_2(\pi_\epsilon)} \leq (1-(\alpha-\epsilon))^n \|\mu - \pi_\epsilon\|_{L_2(\pi_\epsilon)} + \frac{\varphi}{\sqrt{(\alpha-\epsilon)^2 - \varphi^2}}\ ,$$

(ii) if $\pi \in L_\infty(\pi_\epsilon)$ and $\varphi < 1$, then

$$\|\pi - \pi_\epsilon\|_{L_2(\pi_\epsilon)} \leq \frac{\varphi + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon))}{1-\varphi} \ ,$$

and for any $\mu \in L_\infty(\pi_\epsilon)$

$$\begin{aligned}\|\mu P^n_\epsilon - \pi\|_{L_2(\pi_\epsilon)} \leq (1-(\alpha-\epsilon))^n \, \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \, \|\mu - \pi_\epsilon\|_{L_2(\pi_\epsilon)} \\ + \frac{\varphi + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon))}{1-\varphi} \ ,\end{aligned}$$

The proof of this result is found in Section 1.5.2. We turn our attention to bounds on the error of estimation measures of the form $\frac{1}{t}\sum_{k=0}^{t-1} \mu P^k$, and estimates of the form $\frac{1}{t}\sum_{k=0}^{t-1} f(X_k)$. Firstly, when computing Monte Carlo estimates, the bias is controlled by a time-averaged marginal distribution of the form $\frac{1}{t}\sum_{k=0}^{t-1} \mu P^k_\epsilon$. This leads us to the following result.

Theorem 1.4 (Convergence of Time-Averaged Marginal Distributions)**.** *Under the assumptions of Section 1.3.2, suppose $\epsilon < \alpha$ and $\pi_\epsilon \in L_2(\pi)$. Then for any probability distribution $\mu \in L_2(\pi)$,*

$$\left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1} \mu P^k_\epsilon\right\|_{L_2(\pi)} \leq \frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)} \|\pi_\epsilon - \mu\|_{L_2(\pi)} + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}$$

If additionally, $\|P - P_\epsilon\|_{L_2(\pi_\epsilon)} \leq \varphi$ then

(i) if P_ϵ is π_ϵ-reversible, and $\varphi < \alpha - \epsilon$ then

$$\left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1} \mu P^k_\epsilon\right\|_{L_2(\pi_\epsilon)} \leq \frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)} \|\pi_\epsilon - \mu\|_{L_2(\pi_\epsilon)} + \frac{\varphi}{\sqrt{(\alpha-\epsilon)^2-\varphi^2}}$$

(ii) if $\pi \in L_\infty(\pi_\epsilon)$ and $\varphi < 1$, and if $\mu \in L_\infty(\pi_\epsilon)$ then

$$\begin{aligned}\left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1} \mu P^k_\epsilon\right\|_{L_2(\pi_\epsilon)} \leq \frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)} \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \|\pi_\epsilon - \mu\|_{L_2(\pi_\epsilon)} \\ + \frac{\varphi + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon))}{1-\varphi}\end{aligned}$$

The proof of this result is found in Section 1.5.3.1. Relative to the uniform closeness of kernels (in total variation) required [52], our assumption that the approximating kernel is close in the operator norm induced by $L_2(\pi)$ is non-comparable. This is because our bound is in terms of the L_2 distance which always upper-bounds the total variation distance (up to a constant factor of 1/2), but our assumption also does not require spatial uniformity which [52]'s does. Thus, this chapter's assumptions are not weaker nor stronger than those in [52]. Comparing the above results to the corresponding L_1 result of [52], we see that the transient phase bias part of our L_2 bounds differ from their L_1 transient phase bias bound only by a factor which is constant in time, but varies with the initial distribution (as is to be expected when moving from uniform ergodicity to geometric ergodicity).

1.3.4 Mean Squared Error Bounds for Monte Carlo Estimates

Suppose that $(X_k^\epsilon)_{k\in\mathbb{N}\cup\{0\}}$ is a realization of the Markov chain with transition kernel P_ϵ and initial distribution μ. The mean squared error of a Monte Carlo estimate of πf made using $(X_k^\epsilon)_{k\leq t}$ is given by

$$\mathrm{MSE}_t^\epsilon(\mu, f) = \mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \tag{1.6}$$

Theorem 1.5 (Mean Squared Error of Monte Carlo Estimates from the Perturbed Chain)**.** *Under the assumptions of Section 1.3.2, if $\epsilon < \alpha$, $X_0^\epsilon \sim \mu$, P_ϵ is π_ϵ-reversible, and $\rho_2 = (1-(\alpha-\epsilon))$ then for $f \in L_4'(\pi_\epsilon)$*

(i) if $f \in L_2'(\pi)$ as well, then

$$\begin{aligned} MSE_t^\epsilon(\mu, f) \leq{}& \frac{2\left\|f-\pi_\epsilon f\right\|^2_{L_2'(\pi_\epsilon)}}{(1-\rho_2)t} + \frac{2^{7/2}\left\|\mu-\pi_\epsilon\right\|_{L_2(\pi_\epsilon)}\left\|f-\pi_\epsilon f\right\|^2_{L_4'(\pi_\epsilon)}}{(1-\rho_2)^2t^2} \\ &+ \left\|f-\pi_\epsilon f\right\|^2_{L_2'(\pi)}\left(\frac{\epsilon^2}{\alpha^2-\epsilon^2} + 2\frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\frac{1}{t(\alpha-\epsilon)}\left\|\pi_\epsilon-\mu\right\|_{L_2(\pi)}\right) \end{aligned}$$

and

$$MSE_t^{\epsilon}(\mu, f) \leq \frac{4\left\|f - \pi_\epsilon f\right\|^2_{L'_2(\pi_\epsilon)}}{(1-\rho_2)t} + \frac{2^{9/2}\left\|\mu - \pi_\epsilon\right\|_{L_2(\pi_\epsilon)}\left\|f - \pi_\epsilon f\right\|^2_{L'_4(\pi_\epsilon)}}{(1-\rho_2)^2 t^2} + 2\left\|f - \pi_\epsilon f\right\|^2_{L'_2(\pi)}\frac{\epsilon^2}{\alpha^2 - \epsilon^2}\ ,$$

and

(ii) if $\left\|P - P_\epsilon\right\|_{L_2(\pi_\epsilon)} \leq \varphi < (1-\rho_2)$, *then*

$$MSE_t^{\epsilon}(\mu, f) \leq \frac{2^{7/2}\left\|\mu - \pi_\epsilon\right\|_{L_2(\pi_\epsilon)}\left\|f - \pi_\epsilon f\right\|^2_{L'_4(\pi_\epsilon)}}{(1-\rho_2)^2 t^2} + \left\|f - \pi_\epsilon f\right\|^2_{L'_2(\pi_\epsilon)}\left(\frac{\varphi^2}{(1-\rho_2)^2 - \varphi^2} + 2\frac{1 + \frac{\varphi}{\sqrt{(1-\rho_2)^2 - \varphi^2}}}{t(1-\rho_2)}\left\|\pi_\epsilon - \mu\right\|_{L_2(\pi_\epsilon)}\right),$$

and

$$MSE_t^{\epsilon}(\mu, f) \leq \frac{2^{9/2}\left\|\mu - \pi_\epsilon\right\|_{L_2(\pi_\epsilon)}\left\|f - \pi_\epsilon f\right\|^2_{L'_4(\pi_\epsilon)}}{(1-\rho_2)^2 t^2} + \left\|f - \pi_\epsilon f\right\|^2_{L'_2(\pi_\epsilon)}\left(\frac{2\varphi^2}{(1-\rho_2)^2 - \varphi^2} + \frac{4}{t(1-\rho_2)}\left\|\pi_\epsilon - \mu\right\|_{L_2(\pi_\epsilon)}\right)$$

The proof of this result is found in Section 1.5.3.3. Perturbation bounds based upon drift and minorization conditions could provide similar MSE bounds for functions in $L_2(\pi_\epsilon)$ with $\sup_{x \in \mathcal{X}} \frac{|f|}{\sqrt{V}} < \infty$ (where V is the function appearing in the drift condition), as in the work of Johndrow and Mattingly [51]. While that may be a larger class of functions than $L'_4(\pi_\epsilon)$ (depending on what V happens to be), the class $L'_4(\pi_\epsilon)$ is quite rich making this bound still useful. Moreover, the class of functions to which our MSE bounds apply, and the value of the bound itself, depend only on intrinsic features of the Markov chains under consideration. In contrast bounds based on drift and minorization conditions include extrinsic features—introduced by the user for analytic purposes (such as the drift function, V)—of which many choices might exist; each leading to different function classes and different bounds.

1.4 Applications to Markov Chain Monte Carlo

In this section we apply our theoretical results to some specific variants of Markov Chain Monte Carlo (MCMC) algorithms to obtain guarantees for noisy and/or approximate variants of MCMC algorithms. MCMC is used to generate (correlated) samples approximately from a target distribution for which the (unnormalized) density can be evaluated. The key insight is to construct a (typically reversible) Markov chain for which the stationary distribution is the target distribution. This is possible since the reversibility condition is readily verified locally (without integration).

The most commonly used family of MCMC methods is the Metropolis–Hastings algorithm (MH). The chain is initialized from some distribution $X_0 \sim \mu_0$. At each step a *proposal* is drawn from some transition kernel, $Y_t \sim Q(X_{t-1}, \cdot)$. Suppose that the kernel $Q(x, \cdot)$ has density $q(\cdot|x)$. The proposal is *accepted* with probability $a(Y_t|X_{t-1}) = \min\left(1, \frac{\pi(Y_t)q(X_{t-1}|Y_t)}{\pi(X_{t-1})q(Y_t|X_{t-1})}\right)$. If the proposal is accepted then $X_t = Y_t$, and if it is *rejected* (not accepted) then $X_t = X_{t-1}$. The combination of proposal and accept/reject steps yields a π-reversible Markov kernel, and reversibility guarantees that the stationary distribution is the target distribution. The user has freedom in selecting the proposal kernel, Q, and some choice lead to better performance than others. The accept/reject step requires evaluating the target density, π, twice on each step.

A large body of research exists guaranteeing that specific MCMC algorithms will be geometrically ergodic (see for example [42, 67, 95], and many more.). These typically verify geometric ergodicity for a collection of target distributions, π, and for a small family of proposal kernels, Q.

If the target likelihood involves some integral which is computed numerically or by simple Monte Carlo then the numerical and/or stochastic approximation introduces a *perturbation* to the idealized MCMC scheme. This occurs even in standard and widely used statistical models such as generalized linear mixed effect models (GLMMs), since the random effects are nuisance variables which need to be integrated away, either using Laplace or Gaussian quadrature schemes, or by simple Monte Carlo, in order to evaluate the likelihood. Since the Metropolis–Hastings algorithm requires evaluation of the density, these each introduce a perturbation in the acceptance ratio, and hence in the actual transition kernel of the MH scheme. We now consider the extent to which our results from Section 3 can be applied to prove geometric ergodicity for certain approximate MCMC algorithms.

1.4.1 Noisy and Approximate MCMC

The noisy (or approximate) Metropolis–Hastings algorithm (nMH), as found in Alquier et al. [2] (see also Medina–Aguayo et al. [74]) was briefly described above. The algorithm is defined exactly the same way as the Metropolis–Hastings algorithm, except that the *acceptance ratio*, $a(Y_t|X_{t-1})$, is replaced by a (possibly stochastic) approximation $\widehat{a}(Y_t|X_{t-1}, Z_t)$. Here Z_t denotes some random element providing an additional source of randomness, so that $a(Y_t|X_{t-1}, Z_t)$ is not $\sigma(Y_t, X_{t-1})$-measurable when the approximation $\widehat{a}(Y_t|X_{t-1}, Z_t)$ is stochastic. In the case of a deterministic approximation, Z_t can be ignored or treated as a constant. The approximation can typically be though of as replacing the target density in the acceptance ratio with some approximation. This includes most approximate MCMC algorithms which preserve the state space and the Markov property, such as replacing π with a deterministic approximation or and independent stochastic approximation at each step (as in Monte Carlo within Metropolis). It does not include algorithms which retain the Markov property only an augmented state space, such as the Pseudo-Marginal approach of Andrieu and Roberts [4].

For our analysis of these algorithms, P will represent the transition kernel for the MH algorithm while $\widehat{P}$ will represent the kernel for the corresponding nMH chain. The key step in applying our results from Section 1.3 will be to show the $L_2(\pi)$ closeness of the nMH transition kernel to the MH transition kernel. Again, $\|\cdot\|_{L_2(\pi)}$ is the norm on $L_2(\pi)$ and the corresponding operator norm. We will assume that π and $\{Q(x,\cdot)\}_{x\in\mathcal{X}}$ are all absolutely continuous with respect to the Lebesgue measure and have densities π and $\{q(\cdot|x)\}_{x\in\mathcal{X}}$ respectively. All arguments used would still apply if there were an arbitrary dominating measure in place of the Lebesgue measure. Let $F_{y|x}$ be the regular conditional distribution for Z given $X = x$ and $Y = y$, and let $f_{y|x}$ be its Lebesgue density. Define the following *perturbation function* for the nMH algorithm as

$$r(y|x) = \mathop{\mathbb{E}}_{Z\sim F_{y|x}} \left(a(y|x) - \widehat{a}(y|x,Z)\right) = \int \left(a(y|x) - \widehat{a}(y|x,z)\right) f_{y|x}(z)dz$$

Theorem 1.6 (Geometric ergodicity and closeness of stationary distributions noisy or approximate Metropolis–Hastings)**.** *Let P be the transition kernel for a Metropolis–Hastings algorithm with proposal distribution Q, target distribution π, and acceptance ratio $a(\cdot|\cdot)$. Let $\widehat{P}$ be the transition kernel for a corresponding noisy Metropolis–Hastings algorithm with approximate/noisy acceptance ratio $\widehat{a}(\cdot|\cdot,\cdot)$. Let $r(\cdot|\cdot)$ be the corresponding perturbation function.*

If $\|Q\|_{L_2(\pi)} < \infty$ and $\sup_{x,y} |r(y|x)| \le R$ then

$$\left\|\widehat{P} - P\right\|_{L_2(\pi)} \le R(1 + \|Q\|_{L_2(\pi)}) \ . \tag{1.7}$$

Furthermore, if P is reversible and $L_2(\pi)$-geometrically ergodic with geometric contraction factor $(1-\alpha)$, and $\epsilon = R(1 + \|Q\|_{L_2(\pi)}) < \alpha$, then $\widehat{P}$ has a stationary distribution, $\widehat{\pi}$ and the assumptions outlined in Section 1.3.2 hold with $P_\epsilon = \widehat{P}$ and $\pi_\epsilon = \widehat{\pi}$.

Therefore, Theorems 1.1 to 1.5 and Corollary 1.1 can all be applied. In particular, $\widehat{P}$ is $L_2(\pi)$-geometrically ergodic with factor $1-(\alpha-R(1+\|Q\|_{L_2(\pi)}))$, it is a.e.-TV geometrically ergodic, and

$$\|\widehat{\pi} - \pi\|_{L_2(\pi)} \le \frac{R(1 + \|Q\|_{L_2(\pi)})}{\sqrt{\alpha^2 - R^2(1 + \|Q\|_{L_2(\pi)})^2}} \ ; \tag{1.8}$$

and, if $\widehat{P}$ is reversible then it is $L_2(\widehat{\pi})$ geometrically ergodic with factor $(1-(\alpha - R(1 + \|Q\|_{L_2(\pi)})))$.

The above theorem provides an alternative to the analogous result of Corollary 2.3 from [2], relaxing the uniform ergodicity assumption. In particular, it requires that $Q \in \mathcal{B}(L_2(\pi))$ and that $R(1 + \|Q\|_{L_2(\pi)}) < \alpha$. The first of these requirements is not dramatically limiting since the user has control over the choice of Q. The second of these requirements is also not dramatically limiting as control over R may be interpreted as limiting the amount of noise in the nMH algorithm and such control is required regardless in order to ensure the accuracy of approximation in both the geometrically ergodic and uniformly ergodic cases.

1.4.2 Application to Fixed Deterministic Approximations

Suppose we run a fixed Metropolis–Hastings algorithm, but replace the target density with one which is close everywhere. Perhaps this alternative density is easier to compute (e.g. replacing an integral with a Laplace approximation as in Kass et al. [55], or replacing a full sample with a coreset for sub-sampled Bayesian Inference as in Campbell and Broderick [23]). By construction we would know that the approximate target distribution is close to the ideal target distribution. The question still remains whether geometric ergodicity is preserved. We resolve this question in the case that the approximation has constant relative error.

Corollary 1.2. *Suppose we can approximate the unnormalized target density, $C\pi$, by $\widehat{\pi}$, with a θ-bounded relative error;*

$$\sup_{x\in\mathcal{X}}\left|\log\frac{C\pi(x)}{\widehat{\pi}(x)}\right|\le\theta\ . \tag{1.9}$$

If the Metropolis–Hastings algorithm with proposal kernel Q is $L_2(\pi)$-geometrically ergodic with factor $(1-\alpha)$, and if $\theta<\frac{\alpha}{2(1+\|Q\|_{L_2(\pi)})}$, then the corresponding approximate transition kernel, $\hat{P}$, is $L_2(\widehat{\pi})$-geometrically ergodic and

$$\|\widehat{\pi}-\pi\|_{L_2(\pi)}\le\frac{2\theta(1+\|Q\|_{L_2(\pi)})}{\sqrt{\alpha^2-4\theta^2(1+\|Q\|_{L_2(\pi)})^2}}\ ; \tag{1.10}$$

Proof. Since the function $x\mapsto 1\wedge\exp(x)$ is 1-Lipschitz, we have:

$$\begin{aligned}|r(y|x)|&=|a(y|x)-\widehat{a}(y|x)|\\&\le\left|\log\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}-\log\frac{\widehat{\pi}(y)q(x|y)}{\widehat{\pi}(x)q(y|x)}\right|\\&=\left|\log\frac{C\pi(y)}{\widehat{\pi}(y)}-\log\frac{C\pi(x)}{\widehat{\pi}(x)}\right|\\&\le 2\theta\end{aligned} \tag{1.11}$$

So, $\widehat{P}$ will be $L_2(\pi)$-geometrically ergodic as long as P was geometrically ergodic with some factor $0 \leq (1-\alpha) < 1$ and

$$\theta < \frac{\alpha}{2(1+\|Q\|_{L_2(\pi)})} \ . \tag{1.12}$$

Moreover, in this case, $\widehat{P}$ is reversible. Thus, we can use Theorem 1.3 to obtain $L_2(\widehat{\pi})$-geometric ergodicity of $\widehat{P}$, with factor $1-\alpha+2\theta(1+\|Q\|_{L_2(\pi)})$. □

In this scenario, we can also use Theorem 1.5 to get quantitative bounds for the mean-squared error of any Monte Carlo estimates made using $\widehat{P}$, or any of our other results in Theorems 1.1 to 1.4 and Corollary 1.1 as needed.

Example 1.2 (Independence Sampler)**.** *The previous result also immediately gives that if $\frac{\mathrm{d}\widehat{\pi}}{\mathrm{d}\pi}$ is bounded above by $C < \exp(1/4)$ and below by $c > \exp(-1/4)$ then the independence sampler for $\widehat{\pi}$ with proposals from π is geometrically ergodic with factor at most $4\max(\log C, -\log(c))$. This is, however, sub-optimal when compared to Smith and Tierney [110] which only requires a finite upper bound on $\frac{\mathrm{d}\widehat{\pi}}{\mathrm{d}\pi}$ to establish* uniform ergodicity*.* ◁

Example 1.3 (Laplace Approximation for GLMMs)**.** *Generalized linear mixed models (GLMMs) (see Breslow and Clayton [16],McCulloch and Neuhaus [72], etc.) are widely used in the modelling of non-normal response variables under repeated or correlated measurements. They are the natural common extension of generalized linear models and linear mixed effects models. They handle dependence between observations by introducing Gaussian latent variables. These* random effects *are nuisance variables for the purpose of inference. In order to perform Bayesian inference for GLMMs, one requires samples from the marginal posterior distribution of the parameters given the data. The marginal posterior, here, is the posterior for the parameters given the observations, in contrast to the joint posterior of the random effects and the parameters given the data.*

This can be approached in two ways. One option is to obtain samples for the random effects and parameters jointly given the data, and discard the random effects to get marginal posterior samples for the parameters. The second option is to approximate the likelihood by integrating (numerically) over the random effects, and using the resulting approximate likelihood in the calculations involving the unnormalized posterior for the parameters.

In the second case, when the prior for the parameters is compactly supported, if one had *established a result saying that a particular MH procedure for the exact posterior distribution of the parameters would be geometrically ergodic, then one could directly transfer this result to the approximate posterior computed using a Laplace approximation, at least for large enough samples. This is valid since the Laplace approximation has constant relative error on compact sets, and the relative error decreases with sample size (see Tierney and Kadane [116]). Hence, for a large enough sample size Eq. (1.12) will be satisfied regardless of what the proposal kernel Q was (as long as $\|Q\|_{L_2(\pi)}$ was finite).* $\triangleleft$

Example 1.4 (Uniform Coresets)**.** *In Bayesian inference with large samples, an approach to reducing the computational burden of evaluating the likelihood in the unnormalized posterior for MCMC accept/reject steps is to select a representative subsample of the data and to up-weight the contributions of each of the selected samples in a way to best approximate the original likelihood. These up-weighted subsamples are called* coresets*. They naturally give rise to approximate MCMC methods in which the true posterior is replaced by an approximation based upon a coreset. Several methods for coreset construction exist, however relatively little work has been done to assess their impact upon approximate MCMC methods. We will consider the* uniform coreset *construction of Huggins et al. [44] (as so named in [23]).*

Campbell and Broderick [23, Theorem 3.2] provides the guarantee that, with probability $(1-\delta)$, the unnormalized approximate posterior $\hat{C}\hat{\pi}$ based on a uniform coreset of size M will satisfy

$$\sup_{x\in\mathcal{X}} \frac{1}{|\mathcal{L}(x)|}\left|\log\frac{\hat{C}\hat{\pi}(x)}{C\pi(x)}\right| \leq \frac{\sigma}{\sqrt{M}}\left(\frac{3}{2}D + \overline{\eta}\sqrt{2\log\left(1/\delta\right)}\right) \tag{1.13}$$

where $\sigma = \sum_{n=1}^{N}\sigma_n$, N is the number of observations, $\sigma_n = \sup_{x\in\mathcal{X}}\left|\frac{\mathcal{L}_i(x)}{\mathcal{L}(x)}\right|$, $\mathcal{L}_i(x)$ is the log-likelihood of parameter x at the ith observation, $\mathcal{L}(x) = \sum_{i=1}^{N}\mathcal{L}_i(x)$ is the log-likelihood of the dataset

$$\overline{\eta} = \max_{i,j\in\{1,\ldots,N\}}\sup_{x\in\mathcal{X}}\frac{1}{|\mathcal{L}(x)|}\left|\frac{\mathcal{L}_i(x)}{\sigma_i} - \frac{\mathcal{L}_j(x)}{\sigma_j}\right|, \tag{1.14}$$

and D is the approximate dimension *of $\{\mathcal{L}_i\}_{i=1}^{n}$ ([23, Definition 3.1])*

If in addition to assuming that $\{\sigma_i\}_{i=1}^{N}$ are all finite as in [23, Section 3], one were to assume that $|\mathcal{L}(x)|$ is bounded as a function of x, then the uniform coreset result would *imply the conditions of our Corollary 1.2, namely that*

$$\sup_{x\in\mathcal{X}}\left|\log\frac{C\pi(x)}{\widehat{\pi}(x)}\right| \le \frac{\sigma\left\|\mathcal{L}\right\|_{\infty}}{\sqrt{M}}\left(\frac{3}{2}D+\overline{\eta}\sqrt{2\log\left(1/\delta\right)}\right), \tag{1.15}$$

with high probability. Consequently, for any proposal kernel $Q : L_2(\pi) \to L_2(\pi)$ we should be able to choose M sufficiently large so that with high probability

$$\frac{\sigma\left\|\mathcal{L}\right\|_{\infty}}{\sqrt{M}}\left(\frac{3}{2}D+\overline{\eta}\sqrt{2\log\left(1/\delta\right)}\right) < \frac{\alpha}{2(1+\|Q\|_{L_2(\pi)})}. \tag{1.16}$$

Hence the approximating Markov chain will by geometrically ergodic with high probability.

$\triangleleft$

1.4.3 Application to Monte Carlo Within Metropolis

Following Medina–Aguayo et al. [73], we can get bounds for the simple Monte Carlo within Metropolis algorithm (MCwM). This is the special case of nMH where we approximate the likelihood ratio $\frac{\pi(y)}{\pi(x)} = \frac{\mathbb{E}\Pi(y,Z)}{\mathbb{E}\Pi(x,Z)}$ by $\widehat{\left(\frac{\pi(y)}{\pi(x)}\right)} = \frac{\sum_{i=1}^{N}\Pi(y,Z_i)}{\sum_{i=N+1}^{2N}\Pi(x,Z_i)}$ using a new independent sample taken each time the likelihood is evaluated. In the notation of the previous section,

$$\widehat{a}(y|x,z) = 1 \wedge \frac{q(x|y)\sum_{i=1}^{N}\Pi(y,z_i)}{q(y|x\sum_{i=N+1}^{2N}\Pi(x,z_i)} \tag{1.17}$$

Let

$$\begin{aligned} W_k(x) &= \frac{1}{k\pi(x)}\sum_{i=1}^{k}\Pi(x,Z_i) \\ i_k(x)^2 &= \mathbb{E}[W_k(x)^{-2}] \\ s(x) &= \frac{1}{\sqrt{\pi(x)}}\,\mathrm{StdDev}(\Pi(x,Z_1)) \end{aligned} \tag{1.18}$$

[73, Lemma 14] tells us that if there is a $k \in \mathbb{N}$ such that $i_k(x) < \infty$ for all $x \in \mathcal{X}$ then for $N \ge k$

$$\begin{aligned} |r(y|x)| &\le a(y|x)\frac{1}{\sqrt{N}}i_k(y)\left(s(x)+s(y)\right) \\ &\le \frac{1}{\sqrt{N}}i_k(y)\left(s(x)+s(y)\right) \end{aligned} \tag{1.19}$$

Corollary 1.3. *Let P be the Metropolis–Hastings transition kernel for the target density π*

and proposal kernel Q. Let $\widehat{P}_N$ be the corresponding MCwM transition kernel when $\pi(\cdot)$ is approximated by $\frac{1}{N}\sum_{i=1}^{N}\Pi(\cdot, Z_i)$.

Assume that s and i_k as defined above are uniformly bounded for some $k \in \mathbb{N}$. Suppose further that $N_0 = \max\left(k, \frac{4\|i_k\|_\infty^2 \|s\|_\infty^2 (1+\|Q\|_{L_2(\pi)})^2}{\alpha^2}\right)$, and $N \geq \lfloor N_0 \rfloor + 1$.

Then $\widehat{P}_N$ is reversible and $L_2(\pi)$-geometrically ergodic with factor $1-\alpha+\frac{1}{\sqrt{N/N_0}}$, and has a stationary distribution, $\widehat{\pi}_N(x) \propto \frac{\pi(x)}{N\ \mathbb{E}\left[\left(\sum_{i=1}^{N}\Pi(x,Z_i)\right)^{-1}\right]}$ with

$$\|\pi - \widehat{\pi}_N\|_{L_2(\pi)} \leq \sqrt{\frac{N_0}{N\alpha^2 - N_0}} \tag{1.20}$$

Proof. Suppose that $N \geq \lfloor N_0 \rfloor + 1$. From Theorem 1.1, we know that the perturbed chain, $\widehat{P}_N$ is $L_2(\pi)$-geometrically ergodic with factor $1-\alpha+\frac{1}{\sqrt{N/N_0}}$, has a stationary distribution, $\widehat{\pi}_N$ with

$$\|\pi - \widehat{\pi}_N\|_{L_2(\pi)} \leq \sqrt{\frac{N_0}{N\alpha^2 - N_0}}\ . \tag{1.21}$$

Moreover, by inspection, $\widehat{P}_N$ is reversibility with respect to

$$\widehat{\pi}_N(x) \propto \frac{\pi(x)}{N\ \mathbb{E}\left[\left(\sum_{i=1}^{N}\Pi(x,Z_i)\right)^{-1}\right]}.$$

Thus, we can use Theorem 1.3 to obtain $L_2(\widehat{\pi}_N)$-geometric ergodicity of $\widehat{P}_N$, with factor $1-\alpha+\frac{1}{\sqrt{N/N_0}}$. □

Remark 1.6. *A simple scenario under which these i_k and s are uniformly bounded is when the joint density of x and Z is bounded above an below by a multiple of the marginal of x, so that*

$$\frac{\Pi(x,z)}{\pi(x)} \in [c, C] \tag{1.22}$$

for all $(x,z) \in \mathcal{X} \times \mathcal{Z}$. This condition is essentially tight if we wish to take $k=1$ and the base measure to be the Lebesgue measure restricted to $U \subset \mathbb{R}^d$; in this case the condition $\|i_k(x)\|_{L_\infty} < \infty$ implies that

$$\int_U \frac{\pi(x)}{\Pi(x,z)}dz = \mathbb{E}_{Z\sim\frac{\Pi(x,\cdot)}{\pi(x)}} \frac{\pi(x)^2}{\Pi(x,\cdot)^2} < \infty \tag{1.23}$$

for all x. That is, the reciprocal of the conditional density of Z given $X = x$ has a finite integral for each x. $\triangleleft$

Remark 1.7. *More generally, [73, Lemma 23] tells us that if $\mathbb{E}[W_{k_0}(x)^{-p}] < \infty$ for some $k_0 \in \mathbb{N}$ and $p > 0$ then for $k \geq k_0 \left\lceil \frac{2}{p} \right\rceil$, $i_k(x)^2 < \mathbb{E}[W_{k_0}(x)^{-p}]$. Therefore, in order to uniformly bound $i_k(x)$, it is sufficient to bound $\mathbb{E}[W_{k_0}(x)^{-p}]$ uniformly in x for some $k_0 \in \mathbb{N}$, $p > 0$. This is much less restrictive than trying to bound $i_1(x)$. In the case that $p < 1, k_0 = 1$ this is much less restrictive then $p = 2, k_0 = 1$; it is equivalent to requiring that tempered versions of conditional distribution $\frac{\Pi(x,\cdot)}{\pi(x)}$ can be normalized by uniformly bounded normalizing constants. This would be true, if for example $(Z|X = x) \sim \mathcal{N}(\mu(x), \sigma^2(x))$ with $\sigma^2(x)$ uniformly bounded in x. More generally, using $0 < p < 1$, instead of $p = 2$ whenever the conditional law of Z has uniform exp-poly tails, $\frac{\Pi(x,z)}{\pi(x)} \leq \exp(-C\,|z - \mu(x)|^\alpha)$, with $\alpha > 0$, the p-version of the condition would hold.*

$\triangleleft$

We could also use Theorem 1.5 to get quantitative bounds for the mean-squared error of any Monte Carlo estimates made using $\widehat{P}_N$, or any of our other results in Theorems 1.1 to 1.4 and Corollary 1.1 as needed.

In [73], they also consider a case where the the assumption that s and i_k are uniformly bounded is dropped, and instead, the perturbed kernel is restricted to a bounded region. We do not address this case here.

1.5 Proofs

1.5.1 Proof of Theorem 1.1

The following lemma is contained in the remark after Theorem 2.1 of [96]; we prove it here as well since the proof is so simple.

Lemma 1.3 (Remark in [96])**.** *For any probability measure $\mu \in L_2(\pi)$,*

$$\|\mu - \pi\|^2_{L_2(\pi)} = \|\mu\|^2_{L_2(\pi)} - 1$$

Proof.

$$0 \le \|\mu - \pi\|^2_{L_2(\pi)} = \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)^2 \mathrm{d}\pi = \int \left(\left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 - 2\frac{\mathrm{d}\mu}{\mathrm{d}\pi} + 1\right) \mathrm{d}\pi$$
$$= \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi}\right)^2 \mathrm{d}\pi - 2\int \mathrm{d}\mu + \int \mathrm{d}\pi = \|\mu\|^2_{L_2(\pi)} - 1$$

□

We will make use of the following simplified version of Theorem 2.1 from [96] as well:

Proposition 1.2 (Equivalent definitions of $L_2(\pi)$ geometric ergodicity from [96]). *For a reversible Markov chain with kernel P and stationary distribution π on state space $\mathcal{X}$, the following are equivalent (and ρ is equal in both cases):*

(i) P is $L_2(\pi)$-geometrically ergodic with optimal rate ρ and coefficient function $C(\mu) = \|\mu - \pi\|_{L_2(\pi)}$,

(ii) P has $L_{2,0}(\pi)$-spectral radius and norm both equal to ρ;

$$\sup_{\nu \in L_{2,0}(\pi)\setminus\{0\}} \frac{\|\nu P\|_{L_2(\pi)}}{\|\nu\|_{L_2(\pi)}} = \rho = r(P|_{L_{2,0}(\pi)}) \ ,$$

Where

$$r(P|_{L_{2,0}(\pi)}) := \sup\left\{|\rho| : \rho \in \mathbb{C} \ \textit{and} \ \left(P|_{L_{2,0}(\pi)} - \rho I_{L_{2,0}(\pi)}\right) \ \textit{is not invertible}\right\} \quad (1.24)$$

Note that while when the kernel is reversible we may take $C(\mu) = \|\mu - \pi\|_{L_2(\pi)}$ in the bound corresponding $L_2(\pi)$-GE with optimal rate ρ, this is not true for non-reversible chains. By applying the above theorem in our context we have:

Lemma 1.4. *Under the assumptions of Section 1.3.2,*

$$\|\nu_1 P^n - \nu_2 P^n\|_{L_2(\pi)} \le (1-\alpha)^n \|\nu_1 - \nu_2\|_{L_2(\pi)}$$

for any probability distributions $\nu_1, \nu_2 \in L_2(\pi)$. In particular, taking $\nu_2 = \pi$,

$$\|\nu_1 P^n - \pi\|_{L_2(\pi)} \le (1-\alpha)^n \|\nu_1 - \pi\|_{L_2(\pi)} = (1-\alpha)^n \sqrt{\|\nu_1\|^2_{L_2(\pi)} - 1}$$

and applying Cauchy-Schwarz yields

$$\|\nu_1 P^n - \pi\|_{L_1(\pi)} \leq \|\nu_1 P^n - \pi\|_{L_2(\pi)} \leq (1-\alpha)^n \|\nu_1 - \pi\|_{L_2(\pi)}$$

We begin with a first result giving sufficient conditions under which the stationary distribution π_ϵ of the perturbed chain is in $L_2(\pi)$:

Lemma 1.5. *Under the assumptions of Section 1.3.2, if in addition $\epsilon < \alpha$, then P_ϵ has a unique stationary distribution, $\pi_\epsilon \in L_2(\pi)$, and $\|\pi_\epsilon - \pi\|_{L_2(\pi)} \leq \frac{\epsilon}{\alpha - \epsilon}$.*

Proof. Since P_ϵ is π-irreducible and aperiodic, it has at most one stationary distribution, π_ϵ, with $\pi_\epsilon \ll \pi$ (see for example [28, Corollary 9.2.16]).

Suppose for now that πP_ϵ^n has an $L_2(\pi)$ limit, π_ϵ; Then, using the triangle inequality, and the contraction property ($\|P_\epsilon\|_{\mathrm{TV}} = 1$), and Cauchy-Schwarz

$$\begin{aligned}
\|\pi_\epsilon P_\epsilon - \pi_\epsilon\|_{\mathrm{TV}} &\leq \|\pi_\epsilon P_\epsilon - \pi P_\epsilon^n\|_{\mathrm{TV}} + \|\pi P_\epsilon^n - \pi_\epsilon\|_{\mathrm{TV}} \\
&\leq \left\|\pi_\epsilon - \pi P_\epsilon^{n-1}\right\|_{\mathrm{TV}} + \|\pi P_\epsilon^n - \pi_\epsilon\|_{\mathrm{TV}} \\
&\leq \left\|\pi_\epsilon - \pi P_\epsilon^{n-1}\right\|_{L_2(\pi)} + \|\pi P_\epsilon^n - \pi_\epsilon\|_{L_2(\pi)} \overset{n\to\infty}{\to} 0
\end{aligned}$$

we find that π_ϵ must be stationary for P_ϵ.

It remains to verify that $\{\pi P_\epsilon^n\}_{n\in\mathbb{N}}$ is an $L_2(\pi)$-Cauchy sequence, and thus from completeness it must have an $L_2(\pi)$-limit. To this end, define $Q_\epsilon = (P_\epsilon - P)$. Let $\mathbf{2}^k = \{0,1\}^k$ for all $k \in \mathbb{N}$. We will expand $\pi(P + Q_\epsilon)^n$ and use the following facts:

(A) $\forall R \in \mathcal{B}(L_2(\pi))\ [\pi P^n R = \pi R]$

(B) $Q_\epsilon : L_2(\pi) \to L_{2,0}(\pi)$

(C) $P|_{L_{2,0}(\pi)} \in \mathcal{B}(L_{2,0}(\pi))$ and $\left\|P|_{L_{2,0}(\pi)}\right\|_{L_{2,0}(\pi)} \leq (1-\alpha)$

Since the operators P and Q_ϵ do not (necessarily) commute, when we expand $(P+Q)^n$ we must have one distinct term per binary sequence of length n. We can then group terms by the number of leading Ps, and use (A) to cancel the leading terms.

Let $m, n \in \mathbf{N}$ be arbitrary with $m \leq n$.

$$
\begin{aligned}
&\|\pi P_\epsilon^n - \pi P_\epsilon^m\|_{L_2(\pi)} \\
&\quad = \|\pi(P+Q_\epsilon)^n - \pi(P+Q_\epsilon)^m\|_{L_2(\pi)} \\
&\quad = \left\|\pi\left[\left(\sum_{\mathbf{b}\in\mathbf{2}^n}\prod_{j=1}^{n} P^{b_j}Q_\epsilon^{1-b_j}\right) - \left(\sum_{\mathbf{b}\in\mathbf{2}^m}\prod_{j=1}^{m} P^{b_j}Q_\epsilon^{1-b_j}\right)\right]\right\|_{L_2(\pi)} \\
&\quad = \left\|\pi\left[\left(P^n + \sum_{k=0}^{n-1} P^{n-k-1}Q_\epsilon \sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k} P^{b_j}Q_\epsilon^{1-b_j}\right)\right.\right. \\
&\qquad\qquad \left.\left. - \left(P^m + \sum_{k=0}^{m-1} P^{m-k-1}Q_\epsilon \sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k} P^{b_j}Q_\epsilon^{1-b_j}\right)\right]\right\|_{L_2(\pi)} \\
&\quad = \left\|\left(\pi + \sum_{k=0}^{n-1}\pi Q_\epsilon \sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k} P^{b_j}Q_\epsilon^{1-b_j}\right) - \left(\pi + \sum_{k=0}^{m-1}\pi Q_\epsilon \sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k} P^{b_j}Q_\epsilon^{1-b_j}\right)\right\|_{L_2(\pi)} \\
&\quad = \left\|\pi\sum_{k=m}^{n-1} Q_\epsilon \sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k} P^{b_j}Q_\epsilon^{1-b_j}\right\|_{L_2(\pi)} \\
&\quad \le \epsilon\sum_{k=m}^{n-1}\sum_{\mathbf{b}\in\mathbf{2}^k}\prod_{j=1}^{k}(1-\alpha)^{b_j}\epsilon^{1-b_j} \\
&\quad = \epsilon\sum_{k=m}^{n-1}(1-\alpha+\epsilon)^k \\
&\quad \le \frac{\epsilon}{\alpha-\epsilon}(1-\alpha+\epsilon)^m
\end{aligned}
$$

Since this upper bound on $\|\pi P_\epsilon^n - \pi P_\epsilon^m\|_{L_2(\pi)}$ decreases to 0 monotonically with $m = \min(m,n)$ then the sequence must be $L_2(\pi)$-Cauchy.

Now, to bound the norm of π_ϵ we take $m = 0$ and we get that for all $n \in \mathbb{N}$:

$$\|\pi P_\epsilon^n - \pi\|_{L_2(\pi)} \le \frac{\epsilon}{\alpha-\epsilon}$$

From the continuity of norm, it must be the case that $\|\pi_\epsilon - \pi\|_{L_2(\pi)} \le \frac{\epsilon}{\alpha-\epsilon}$ $\square$

Lemma 1.6. *Under the assumptions of Section 1.3.2, if in addition $\epsilon < \alpha$ then*

$$1 \le \|\pi_\epsilon\|_{L_2(\pi)} \le \frac{\alpha}{\sqrt{\alpha^2-\epsilon^2}}$$

and

$$0 \leq \|\pi - \pi_\epsilon\|_{L_2(\pi)} \leq \frac{\epsilon}{\sqrt{\alpha^2 - \epsilon^2}}.$$

Proof. The two lower bounds are immediate from Lemma 1.3 and the positivity of norms:

$$0 \leq \|\pi - \pi_\epsilon\|_{L_2(\pi)}^2 = \|\pi_\epsilon\|_{L_2(\pi)}^2 - 1$$

To derive the first upper bound, we apply Lemma 1.3, our assumptions about the operators P and P_ϵ, and triangle inequality, to $\|\pi - \pi_\epsilon\|_2$:

$$\begin{aligned}
\sqrt{\|\pi_\epsilon\|_{L_2(\pi)}^2 - 1} = \|\pi - \pi_\epsilon\|_{L_2(\pi)} &= \|\pi P - \pi_\epsilon P + \pi_\epsilon P - \pi_\epsilon P_\epsilon\|_{L_2(\pi)} \\
&\leq \|\pi P - \pi_\epsilon P\|_{L_2(\pi)} + \|\pi_\epsilon P - \pi_\epsilon P_\epsilon\|_{L_2(\pi)} \\
&\leq (1-\alpha)\,\|\pi - \pi_\epsilon\|_{L_2(\pi)} + \epsilon\,\|\pi_\epsilon\|_{L_2(\pi)} \\
&= (1-\alpha)\sqrt{\|\pi_\epsilon\|_{L_2(\pi)}^2 - 1} + \epsilon\,\|\pi_\epsilon\|_{L_2(\pi)}
\end{aligned}$$

Collecting the square roots and squaring both sides yields

$$\alpha^2\left(\|\pi_\epsilon\|_{L_2(\pi)}^2 - 1\right) \leq \epsilon^2\,\|\pi_\epsilon\|_{L_2(\pi)}^2$$

which implies that

$$\|\pi_\epsilon\|_{L_2(\pi)}^2 \leq \frac{\alpha^2}{\alpha^2 - \epsilon^2}$$

Finally, the second upper bound is derived from the first one, again using Lemma 1.3:

$$\|\pi - \pi_\epsilon\|_{L_2(\pi)}^2 = \|\pi_\epsilon\|_{L_2(\pi)}^2 - 1 \leq \frac{\alpha^2}{\alpha^2 - \epsilon^2} - 1 = \frac{\epsilon^2}{\alpha^2 - \epsilon^2}$$

□

We next observe that our assumptions imply that for small enough perturbations, the perturbed chain P_ϵ is geometrically ergodic in the $L_2(\pi)$ norm.

Lemma 1.7. *Under the assumptions of Section 1.3.2, if $\epsilon < \alpha$, then we have that P_ϵ is $L_2(\pi)$-geometrically ergodic, with factor $\leq 1 - (\alpha - \epsilon)$.*

Proof. Suppose that $\nu \in L_{2,0}(\pi)$. Then

$$\begin{aligned}\|\nu P_\epsilon\|_{L_2(\pi)} &\leq \|\nu(P_\epsilon - P)\|_{L_2(\pi)} + \|\nu P\|_{L_2(\pi)} \\ &\leq \epsilon \|\nu\|_{L_2(\pi)} + (1-\alpha) \|\nu\|_{L_2(\pi)} \\ &= (1 - \alpha + \epsilon) \|\nu\|_{L_2(\pi)} .\end{aligned}$$

Thus, for any probability measure $\mu \in L_2(\pi)$, since $\pi_\epsilon \in L_2(\pi)$ we have

$$\begin{aligned}\|\mu P_\epsilon^n - \pi_\epsilon\|_{L_2(\pi)} &= \|(\mu - \pi_\epsilon) P_\epsilon^n\|_{L_2(\pi)} \\ &\leq (1 - (\alpha - \epsilon))^n \|\mu - \pi_\epsilon\|_{L_2(\pi)} .\end{aligned}$$

□

Combining Lemmas 1.5 to 1.7 together with the triangle inequality immediately yields Theorem 1.1.

1.5.2 Proofs of Theorem 1.2, Theorem 1.3 and Corollary 1.1

Definition 1.3. Following [96], a subset $S \subset \mathcal{X}$ is called *hyper-small* for the π-irreducible Markov kernel P with stationary measure π if $\pi(S) > 0$ and there exists $\delta_S > 0$ and $k \in \mathbb{N}$ such that $\frac{\mathrm{d}P^k(x,\cdot)}{\mathrm{d}\pi}(y) \geq \delta_S \mathbf{1}_S(x) \mathbf{1}_S(y)$ or equivalently $P^k(x, A) \geq \delta_S \pi(A)$ for all $x \in S$ and $A \subset S$ measurable.

Lemma 4 of Jain and Jamison [48] states that on a countably generated state space (as we have assumed herein), every set of positive π-measure contains a hyper-small subset.

Lemma 1.8 (Existence of Hyper-Small Subsets from [48])**.** *Suppose that $(\mathcal{X}, \Sigma)$ is countably generated. Suppose that X is a ϕ-irreducible Markov chain on $\mathcal{X}$ with kernel P for some σ-finite measure ϕ on $\mathcal{X}$. Then any set $K \subset \mathcal{X}$ with $\phi(K) > 0$ contains a set S_K such that (for some $n_K \in \mathbb{N}$)*

$$\inf_{(x,y) \in S_K \times S_K} \frac{\mathrm{d}P^{n_K}(x,\cdot)}{\mathrm{d}\pi}(y) = \delta > 0$$

In the case that a stationary distribution, π, for P exists, without loss of generality we can take $\phi = \pi$. In this case, it is immediate that any set (S_K, n_K) satisfying Lemma 1.8 also satisfies Definition 1.3.

Also of importance to us is the following variant of Proposition 2.1 of [96], which provides a characterization of geometric ergodicity in terms of convergence to a hyper-small set.

Proposition 1.3 (Equivalent characterizations of π-a.e.-TV geometric ergodicity from [96] and Nummelin and Tweedie [86])**.** *Suppose that (Ω, Σ) is countably generated, and that X is a a ϕ-irreducible Markov chain on $\mathcal{X}$ with kernel P with stationary distribution π. Then the following are equivalent:*

(i) There exists $\rho_{TV} \in (0,1)$ such that P is π-a.e.-TV geometrically ergodic with factor ρ_{TV}

(i′) There exists a hyper-small set $S \subset \mathcal{X}$, and constants $\rho_S < 1$, $C_S \in \mathbb{R}_+$ such that:

$$\left\| \int \frac{\mathbf{1}_S(y)\pi(\mathrm{d}y)}{\pi(S)} P^n(y,\cdot) - \pi \right\|_{TV} \leq C_S \rho_S^n \qquad \forall n \in \mathbb{N}$$

(ii) There exists a π-a.e. finite, measurable function $V : \mathcal{X} \to [1,\infty]$ with $\pi(V^2) < \infty$, and $\rho_V \in (0,1)$, and $C > 0$ such that:

$$2 \left\| \delta_x P^n - \pi \right\|_{TV} \leq \left\| \delta_x P^n - \pi \right\|_V \leq C V(x) \rho_V^n$$

where $\|\mu\|_V = \sup_{|f|\leq V} |\mu(f)|$.

Proof of Theorem 1.2. (i) Let S be a hyper-small set for P_ϵ (which exists from Lemma 1.8, since P_ϵ is π_ϵ-irreducible). Then the measure μ_S defined by $\frac{\mathrm{d}\mu_S}{\mathrm{d}\pi} = \frac{\mathbb{1}_S}{\pi_\epsilon(S)} \frac{\mathrm{d}\pi_\epsilon}{\mathrm{d}\pi}$ has (by Hölder's inequality, and since $\pi_\epsilon \in L_2(\pi)$) that

$$\|\mu_S\|_{L_2(\pi)}^2 \leq \|\pi_\epsilon\|_{L_2(\pi)}^2 \pi_\epsilon(S)^{-2} < \infty,$$

and hence $\mu_S \in L_2(\pi)$. Then (by Cauchy-Schwarz again):

$$\left\| \int \frac{\mathbf{1}_S(y)\pi_\epsilon(\mathrm{d}y)}{\pi_\epsilon(S)} P_\epsilon^n(y,\cdot) - \pi_\epsilon \right\|_{\mathrm{TV}} \leq \frac{1}{2} \left\| \mu_S P_\epsilon^n - \pi_\epsilon \right\|_{L_2(\pi)} \leq \left\| \mu_S - \pi_\epsilon \right\|_{L_2(\pi)} (1-\alpha+\epsilon)^n$$

which, along with Proposition 1.3, establishes that P_ϵ is π_ϵ-a.e.-TV geometrically ergodic with some factor $\rho_{\mathrm{TV}} \in (0,1)$.

(ii) Suppose that $\mu \in L_\infty(\pi_\epsilon)$. Then $\mu \in L_2(\pi)$ since $\frac{\mathrm{d}\mu}{\mathrm{d}\pi} \leq \|\mu\|_{L_\infty(\pi_\epsilon)} \frac{\mathrm{d}\pi_\epsilon}{\mathrm{d}\pi}$. Since $\mu P_\epsilon^n - \pi_\epsilon \in L_1(\pi_\epsilon) \subset L_1(\pi)$ then

$$\|\mu P_\epsilon^n - \pi_\epsilon\|_{L_1(\pi_\epsilon)} = \|\mu P_\epsilon^n - \pi_\epsilon\|_{L_1(\pi)} = 2\,\|\mu P_\epsilon^n - \pi_\epsilon\|_{\mathrm{TV}}\,. \tag{1.25}$$

Applying this equality as well as Cauchy-Schwarz we get

$$\begin{aligned}\|\mu P_\epsilon^n - \pi_\epsilon\|_{L_1(\pi_\epsilon)} &= \|\mu P_\epsilon^n - \pi_\epsilon\|_{L_1(\pi)} \\ &\leq \|\mu P_\epsilon^n - \pi_\epsilon\|_{L_2(\pi)} \\ &\leq \|\mu - \pi_\epsilon\|_{L_2(\pi)}\,(1-\alpha-\epsilon)^n\end{aligned} \tag{1.26}$$

(iii) If $\pi \in L_\infty(\pi_\epsilon)$ and $\mu \in L_2(\pi_\epsilon)$ then

$$\begin{aligned}\|\mu P_\epsilon^n - \pi_\epsilon\|^2_{L_2(\pi_\epsilon)} &= \int \left(\frac{\mathrm{d}\mu P_\epsilon^n - \pi_\epsilon}{\mathrm{d}\pi_\epsilon}\right)^2 d\pi_\epsilon \\ &= \int \left(\frac{\mathrm{d}\mu P_\epsilon^n - \pi_\epsilon}{\mathrm{d}\pi}\right)^2 \frac{\mathrm{d}\pi}{\mathrm{d}\pi_\epsilon} d\pi \\ &\leq \|\pi\|_{L_\infty(\pi_\epsilon)} \int \left(\frac{\mathrm{d}\mu P_\epsilon^n - \pi_\epsilon}{\mathrm{d}\pi}\right)^2 d\pi \\ &= \|\pi\|_{L_\infty(\pi_\epsilon)}\,\|\mu P_\epsilon^n - \pi_\epsilon\|^2_{L_2(\pi)} \\ &\leq \|\pi\|_{L_\infty(\pi_\epsilon)}\,\|\mu - \pi_\epsilon\|^2_{L_2(\pi)}\,(1-(\alpha-\epsilon))^{2n}\end{aligned} \tag{1.27}$$

□

Proof of Theorem 1.3. From Baxter and Rosenthal [10, Lemma 1], since P_ϵ has stationary measure π_ϵ, then $P_\epsilon : L_2(\pi_\epsilon) \to L_2(\pi_\epsilon)$. Since P_ϵ is $(L_\infty(\pi_\epsilon), \|\cdot\|_{L_1(\pi_\epsilon)})$-GE with factor $\rho_1 \leq (1-(\alpha-\epsilon))$ (as established by Theorem 1.2) and P_ϵ is reversible, then it must also be $L_2(\pi_\epsilon)$-geometrically ergodic with factor $\rho = \rho_1$ by Lemma 1.2. □

Proof of Corollary 1.1. Note that the assumption that $\|P - P_\epsilon\|_{L_2(\pi_\epsilon)} < \varphi$ implies $P - P_\epsilon : L_2(\pi_\epsilon) \to L_2(\pi_\epsilon)$.

(i) Since P_ϵ is $L_2(\pi_\epsilon)$-geometrically ergodic with factor $(1-(\alpha-\epsilon))$ and π_ϵ-reversible, we can reverse the roles of P and P_ϵ, so the result follows by Theorem 1.1.

(ii) Taking $\mu = \pi$ and $n = 1$ in Theorem 1.2 *(iii)*,

$$\begin{aligned} \|\pi P_\epsilon - \pi_\epsilon\|^2_{L_2(\pi_\epsilon)} &\le \|\pi\|_{L_\infty(\pi_\epsilon)} \|\pi - \pi_\epsilon\|^2_{L_2(\pi)} (1-(\alpha-\epsilon))^2 \\ &\le \|\pi\|_{L_\infty(\pi_\epsilon)} \frac{\epsilon^2}{\alpha^2-\epsilon^2}(1-(\alpha-\epsilon))^2 \end{aligned} \tag{1.28}$$

Hence,

$$\begin{aligned} \|\pi - \pi_\epsilon\|_{L_2(\pi_\epsilon)} &\le \|\pi P - \pi P_\epsilon\|_{L_2(\pi_\epsilon)} + \|\pi P_\epsilon - \pi_\epsilon\|_{L_2(\pi_\epsilon)} \\ &\le \varphi \|\pi\|_{L_2(\pi_\epsilon)} + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon)) \\ &= \varphi\sqrt{\|\pi-\pi_\epsilon\|^2_{L_2(\pi_\epsilon)}+1} + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon)) \\ &\le \varphi(\|\pi-\pi_\epsilon\|_{L_2(\pi_\epsilon)}+1) + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon)) \end{aligned} \tag{1.29}$$

Hence,

$$\|\pi - \pi_\epsilon\|_{L_2(\pi_\epsilon)} \le \frac{\varphi + \|\pi\|^{1/2}_{L_\infty(\pi_\epsilon)} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}(1-(\alpha-\epsilon))}{1-\varphi} \tag{1.30}$$

Finally,

$$\|\mu P^n_\epsilon - \pi\|_{L_2(\pi_\epsilon)} \le \|\mu P^n_\epsilon - \pi_\epsilon\|_{L_2(\pi_\epsilon)} + \|\pi_\epsilon - \pi\|_{L_2(\pi_\epsilon)},$$

The first term is bounded by Theorem 1.2 *(iii)*, and the second term is bounded by Eq. (1.30)

□

1.5.3 Proofs of Theorem 1.4 and Theorem 1.5

1.5.3.1 Time-Averaging of Marginal Distributions

Proof of Theorem 1.4. The first result of Theorem 1.4 follows from the triangle inequality and Theorem 1.1,

$$\begin{aligned} \left\|\pi - \frac{1}{t}\sum_{k=0}^{t-1}\mu P^k_\epsilon\right\|_{L_2(\pi)} &\le \frac{1}{t}\sum_{k=0}^{t-1}\left\|\pi - \mu P^k_\epsilon\right\|_{L_2(\pi)} \\ &\le \frac{1}{t}\sum_{k=0}^{t-1}\left[(1-(\alpha-\epsilon))^k \|\pi_\epsilon - \mu\|_{L_2(\pi)} + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\right] \\ &\le \frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)} \|\pi_\epsilon - \mu\|_{L_2(\pi)} + \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}} . \end{aligned}$$

The subsequent results follows from similarly via Theorems 1.2 and 1.3 and Corollary 1.1.

□

1.5.3.2 Covariance Bounds

We turn our attention to the covariance structure of the original and perturbed chains. There is an obvious isometric isomorphism between the space of measures $L_2(\pi)$ and the function space

$$L_2'(\pi) = \left\{ f : \mathcal{X} \to \mathbb{R} \text{ s.t. } \int f(x)^2 \pi(dx) < \infty \right\}$$

equipped with the norm $\|f\|^2_{L_2'(\pi)} = \int f(x)^2 \pi(dx)$ where a measure μ is mapped to its Radon–Nikodym derivative $\mu \mapsto \frac{\mathrm{d}\mu}{\mathrm{d}\pi}$. For this reason, we need not distinguish between these spaces, and when dealing with a function $f \in L_2'(\pi)$ we may occasionally abuse notation and treat it as its associated measure. Let X_t and X_t^ϵ denote the original and perturbed chains run from some initial measure $\mu \in L_2(\pi)$.

Corollary 1.4. *Under the assumptions of Section 1.3.2,*

(a) if $X_0 \sim \pi$ (the initial distribution is the stationary distribution), then for $f, g \in L_2'(\pi)$

$$\operatorname{Cov}[f(X_t), g(X_s)] \leq (1-\alpha)^{|t-s|} \|f - \pi f\|_{L_2'(\pi)} \|g - \pi g\|_{L_2'(\pi)} \ , \tag{1.31}$$

(b) if $\epsilon < \alpha$, and P_ϵ is π_ϵ-reversible, $\rho_2 = (1-(\alpha-\epsilon))$, and $X_0^\epsilon \sim \pi_\epsilon$, then for $f, g \in L_2'(\pi_\epsilon)$

$$\operatorname{Cov}[f(X_t^\epsilon), g(X_s^\epsilon)] \leq \rho_2^{|t-s|} \|f - \pi_\epsilon f\|_{L_2'(\pi_\epsilon)} \|g - \pi_\epsilon g\|_{L_2'(\pi_\epsilon)} \ , \tag{1.32}$$

where for a function $h : \mathcal{X} \to \mathbb{R}$, πh is the constant function equal to $\int h(s)\pi(ds)$ everywhere.

Proof. The proof of this result follows that of Corollary B.5 in [52]. We only show the proof for the original chain, however the proof for the perturbed chain is the same, since it is reversible and $L_2(\pi_\epsilon)$ geometrically ergodic with the appropriate factor, from Theorem 1.3.

Define the subspace

$$L_{2,0}'(\pi) = \{h \in L_2'(\pi) : \int h(s)\pi(ds) = 0\} \ ,$$

and the operator $F \in \mathcal{B}(L'_{2,0}(\pi))$ by

$$[Ff](x) = \int P(x, dy)f(y) = \mathbb{E}[f(X_1)|X_0 = x]$$

From Lemma 12.6.4 of Liu [65],

$$\sup_{f,g \in L'_2(\pi)} \operatorname{corr}(f(X_0), g(X_t)) = \sup_{\substack{\|f\|_{L'_2(\pi)}=1=\|g\|_{L'_2(\pi)} \\ f,g \in L'_{2,0}(\pi)}} \langle f, F^t g\rangle = \left\|F^t\right\|_{L'_{2,0}(\pi)}$$

Consider the canonical isomorphism between $L_2(\pi)$ and $L'_2(\pi)$. The restriction of this isomorphism (on the right) to elements of $L'_{2,0}(\pi)$ yields $L_{2,0}(\pi)$ (on the left) – the signed measures with total measure 0. The image of F under the restricted isomorphism is the adjoint operator of P restricted to $L_{2,0}(\pi)$. Since P is π-reversible, it is self-adjoint, in $L_2(\pi)$ so $\|F\|_{L'_{2,0}(\pi)} = \|P\|_{L_{2,0}(\pi)}$.

$$\left\|F^t\right\|_{L'_{2,0}(\pi)} \le \|F\|^t_{L'_{2,0}(\pi)} = \left\|P|_{L_{2,0}(\pi)}\right\|^t \le (1-\alpha)^t$$

Therefore

$$\operatorname{Cov}(f(X_0), g(X_t)) \le \|f - \pi f\|_{L'_2(\pi)} \|g - \pi g\|_{L'_2(\pi)} (1-\alpha)^t$$

Since Cov is symmetric, the shifted and symmetrized result holds for any $f, g \in L'_2(\pi)$:

$$\operatorname{Cov}[f(X_t), g(X_s)] \le (1-\alpha)^{|t-s|} \|f - \pi f\|_{L'_2(\pi)} \|g - \pi g\|_{L'_2(\pi)} \tag{1.33}$$

□

We present further bounds for the case that the initial distribution is not the stationary distribution in Corollary 1.5.

Remark 1.8. Note in Corollary 1.4 that

$$\|h - \pi h\|_{L'_2(\pi)} = \sqrt{\|h\|^2_{L'_2(\pi)} - (\pi h)^2} \le \|h\|_{L'_2(\pi)} \ .$$

Also note that

$$\|h\|_{L_2'(\pi)} \leq \|h - \pi(h)\|_{L_2'(\pi)} + |\pi(h)| \ . \tag{1.34}$$

$\triangleleft$

Corollary 1.5. *Under the assumptions of Section 1.3.2,*

(a) if $X_0 \sim \mu$, then for $f, g \in L_4'(\pi)$

$$\begin{aligned}
&\operatorname{Cov}(f(X_t), g(X_{t+s})) \\
&\quad \leq (1-\alpha)^s \|f - \pi f\|_{L_2'(\pi)} \|g - \pi g\|_{L_2'(\pi)} \\
&\qquad + 2^{3/2}(1-\alpha)^{t+s/2} \|\mu - \pi\|_{L_2(\pi)} \|f - \pi f\|_{L_4'(\pi)} \|g - \pi g\|_{L_4'(\pi)} \\
&\qquad - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right)
\end{aligned}$$

(b) if $\epsilon < \alpha$, and P_ϵ is π_ϵ-reversible, $\rho_2 = (1-(\alpha-\epsilon))$, and $X_0^\epsilon \sim \mu$, then for $f, g \in L_4'(\pi_\epsilon)$

$$\begin{aligned}
&\operatorname{Cov}(f(X_t^\epsilon), g(X_{t+s}^\epsilon)) \\
&\quad \leq \rho_2^s \|f - \pi_\epsilon f\|_{L_2'(\pi_\epsilon)} \|g - \pi_\epsilon g\|_{L_2'(\pi_\epsilon)} \\
&\qquad + 2^{3/2}\rho_2^{t+s/2} \|\mu - \pi\|_{L_2(\pi_\epsilon)} \|f - \pi_\epsilon f\|_{L_4'(\pi_\epsilon)} \|g - \pi_\epsilon g\|_{L_4'(\pi_\epsilon)} \\
&\qquad - (\mu P_\epsilon^t f - \pi_\epsilon f)\left(\mu P_\epsilon^{t+s} g - \pi_\epsilon g\right)
\end{aligned}$$

Proof. This will use the following shorthand notation. Let

$$\begin{aligned}
f_0 &= f - \pi f \\
g_0 &= g - \pi g \\
\|h\|_\star &= \left(\int (h(x) - \pi h)^2 \pi(dx)\right)^{1/2} \\
\|h\|_{\star\star} &= \left(\int (h(x) - \pi h)^4 \pi(dx)\right)^{1/4} \\
C_\mu &= \|\mu - \pi\|_2
\end{aligned}$$

$\|\cdot\|_{\star\star}$ can be interpreted as a centred 4-norm. It is certainly bounded above by $\|\cdot\|_4$, the norm on $L_4'(\pi)$. For some results regarding the properties of a Markov transition kernel as an operator on $L_p'(\pi)$ for general p given an L_2-spectral gap (as is implied by L_2-geometric ergodicity) please refer to Rudolf [103].

We only show the proof for the original chain. The result for the perturbed chain has essentially the same proof.

By definition we can express the covariance by the triple integral below. We re-express this integral as a sum of two integrals involving the chain run from stationarity. This will allow us to apply Corollary 1.4.

$$
\begin{aligned}
&\mathrm{Cov}(f(X_t), g(X_{t+s})) \\
&= \iiint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g)\mu(dx)P^t(x, dy)P^s(y, dz) \\
&= \iiint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g)\left[\frac{d\mu}{d\pi}(x) - 1\right]\pi(dx)P^t(x, dy)P^s(y, dz) \\
&\quad + \iiint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g)\pi(dx)P^t(x, dy)P^s(y, dz)
\end{aligned}
$$

We will simplify each of these expressions separately, starting with the second term:

$$
\begin{aligned}
&\iiint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g)\pi(dx)P^t(x, dy)P^s(y, dz) \\
&\quad = \iint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g)\pi(dy)P^s(y, dz) \\
&\quad = \iint f(y)g(z)\pi(dy)P^s(y, dz) \\
&\qquad - (\mu P^t f)(\pi g) - (\pi f)(\mu P^{t+s} g) + (\mu P^t f)(\mu P^{t+s} g) \\
&\quad = \iint f_0(y)g_0(z)\pi(dy)P^s(y, dz) + (\pi f)(\pi g) \\
&\qquad - (\mu P^t f)(\pi g) - (\pi f)(\mu P^{s+t} g) + (\mu P^t f)(\mu P^{t+s} g) \\
&\quad = \langle f_0, F^s g_0\rangle + (\mu P^t f - \pi f)(\mu P^{s+t} g - \pi g)
\end{aligned}
$$

For the first term we find that:

$$
\begin{aligned}
&\iiint (f(y) - \mu P^t f)(g(z) - \mu P^{t+s} g) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^t(x, dy) P^s(y, dz) \\
&\quad = \iiint f(y) g(z) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^t(x, dy) P^s(y, dz) \\
&\qquad - (\mu P^t f) \iint g(z) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^{t+s}(x, dz) \\
&\qquad - (\mu P^{s+t} g) \iint f(y) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^t(x, dy) \\
&\qquad + (\mu P^t f)(\mu P^{s+t} g) \int \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) \\
&\quad = \iiint f_0(y) g_0(z) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^t(x, dy) P^s(y, dz) \\
&\qquad - (\mu P^t f - \pi f) \iint g(z) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^{t+s}(x, dz) \\
&\qquad - (\mu P^{s+t} g - \pi g) \iint f(y) \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) P^t(x, dy) \\
&\qquad - (\pi f)(\pi g) \int \left(\frac{d\mu}{d\pi}(x) - 1\right) \pi(dx) \\
&\quad = \left\langle \frac{d\mu}{d\pi} - 1, F^t(f_0 \otimes (F^s g_0)) \right\rangle \\
&\qquad - (\mu P^t f - \pi f) \left\langle \frac{d\mu}{d\pi} - 1, F^{t+s} g \right\rangle - (\mu P^{t+s} g - \pi g) \left\langle \frac{d\mu}{d\pi} - 1, F^t f \right\rangle \\
&\quad = \left\langle \frac{d\mu}{d\pi} - 1, F^t(f_0 \otimes (F^s g_0)) \right\rangle - 2(\mu P^t f - \pi f) \left(\mu P^{t+s} g - \pi g\right)
\end{aligned}
$$

Where $f_0 \otimes F^s g_0$ is defined by

$$
[f_0 \otimes F^s g_0](y) = f_0(y) \int g_0(z) P^s(y, dz)
$$

Putting these together,

$$
\begin{aligned}
&\mathrm{Cov}(f(X_t), g(X_{t+s})) \\
&\quad = \langle f_0, F^s g_0\rangle + (\pi f - \mu P^t f)(\pi g - \mu P^{s+t} g) \\
&\qquad + \left\langle \frac{d\mu}{d\pi} - 1, F^t(f_0 \otimes (F^s g_0))\right\rangle - 2(\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right) \\
&\quad = \langle f_0, F^s g_0\rangle + \left\langle \frac{d\mu}{d\pi} - 1, F^t(f_0 \otimes (F^s g_0))\right\rangle - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right) \\
&\quad \le (1-\alpha)^s \|f\|_\star \|g\|_\star + (1-\alpha)^t \|\mu - \pi\|_2 \|f_0 \otimes F^s g_0\|_2 \\
&\qquad - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right) \\
&\quad \le (1-\alpha)^s \|f\|_\star \|g\|_\star + (1-\alpha)^t \|\mu - \pi\|_2 \|f_0\|_4 \|F^s g_0\|_4 \\
&\qquad - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right) \\
&\quad \le (1-\alpha)^s \|f\|_\star \|g\|_\star + (1-\alpha)^t \|\mu - \pi\|_2 \|f_0\|_4 \|g_0\|_4 \left\| F^s|_{L'_{4,0}} \right\|_4 \\
&\qquad - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right) \\
&\quad \le (1-\alpha)^s \|f\|_\star \|g\|_\star + 2^{3/2}(1-\alpha)^{t+s/2} \|\mu - \pi\|_2 \|f\|_{\star\star} \|g\|_{\star\star} \\
&\qquad - (\mu P^t f - \pi f)\left(\mu P^{t+s} g - \pi g\right)
\end{aligned}
$$

The $\langle f_0, F^s g_0\rangle$ term is bounded using Corollary 1.4 where we have taken the result in its equivalent form using the $\langle \cdot, \cdot \rangle$ notation and the forward operator F. Next, the

$$
\left\langle \frac{d\mu}{d\pi} - 1, F^t(f_0 \otimes (F^s g_0))\right\rangle
$$

term is bounded following the methodology of the proof of [103], Lemma 3.39 (in order the inequalities are: Cauchy-Schwarz, $\|F^s g_0\| \le \|F^s\|\|g_0\|$ for any norm $\|\cdot\|$, and Proposition 3.17 of [103]). $\square$

The main motivation in establishing the covariance bounds in Corollaries 1.4 and 1.5 is that we will need to sum up covariances in order to establish bounds on the variance component of mean-squared error for estimation of $\pi(f)$ via the dependent sample means $\frac{1}{t}\sum_{j=0}^{t-1} f(X_j)$ and $\frac{1}{t}\sum_{j=0}^{t-1} f(X_j^\epsilon)$ for an arbitrary starting measure. To this end we will be interested in the following summation result.

Corollary 1.6. *Under the assumptions of Section 1.3.2,*

(a) if $X_0 \sim \mu$, then for $f, g \in L_4'(\pi)$

$$\frac{1}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}\mathrm{Cov}(f(X_j), f(X_k)) \\ \leq \frac{2\left\|f-\pi f\right\|_{L_2'(\pi)}^2}{\alpha t} + \frac{2^{7/2}\left\|\mu-\pi\right\|_{L_2(\pi)}\left\|f-\pi f\right\|_{L_4'(\pi)}^2}{\alpha^2 t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}\mu P^m f - \pi f\right)^2$$

(b) if $\epsilon < \alpha$, and P_ϵ is π_ϵ-reversible, $\rho_2 = (1-(\alpha-\epsilon))$, and $X_0^\epsilon \sim \mu$, then for $f, g \in L_4'(\pi_\epsilon)$

$$\frac{1}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}\mathrm{Cov}(f(X_j^\epsilon), f(X_k^\epsilon)) \\ \leq \frac{2\left\|f-\pi f\right\|_{L_2'(\pi_\epsilon)}^2}{(1-\rho_2)t} + \frac{2^{7/2}\left\|\mu-\pi\right\|_{L_2(\pi_\epsilon)}\left\|f-\pi f\right\|_{L_4'(\pi_\epsilon)}^2}{(1-\rho_2)^2 t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}\mu P_\epsilon^m f - \pi f\right)^2$$

Proof. We only show the proof for the original chain. The results for the perturbed chain have essentially the same proof. The proof is largely an exercise in summation of geometric series and meticulous bookkeeping. The first inequality is due to Corollary 1.5. The second inequality makes use of the fact $0 < \alpha < 1$. To simplify notation, $C_\mu = \left\|\mu-\pi\right\|_{L_2(\pi)}$.

$$
\begin{aligned}
&\frac{1}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}\mathrm{Cov}(f(X_j),f(X_k)) \\
&\quad= \frac{\|f\|_\star^2}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}(1-\alpha)^{|m-n|} - \frac{1}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}(\mu P^m f - \pi f)\,(\mu P^n f - \pi f) \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\sum_{m=0}^{t-1}\sum_{n=0}^{t-1}(1-\alpha)^{(m+n)/2} \\
&\quad= \frac{\|f\|_\star^2}{t^2}\sum_{m=0}^{t-1}\left(1+2\sum_{s=1}^{t-m-1}(1-\alpha)^s\right) - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2 \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\sum_{m=0}^{t-1}(1-\alpha)^m\left(1+2\sum_{s=1}^{t-m-1}(1-\alpha)^{s/2}\right) \\
&\quad= \frac{\|f\|_\star^2}{t^2}\sum_{m=0}^{t-1}\left(1+2\frac{(1-\alpha)-(1-\alpha)^{t-m}}{\alpha}\right) - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2 \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\sum_{m=0}^{t-1}(1-\alpha)^m\left(1+2\frac{\sqrt{1-\alpha}-\sqrt{1-\alpha}^{t-m}}{1-\sqrt{1-\alpha}}\right) \\
&\quad= \frac{\|f\|_\star^2}{t^2}\sum_{m=0}^{t-1}\left(\frac{2-\alpha}{\alpha}-\frac{2}{\alpha}(1-\alpha)^{t-m}\right) - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2 \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\sum_{m=0}^{t-1}\left((1-\alpha)^m\frac{1+\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}} - 2\frac{\sqrt{1-\alpha}^{t+m}}{1-\sqrt{1-\alpha}}\right) \\
&\quad= \frac{\|f\|_\star^2}{t^2}\left(\frac{2-\alpha}{\alpha}t - \frac{2}{\alpha}\frac{(1-\alpha)-(1-\alpha)^{t+1}}{\alpha}\right) - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2 \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\left(\left[\frac{1+\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}}\right]\left[\frac{1-(1-\alpha)^t}{\alpha}\right]\right. \\
&\qquad\qquad\qquad \left. - \left[\frac{2\sqrt{1-\alpha}^t}{1-\sqrt{1-\alpha}}\right]\left[\frac{1-\sqrt{1-\alpha}^t}{1-\sqrt{1-\alpha}}\right]\right) \\
&\quad= (2-\alpha)\frac{\|f\|_\star^2}{\alpha t} - 2(1-\alpha)\frac{1-(1-\alpha)^t}{\alpha^2 t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2 \\
&\qquad + \frac{2^{3/2}C_\mu\|f\|_{\star\star}^2}{t^2}\left(\frac{1+\sqrt{1-\alpha}}{\alpha}\right)^2(1-(1-\alpha)^{t/2})^2 \\
&\quad\le \frac{2\|f\|_\star^2}{\alpha t} + \frac{2^{7/2}C_\mu\|f\|_{\star\star}^2}{\alpha^2 t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}(\mu P^m f - \pi f)\right)^2
\end{aligned}
$$

$\square$

1.5.3.3 Mean Squared Error Bonds

Theorem 1.7. *Under the assumptions of Section 1.3.2, if $X_0 \sim \mu \in L_2(\pi)$, then*

$$\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k)\right)^2\right] \leq \frac{2\|f-\pi f\|_2^2}{\alpha t} + \frac{2^{7/2}\|\mu-\pi\|_2\|f-\pi f\|_4^2}{\alpha^2 t^2}$$

Proof. The proof proceeds by partitioning the MSE via the bias-variance decomposition then bounding variance term and noting that our bond for the variance contains an expression which exactly cancels the bias term. We compute that

$$\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k)\right)^2\right] \\
&= \mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f) - \frac{1}{t}\sum_{k=0}^{t-1}(f(X_k) - [\mu P^k](f))\right)^2\right] \\
&= \left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f)\right)^2 + \mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1}(f(X_k) - [\mu P^k](f))\right)^2\right] \\
&= \left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P^k](f)\right)^2 + \frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathrm{Cov}(f(X_j), f(X_k))
\end{aligned}$$

The variance term is bounded using Corollary 1.6:

$$\begin{aligned}
&\frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathrm{Cov}(f(X_j), f(X_k)) \\
&\qquad \frac{2\|f-\pi f\|_2^2}{\alpha t} + \frac{2^{7/2}\|\mu-\pi\|_2\|f-\pi f\|_4^2}{\alpha^2 t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}\mu P^m f - \pi f\right)^2
\end{aligned}$$

Putting these together yields the desired result. $\square$

Remark 1.9. We note that, as per Remark 1.8, $\|f-\pi f\| \leq \|f\|_2$. Similarly $\|f-\pi f\|_4 \leq \|f\|_4$. Also in the case that f is is π-essentially bounded, $\|f\|_2 \leq \|f\|_\infty$ and $\|f\|_4 \leq \|f\|_\infty$. These alternative norms may be substituted into the result as necessary in order to make the bounds tractable for a given application. $\triangleleft$

Remark 1.10. Comparing our above geometrically ergodic results to the L_1 results of [52] in the uniformly ergodic case, we see that the L_2 and L_1 bounds we establish above differ from the corresponding L_1 bound of [52] only by a factor, which is constant in time, but varies with the initial distribution (as is to be expected when moving from uniform ergodicity to geometric ergodicity). For the Mean-Squared-Error results, the $\|\cdot\|_\star$-norm in that work is based on the midrange-centred infinity norm, which as per Remark 1.9 is an upper bound on what we have. $\triangleleft$

Proof of Theorem 1.5. For the first result, we proceed via bias-variance decomposition, as in the corresponding result for the exact chain. However, now the bias under consideration is itself decomposed as the square of a sum of two components. The squared sum is expanded simultaneously with the bias-variance expansion. We compute that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \\
&= \mathbb{E}\left[\left(\pi(f) - \pi_\epsilon(f) + \frac{1}{t}\sum_{k=0}^{t-1}\left[\pi_\epsilon - \mu P_\epsilon^k\right](f) - \frac{1}{t}\sum_{k=0}^{t-1}(f(X_k^\epsilon) - [\mu P_\epsilon^k](f))\right)^2\right] \\
&= ([\pi - \pi_\epsilon](f))^2 + 2\left([\pi - \pi_\epsilon](f)\right)\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right) \\
&\qquad + \left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 + \frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathrm{Cov}(f(X_j^\epsilon), f(X_k^\epsilon))
\end{aligned}
$$

We bound the first component of the bias term using versions of Lemma 1.6

$$
\begin{aligned}
([\pi - \pi_\epsilon](f))^2 &= ([\pi - \pi_\epsilon](f - \pi_\epsilon f))^2 \\
&\leq \begin{cases} \|\pi - \pi_\epsilon\|_{L_2(\pi)}^2 \|f - \pi_\epsilon f\|_{L_2'(\pi)}^2 \\ \|\pi - \pi_\epsilon\|_{L_2(\pi_\epsilon)}^2 \|f - \pi_\epsilon f\|_{L_2'(\pi_\epsilon)}^2 \end{cases} \\
&\leq \begin{cases} \frac{\epsilon^2}{\alpha^2 - \epsilon^2} \|f - \pi_\epsilon f\|_{L_2'(\pi)}^2 \\ \frac{\varphi^2}{(1-\rho_2)^2 - \varphi^2} \|f - \pi_\epsilon f\|_{L_2'(\pi_\epsilon)}^2 & : \text{given } (*) \end{cases}
\end{aligned}
$$

We bound the variance term using Corollary 1.6:

$$\frac{1}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\mathrm{Cov}(f(X_j^\epsilon), f(X_k^\epsilon))$$
$$\leq \frac{2\,\|f-\pi_\epsilon f\|^2_{L_2'(\pi_\epsilon)}}{(1-\rho_2)t} + \frac{2^{7/2}\,\|\mu-\pi_\epsilon\|_{L_2(\pi_\epsilon)}\,\|f-\pi_\epsilon f\|_{L_4'(\pi_\epsilon)}}{(1-\rho_2)^2t^2} - \left(\frac{1}{t}\sum_{m=0}^{t-1}\mu P_\epsilon^m f - \pi_\epsilon f\right)^2$$

The negative term in this expression exactly cancels out the third bias term in the expansion.

Finally, we bound the second bias term using Lemma 1.6 and Theorem 1.4:

$$2\left([\pi-\pi_\epsilon](f)\right)\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)$$
$$= 2\left([\pi-\pi_\epsilon](f-\pi_\epsilon f)\right)\left(\left[\pi_\epsilon - \frac{1}{t}\sum_{k=0}^{t-1}\mu P_\epsilon^k\right](f-\pi_\epsilon f)\right)$$
$$\leq 2\begin{cases} \frac{\varphi}{\sqrt{(1-\rho_2)^2-\varphi^2}}\,\|f-\pi_\epsilon f\|_{L_2'(\pi)}\,\frac{1-(1-(\alpha-\epsilon))^t}{t(\alpha-\epsilon)}\,\|\pi_\epsilon-\mu\|_{L_2(\pi)}\,\|f-\pi_\epsilon f\|_{L_2'(\pi)} \\ \frac{\varphi^2}{(1-\rho_2)^2-\varphi^2}\,\|f-\pi_\epsilon f\|^2_{L_2'(\pi_\epsilon)}\,\frac{1-\rho_2^t}{t(1-\rho_2)}\,\|\pi_\epsilon-\mu\|_{L_2(\pi_\epsilon)}\,\|f-\pi_\epsilon f\|_{L_2'(\pi_\epsilon)} & : \text{given } (*) \end{cases}$$
$$\leq 2\begin{cases} \frac{\epsilon}{\sqrt{\alpha^2-\epsilon^2}}\,\frac{1}{t(\alpha-\epsilon)}\,\|\pi_\epsilon-\mu\|_{L_2(\pi)}\,\|f-\pi_\epsilon f\|^2_{L_2'(\pi)} \\ \frac{\varphi}{\sqrt{(1-\rho_2)^2-\varphi^2}}\,\frac{1}{t(1-\rho_2)}\,\|\pi_\epsilon-\mu\|_{L_2(\pi_\epsilon)}\,\|f-\pi_\epsilon f\|^2_{L_2'(\pi_\epsilon)} & : \text{given } (*) \end{cases}$$

Putting these together yields the first and third results.

For the second and fourth result we use the fact that for any random variable, Z, and for any $a, b \in \mathbb{R}$ the following holds:

$$\mathbb{E}[(Z-a)^2] = 2\mathbb{E}[(Z-b)^2] + 2(a-b)^2 - \mathbb{E}[(Z+a-2b)^2]$$
$$\leq 2\mathbb{E}[(Z-b)^2] + 2(a-b)^2$$

$$
\begin{aligned}
&\mathbb{E}\left[\left(\pi(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \\
&\quad \le 2([\pi - \pi_\epsilon](f))^2 + 2\mathbb{E}\left[\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1} f(X_k^\epsilon)\right)^2\right] \\
&\quad = 2([\pi - \pi_\epsilon](f - \pi_\epsilon f))^2 \\
&\qquad + 2\mathbb{E}\left[\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f) - \frac{1}{t}\sum_{k=0}^{t-1}(f(X_k^\epsilon) - [\mu P_\epsilon^k](f))\right)^2\right] \\
&\quad = 2([\pi - \pi_\epsilon](f - \pi_\epsilon f))^2 + 2\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 \\
&\qquad + 2\mathbb{E}\left[\left(\frac{1}{t}\sum_{k=0}^{t-1}(f(X_k^\epsilon) - [\mu P_\epsilon^k](f))\right)^2\right] \\
&\quad = 2([\pi - \pi_\epsilon](f - \pi_\epsilon f))^2 + 2\left(\pi_\epsilon(f) - \frac{1}{t}\sum_{k=0}^{t-1}[\mu P_\epsilon^k](f)\right)^2 \\
&\qquad + \frac{2}{t^2}\sum_{j=0}^{t-1}\sum_{k=0}^{t-1}\operatorname{Cov}(f(X_j^\epsilon), f(X_k^\epsilon))
\end{aligned}
$$

Applying Corollary 1.5 to bound the sum of covariances, we find that we are able to exactly cancel the second term in the final expression above. Using the same bound as before for the first expression, we get the final result. □

1.5.4 Proof of Theorem 1.6

Let

$$
\begin{aligned}
\gamma(x) &= \mathbb{E}_{y\sim q(y|x)} r(y|x) = \int r(y|x)q(y|x)dy \\
[\nu\Gamma](dy) &= \nu(y)\gamma(y)dy \\
[\nu Z](dy) &= \left[\int r(y|x)q(y|x)\nu(x)dx\right] dy
\end{aligned}
$$

Lemma 1.9. $P - \hat{P} = Z - \Gamma$

Proof. We first give expressions for the elements of measure for transitions of the original chain. The first formula is the element of measure for transition from an arbitrary, fixed initial point. It is defined for us by the mechanics of the Metropolis–Hastings algorithm. The second expression is the element of measure for transition from a sample from an initial distribution, ν. It is derived from the first expression by integrating over the sample from

ν.

$$
\begin{aligned}
P(x, dx') &= \delta_x(dx')\left[1 - \int (a(y|x)q(y|x)dy\right] + a(x'|x)q(x'|x)dx' \\
[\nu P](dx') &= \int \left[\delta_x(dx')\left[1 - \int a(y|x)q(y|x)dy\right] + a(x'|x)q(x'|x)dx'\right]\nu(x)dx \\
&= \left[\left[1 - \int a(y|x')q(y|x')dy\right]\nu(x') + \int a(x'|x)q(x'|x)\nu(x)dx\right]dx'
\end{aligned}
$$

The second form of the second expression is an application of Fubini's theorem. The exchange of the order of integration for the second term in the expression is immediate. For the first term, for arbitrary non-negative functions f,

$$
\int_s \int_t f(s,t)\delta_t(ds)dt = \int_t \int_s f(s,t)\delta_t(ds)dt = \int_t f(t,t)dt = \int_s f(s,s)ds
$$

Where the first equality is Fubini's theorem, the second comes from integrating with respect to s, and the third comes from a change of dummy variable.

Similarly, the elements of measure for transitions from the approximating kernel are expressed below. The first expression, as above, is the element of measure for transition from an arbitrary, fixed initial point. It is defined for us by the mechanics of the noisy Metropolis–Hastings algorithm. The second expression is again derived by integrating the first against an initial measure, ν.

$$
\begin{aligned}
\hat{P}(x, dx') &= \delta_x(dx')\left[1 - \iint \hat{a}(y|x,z)q(y|x)f_y(z)dzdy\right] \\
&\qquad + \int \hat{a}(x'|x,z)q(x'|x)f_{x'}(z)dzdx' \\
\left[\nu\hat{P}\right](dx') &= \int \Bigg(\delta_x(dx')\left[1 - \iint \hat{a}(y|x,z)q(y|x)f_y(z)dzdy\right] \\
&\qquad\qquad + \int \hat{a}(x'|x,z)q(x'|x)f_{x'}(z)dzdx'\Bigg)\nu(x)dx \\
&= \left[1 - \iint \hat{a}(y|x',z)q(y|x')f_y(z)dzdy\right]\nu(x')dx' \\
&\qquad + \left[\iint \hat{a}(x'|x,z)q(x'|x)f_{x'}(z)\nu(x)dzdx\right]dx'
\end{aligned}
$$

The same applications of Fubini's theorem occur as above.

We may now leverage our notation defined above to simplify the difference of these elements of measure.

$$\begin{aligned}
&\left[\nu(P-\hat{P})\right](dx') \\
&= \left[\iint \Big(\hat{a}(y|x',z) - a(y|x')\Big) q(y|x') f_y(z) dz dy\right] \nu(x') dx' \\
&\quad + \left[\iint \Big(a(x'|x) - \hat{a}(x'|x,z)\Big) q(x'|x) f_{x'}(z) \nu(x) dz dx\right] dx' \\
&= \left[\int r(x'|x) q(x'|x) \nu(x) dx\right] dx' - \left[\int r(y|x') q(y|x') dy\right] \nu(x') dx' \\
&= [\nu(Z-\Gamma)](dx')
\end{aligned}$$

From this one may conclude that $\left(P - \hat{P} = Z - \Gamma\right)$ as operators. □

Proof of Theorem 1.6. It is obvious that if $|r(y|x)| \leq R$ uniformly in $(x,y) \in \mathcal{X}^2$ then

$$\left(\|\Gamma\|_{L_2(\pi)} \leq R\right) , \tag{1.35}$$

and

$$\left(\|Z\|_{L_2(\pi)} \leq R \|Q\|_{L_2(\pi)}\right) . \tag{1.36}$$

By applying the previous lemma, given the assumptions stated,

$$\left\|P - \hat{P}\right\|_{L_2(\pi)} \leq R(1 + \|Q\|_{L_2(\pi)}) . \tag{1.37}$$

□

1.5.5 $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE is distinct from L_2-GE for non-reversible chains

Let $\mathcal{X} = \mathbb{N} \cup \{0\}$, and let a be a probability mass function on $\mathcal{X}$. Define transition probabilities by

$$p_{ij} = \begin{cases} a_j & : i = 0 \\ 1 & : i > 0,\ j = i-1 \\ 0 & : \text{otherwise} \end{cases} \tag{1.38}$$

Let $b_j = \sum_{i=j}^{\infty} a_i$. It is easy to verify that if $\sum_{j=1}^{\infty} b_j < \infty$ then $\pi_j = \frac{b_j}{\sum_{j=1}^{\infty} b_j}$ is the unique stationary probability mass function for $P = [p_{ij}]_{ij \in \mathcal{X}^2}$.

In the special case where $a_j = 2^{-j-1}$, we have $\pi = a$. We continue this example working exclusively with this choice of a. Now,

$$\delta_j P^n = \begin{cases} \pi & : n \geq j+1 \\ \delta_{n-j} & : n \leq j \end{cases} \tag{1.39}$$

Thus, for any initial probability mass function, μ,

$$[\mu P^n]_j = \sum_{i=0}^{n-1} \mu_i \pi_j + \mu_{j+n} \tag{1.40}$$

If $\frac{\mathrm{d}\mu}{\mathrm{d}\pi}(j) = \frac{\mu_j}{\pi_j} \leq \|\mu\|_{L_\infty(\pi)} < \infty$ for all $j \in \mathcal{X}$ then

$$\begin{aligned} \|\mu P^n - \pi\|_{L_2(\pi)}^2 &= \sum_{j=0}^{\infty} \pi_j \left(\sum_{i=0}^{n-1} \mu_i + \frac{\mu_{j+n}}{\pi_j} - 1 \right)^2 \\ &= \sum_{j=0}^{\infty} \pi_j \left(-\sum_{i=n}^{\infty} \mu_i + \frac{\mu_{j+n}}{\pi_{j+n}} \frac{\pi_{j+n}}{\pi_j} \right)^2 \\ &= \sum_{j=0}^{\infty} \pi_j \left(-\sum_{i=n}^{\infty} \frac{\mu_i}{\pi_i} \pi_i + \frac{\mu_{j+n}}{\pi_{j+n}} \frac{\pi_{j+n}}{\pi_j} \right)^2 \\ &\leq \sum_{j=0}^{\infty} \pi_j \left(\sum_{i=n}^{\infty} \frac{\mu_i}{\pi_i} \pi_i + \frac{\mu_{j+n}}{\pi_{j+n}} \frac{\pi_{j+n}}{\pi_j} \right)^2 \\ &\leq \sum_{j=0}^{\infty} 2^{-j-1} \left(\sum_{i=n}^{\infty} \|\mu\|_{L_\infty(\pi)} 2^{-i-1} + \|\mu\|_{L_\infty(\pi)} 2^{-n} \right)^2 \\ &= \|\mu\|_{L_\infty(\pi)}^2 \sum_{j=0}^{\infty} 2^{-j-1} (2^{-n+1})^2 \\ &= 4 \|\mu\|_{L_\infty(\pi)}^2 (2^{-n})^2 \end{aligned} \tag{1.41}$$

Hence P is $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE with optimal rate no larger than 1/2.

For any $\alpha < \sqrt{0.5}$, let $\nu_j = (1-\alpha)(\alpha)^j$. Then $\nu \in L_2(\pi)$, since

$$\begin{aligned}\|\nu\|^2_{L_2(\pi)} &= \sum_{i=0}^{\infty} 0.5^{i+1}\left(\frac{(1-\alpha)(\alpha)^i}{0.5^{i+1}}\right)^2 \\ &= 2(1-\alpha)^2\sum_{i=0}^{\infty}(2\alpha^2)^i = \frac{2(1-\alpha)^2}{1-2\alpha^2}\end{aligned} \tag{1.42}$$

Moreover,

$$\begin{aligned}\|\nu P^n - \pi\|^2_{L_2(\pi)} &= \sum_{j=0}^{\infty}\pi_j\left(\sum_{i=0}^{n-1}\nu_i + \frac{\nu_{j+n}}{\pi_j} - 1\right)^2 \\ &= \sum_{j=0}^{\infty} 0.5^{j+1}\left(-\sum_{i=n}^{\infty}(1-\alpha)\alpha^i + (1-\alpha)\alpha^{j+n}(0.5)^{-j-1}\right)^2 \\ &= \alpha^{2n}\sum_{j=0}^{\infty}0.5^{j+1}\left(-1+2(1-\alpha)(2\alpha)^j\right)^2 \\ &= \frac{\alpha^{2n}}{2}\sum_{j=0}^{\infty}(0.5^j - 4(1-\alpha)\alpha^j + 4(1-\alpha)^2(2\alpha^2)^j) \\ &= \frac{\alpha^{2n}}{2}\left(2 - \frac{4(1-\alpha)}{1-\alpha} + \frac{4(1-\alpha)^2}{1-2\alpha^2}\right) \\ &= \frac{(2\alpha-1)^2}{1-2\alpha^2}\alpha^{2n}\end{aligned} \tag{1.43}$$

Thus the convergence rate starting from this initial measure is α.

Since this is true for any $\alpha < 1/\sqrt{2}$, this shows that the $L_2(\pi)$-GE optimal rate is no smaller than $\sqrt{0.5}$. Hence the $(L_\infty(\pi), \|\cdot\|_{L_2(\pi)})$-GE and $L_2(\pi)$-GE optimal rates are different.

1.5.6 Proofs of Lemma 1.1 and Lemma 1.2

Proof of Lemma 1.1. Let

$$\rho^\star = \inf\left\{\rho > 0 : \exists C : V \to \mathbb{R}_+ \text{ s.t. } \forall n \in \mathbb{N}, \nu \in V \cap \mathcal{M}_{+,1} \quad |||\nu P^n - \pi||| \le C(\nu)\rho^n\right\} \;,$$
$$\hat{\rho} = \sup_{\mu \in V \cap \mathcal{M}_{+,1}} \limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n} \tag{1.44}$$

$(\hat{\rho} \le \rho^\star)$: Let $\epsilon > 0$

$$\begin{aligned}
\hat{\rho} &= \sup_{\mu \in V \cap \mathcal{M}_{+,1}} \limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n} \\
&\le \sup_{\mu \in V \cap \mathcal{M}_{+,1}} \limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n} \\
&\le \sup_{\mu \in V \cap \mathcal{M}_{+,1}} \limsup_{n\to\infty} (C_\epsilon(\mu)(\rho^\star + \epsilon)^n)^{1/n} \\
&= \rho^\star + \epsilon \;.
\end{aligned} \tag{1.45}$$

Since ϵ is arbitrary, $\hat{\rho} \le \rho^\star$.

$(\hat{\rho} \ge \rho^\star)$: For all $\nu \in V \cap \mathcal{M}_{+,1}$, $\limsup_{n\to\infty} |||\mu P^n - \pi|||^{1/n} \le \hat{\rho}$. Let $\epsilon > 0$. Then for all $\mu \in V \cap \mathcal{M}_{+,1}$, $|||\mu P^n - \pi|||^{1/n} > \hat{\rho} + \epsilon$ for at most finitely many $n \in \mathbb{N}$. Let $C_\epsilon(\mu) = \max_{n\in\mathbb{N}} \left(1 \vee \frac{|||\mu P^n - \pi|||}{(\rho+\epsilon)^n}\right)$. Then $C_\epsilon(\mu) < \infty$ since the maximum is over finitely many distinct elements. Therefore $|||\mu P^n - \pi||| \le C_\epsilon(\mu)(\hat{\rho} + \epsilon)^n$ for all $n \in \mathbb{N}$. This implies that $\hat{\rho} + \epsilon \ge \rho^\star$. Since ϵ is arbitrary, $\hat{\rho} \ge \rho^\star$. ☐

Proof of Lemma 1.2. [*(iii)* $\iff$ *(iv)*] is proven in [96, Theorem 2.1]. [*(iii)* $\implies$ *(ii)*] follows from the inclusion $L_\infty(\pi) \subset L_2(\pi)$. [*(ii)* $\implies$ *(i)*] follows from Cauchy-Schwarz.

[*(ii)* $\implies$ *(iii)*]:

Without loss of generality, assume that ρ is the optimal rate of $(L_\infty(\pi_\epsilon), \|\cdot\|_{L_2(\pi)})$-geometric ergodicity;

$$\rho = \sup_{\nu \in L_{\infty,0}(\pi)} \limsup_{t\to\infty} \left\|\nu P^t\right\|_{L_2(\pi)}^{1/t} \;. \tag{1.46}$$

From the proof of Roberts and Tweedie [100, Theorem 1], P is π-almost-everywhere geometrically ergodic with some unknown optimal rate. From [96, Theorem 2.1], P is $L_2(\pi)$-geometrically ergodic with some unknown optimal rate, ρ_2, which is equivalent to the spectral radius of $P|_{L_{2,0}(\pi)}$; $\rho_2 = r(P|_{L_{2,0}(\pi)})$.

It remains to be shown that $\rho_2 \leq \rho$. We will use the spectral measure decomposition of P, as in [96]. Suppose, for a contradiction, that $\rho_2 > \rho$. Let $\overline{\rho} = \frac{\rho+\rho_2}{2}$. Let $\mathcal{E}$ be the spectral measure of P, so that $\mu P^t = \int_{-1}^{1} \lambda^t \mu \mathcal{E}(d\lambda)$. If $\rho_2 > \rho$ then either $\mathcal{E}([-\rho_2, -\overline{\rho})) \neq \mathbf{0}$ or $\mathcal{E}((\overline{\rho}, \rho_2])) \neq \mathbf{0}$. Assume (replacing P by P^2, ρ by ρ^2, and ρ_2 by ρ_2^2 if necessary) that $\mathcal{E}((\overline{\rho}, \rho_2]) \neq \mathbf{0}$ and $\mathcal{E}((-1, 0)) = \mathbf{0}$. Then there is some non-zero signed measure, ν, in the range of $\mathcal{E}((\overline{\rho}, \rho_2])$. Since the spectral projections are orthogonal and $\{1\} \cap (\overline{\rho}, \rho_2] = \emptyset$, then $\nu \perp \pi$, and hence $\nu(\mathcal{X}) = 0$. Since $L_{\infty,0}(\pi)$ is dense in $L_{2,0}(\pi)$, there is a $\mu \in L_{\infty,0}(\pi)$ with $\|\mu - \nu\|_{L_2(\pi)} < \|\nu\|_{L_2(\pi)}/2$. Then, from the polarization identity, $\langle \nu,\ \mu \rangle_{L_2(\pi)} \geq \frac{3}{8}\|\nu\|^2_{L_2(\pi)} > 0$, and $\mu \neq 0$.

Let $R = \mathrm{range}(\int_{(\overline{\rho},\rho_2]} \mathcal{E}(d\lambda))$. Then $\mathrm{span}(\nu) \subset R$, so

$$\|\mathrm{proj}_R\, \mu\|_{L_2(\pi)} \geq \|\mathrm{proj}_\nu\, \mu\|_{L_2(\pi)} \geq \frac{3}{8}\|\nu\|_{L_2(\pi)} \tag{1.47}$$

Then

$$\begin{aligned}
\left\|\mu P^k\right\|^2_{L_2(\pi)} &= \left\langle \mu P^k,\ \mu P^k \right\rangle_{L_2(\pi)} \\
&= \left\langle \mu,\ \mu P^{2k} \right\rangle_{L_2(\pi)} \\
&= \left\langle \mu,\ \mu \int_{(0,\rho_2]} \lambda^{2k} \mathcal{E}(d\lambda) \right\rangle_{L_2(\pi)} \\
&\geq \left\langle \mu,\ \mu \int_{(\overline{\rho},\rho_2]} \lambda^{2k} \mathcal{E}(d\lambda) \right\rangle_{L_2(\pi)} \\
&\geq \left\langle \mu,\ \mu \int_{(\overline{\rho},\rho_2]} \overline{\rho}^{2k} \mathcal{E}(d\lambda) \right\rangle_{L_2(\pi)} \\
&= \overline{\rho}^{2k}\|\mathrm{proj}_R\, \mu\|^2_{L_2(\pi)} \\
&\geq \overline{\rho}^{2k}\frac{9}{64}\|\nu\|^2_{L_2(\pi)}\ .
\end{aligned} \tag{1.48}$$

Hence $\rho \geq \overline{\rho}$. This contradicts $\rho_2 > \rho$.

[*(i)* $\implies$ *(ii)*]:

Let the optimal rates of $(L_\infty(\pi_\epsilon), \|\cdot\|_{L_1(\pi)})$-GE and $(L_\infty(\pi_\epsilon), \|\cdot\|_{L_2(\pi)})$-GE be (respectively)

$$\rho = \sup_{\mu \in L_{\infty,0}(\pi)} \limsup_{n\to\infty} \|\mu P^n\|_{L_1(\pi)}^{1/n} \ , \quad \rho_2 = \sup_{\mu \in L_{\infty,0}(\pi)} \limsup_{n\to\infty} \|\mu P^n\|_{L_2(\pi)}^{1/n} \ . \tag{1.49}$$

We want to show that $\rho_2 \leq \rho$.

Let $\epsilon > 0$ be arbitrary. Let $\nu_\epsilon \in L_{\infty,0}(\pi)$ with

$$\limsup_{n\to\infty} \|\nu_\epsilon P^n\|_{L_2(\pi)}^{1/n} \geq \rho_2 - \epsilon \ . \tag{1.50}$$

Then, for some $c(\nu_\epsilon) > 0$, for infinitely many $n \in \mathbb{N}$

$$\|\nu_\epsilon P^n\|_{L_2(\pi)} \geq (\rho_2 - 2\epsilon)^n \ . \tag{1.51}$$

Using the fact that $\|\mu\|_{L_1(\pi)} = \sup_{\substack{f \in L'_\infty(\pi) \\ \|f\|_{L'_\infty(\pi)}}} \mu f$, and using the self-adjointness of P in $L_2(\pi)$ (since P is reversible), and using the fact that (a version of) $\frac{\mathrm{d}\nu_\epsilon}{\mathrm{d}\pi}$ is some bounded function with $\left\|\frac{\mathrm{d}\nu_\epsilon}{\mathrm{d}\pi}\right\|_{L'_\infty(\pi)} = \|\nu_\epsilon\|_{L_\infty(\pi)}$, then for infinitely many $n \in \mathbb{N}$,

$$\begin{aligned}
\left\|\nu_\epsilon P^{2n}\right\|_{L_1(\pi)} &= \sup_{\|f\|_\infty \leq 1} \nu_\epsilon P^{2n} f \\
&\geq \frac{1}{\|\nu_\epsilon\|_{L_\infty(\pi)}} \nu_\epsilon P^{2n} \frac{\mathrm{d}\nu_\epsilon}{\mathrm{d}\pi} \\
&= \frac{1}{\|\nu_\epsilon\|_{L_\infty(\pi)}} \left\langle \nu_\epsilon P^{2n},\ \nu_\epsilon \right\rangle \\
&= \frac{1}{\|\nu_\epsilon\|_{L_\infty(\pi)}} \left\langle \nu_\epsilon P^n,\ \nu_\epsilon P^n \right\rangle \\
&= \frac{1}{\|\nu_\epsilon\|_{L_\infty(\pi)}} \|\nu_\epsilon P^n\|_{L_2(\pi)}^2 \\
&\geq \frac{1}{\|\nu_\epsilon\|_{L_\infty(\pi)}} (\rho_2 - 2\epsilon)^{2n}
\end{aligned} \tag{1.52}$$

Thus $\rho_2 - 2\epsilon \leq \rho$. Since ϵ was arbitrary, we find that $\rho_2 \leq \rho$.

□

Chapter 2

Integration by Parts and the Geometry of Probability Density Functions

2.1 Introduction

Integration by parts formulas are indispensable tools in analysis. They are commonly used to evaluate otherwise unapproachable expressions, and so are one of the first things students learn in elementary calculus courses. In more advanced analysis, integration by parts formulas are also used to define weak derivatives (see, for example Maggi [69]) or the infinitesimal generator associated with a Markov diffusion triple (as in Bakry et al. [7]). Versions of integration by parts specialized to probability densities are commonly used in theoretical and applied probability theory. They appear in the study of continuous time Markov processes (as in [7]), convergence of probability measures (as in Chen et al. [25]), spin glass (as in Panchenko [87]) and other applications.

The seminal work of Stein [112] characterizes a normal distribution as the unique probability measure satisfying an integration by parts formula, and uses this characterization to establish a quantitative central limit theorem. From this seminal work, an entire subfield of probability was born – commonly referred to as “Stein’s Method”. In its simplest form, Stein's method for normal approximation relies on the following lemma, stated here as in [25, Lemma 2.1].

Lemma 2.1 (Stein's Lemma)**.** *Let $\mathcal{F}$ be the class of absolutely continuous functions with $\underset{Z\sim\mathcal{N}(0,1)}{\mathbb{E}} |f'(Z)| < \infty$. Then*

$$[W \sim \mathcal{N}(0,1)] \iff [\mathbb{E}f'(W) = \mathbb{E}Wf(W) \quad \forall f \in \mathcal{F}]$$

While in the literature of Stein's method "Stein's Lemma" often refers to this bidirectional result, for our purposes it will refer to the one directional result

$$[W \sim \mathcal{N}(0,1)] \implies [\mathbb{E}f'(W) = \mathbb{E}Wf(W) \quad \forall f \in \mathcal{F}].$$

It is this one directional result which we will generalize. In particular, in Theorem 2.1, we show that for a large class of densities on $(\mathbb{R}^n, \mathcal{R}^n)$ and for a large function class, $\mathcal{F}$,

$$[W \sim \pi] \implies [\mathbb{E}\nabla f(W) = -\mathbb{E}\nabla \log \pi(W) f(W) \quad \forall f \in \mathcal{F}].$$

Related Work

Other generalizations of Stein's lemma exist in the literature. For example [25, Ch. 13] and Stein et al. [113, Prop. 1.4] handle densities on $\mathbb{R}$ which may be discontinuous at the boundary of the support, such as the exponential distribution. The present work is not a direct generalization of that result, since we do not accommodate jump-discontinuities at the boundary of the support. Another example, Landsman [61], provides a version for multivariate elliptic distributions. Since not all elliptic distributions have densities, our result is not a strict generalization of theirs. Our result is a strict generalization of their result restricted to distributions with weakly differentiable densities $\pi : \mathbb{R}^n \to \overline{\mathbb{R}}$.

The result most similar to the present work is Gorham et al. [37, Prop. 3]. Their version handles continuously differentiable densities on $\mathbb{R}^n$; and integrands which are continuously differentiable and absolutely integrable, with absolutely integrable gradients. Our result is a strict generalization of theirs, over both the densities and the integrands to which it applies. The proof of Theorem 2.1 is more similar to that in [25, Lemma 2.1], while the proof in [37] is similar to Proposition 2.1 since our proof uses an intrinsic foliation of the density along its own level sets, while the proof of [37, Prop. 3] verifies that the boundary integral in integration by parts vanishes given the assumptions.

As in [37], variants of the integration by parts formula for a distribution π arise from the Fokker-Planck equation of a Langevin diffusion process. However, this only yields second-order versions of the formula. Taking $\mathcal{F}$ to be the domain of the generator of a Langevin diffusion with stationary measure π, the Fokker-Planck equation tell us that

$$[W \sim \pi] \implies [\mathbb{E}\Delta f(W) + \mathbb{E}\nabla \log \pi(W)' \nabla f(W) = 0 \quad \forall f \in \mathcal{F}].$$

This formula is follows from Corollary 2.1, applied to ∇f. Corollary 2.1 is not implied by the Fokker-Planck equation since not all functions are the gradient of a scalar field.

Outline

The present chapter proceeds as follows. Section 2.2 reviews the existing results for the univariate Gaussian case and their proofs. It serves as a blueprint for the proof of our main result, and provides intuition for the key steps. Section 2.3 states and proves our main result. That proof relies on some geometric properties of densities which satisfy our key assumptions stated in Theorem 2.1. Those geometric properties are established in Section 2.4. Finally, Section 2.5 applies our result to densities, π, such that $\nabla \log \pi$ is L-Lipschitz to demonstrate that if $X \sim \pi$ then $\mathrm{Cov}(\nabla \log \pi(X)) = -\mathbb{E}\nabla^2 \log \pi(X)$ and that $\nabla \log \pi(X)$ is sub-Gaussian with dimension-free sub-Gaussian constant L.

2.2 The Univariate Gaussian Case

Stein's Lemma for the univariate Gaussian tells us that if $X \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\sigma^2 \mathbb{E} f'(X) = \mathbb{E}[(X - \mu) f(X)]$$

for “suitable” f. The result can be proved in (at least) two different ways, leading to different conditions needed to verify that f is “suitable”. First, we use that the formula resembles integration by parts without the boundary term.

Proposition 2.1. *If π is a probability density on $\mathbb{R}$ with support S, which (as a function) is absolutely continuous, $f : \mathbb{R} \to \mathbb{R}$ is absolutely continuous, and $\lim_{x\to x'} f(x)\pi(x) = 0$ for all $x' \in \partial S \cup \{-\infty, \infty\}$*

$$\mathop{\mathbb{E}}_{Z\sim\pi}\left([\log \pi]'(Z)\ f(Z)\right) = -\mathop{\mathbb{E}}_{Z\sim\pi}[f'(Z)]$$

The result is just integration by parts, with the recognition that $\pi' = \pi[\log \pi]'$ and that the assumptions directly imply the boundary terms of integration by parts vanish. These results are unsatisfactory in some cases because of the condition that the product $f \cdot \pi$ must vanish at the boundary of the support of π. In one dimension this could amount to evaluating a countable set of limits. The problem is made worse in higher dimensions where the integration by parts formula gives a limits of surface integrals instead of limits of function evaluations. The second, and more widely used, variant Stein’s Lemma for the univariate Gaussian (Proposition 2.2) gives a measure-theoretic constraint which essentially says that when $\mathbb{E}[f'(Z)]$ and $\mathbb{E}[Zf(Z)]$ are well-defined the result holds. The proof of this version will guide us in the multivariate setting.

Proposition 2.2 (Chen et al. [25], Lemma 2.1)**.** *If $f : \mathbb{R} \to \mathbb{R}$ is abs. continuous with $f'(Z)$, $Zf(Z)$ extended integrable, and $Z \sim \mathcal{N}(0, 1)$, then*

$$\mathop{\mathbb{E}}_{Z\sim\pi}[Zf(Z)] = \mathop{\mathbb{E}}_{Z\sim\pi}[f'(Z)]$$

The proof presented here proof is a slight modification of versions seen elsewhere, and so is not original. For example, a variant of this proof appears in [25].

Proof. Let $\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. Then

$$\begin{aligned}
\int_{-\infty}^{\infty} f'(x)\phi(x)\ dx &= \int_{-\infty}^{\infty} f'(x) \int_0^{\phi(x)} 1\ dr\ dx \\
&= \int_0^{\frac{1}{\sqrt{2\pi}}} \int_{-\sqrt{-2\log(\sqrt{2\pi}\ r)}}^{\sqrt{-2\log(\sqrt{2\pi}\ r)}} f'(x)dx\ dr && \text{(Key step 1)} \\
&= \int_0^{\frac{1}{\sqrt{2\pi}}} \left(f(\sqrt{-2\log(\sqrt{2\pi}\ r)}) - f(-\sqrt{-2\log(\sqrt{2\pi}\ r)})\right) dr && \text{(Key step 2)} \\
&= \int_0^{\frac{1}{\sqrt{2\pi}}} \left[\sum_{x\in\phi^{-1}(\{r\})} -\operatorname{sign}(\phi'(x))f(x)\right] dr && \text{(Key step 3)} \\
&= \int_{-\infty}^{\infty} -\operatorname{sign}(\phi'(x))f(x)\, |\phi'(x)|\ dx && \text{(Key step 4)} \\
&= \int_0^{\infty} f(x)x\frac{\exp(-x^2/2)}{\sqrt{2\pi}}\ du
\end{aligned}$$

Key step 1 is the Fubini-Tonelli theorem. Key step 2 is is the fundamental theorem of (Lebesgue integral) calculus. Key step 3 uses the fact that boundary of the super level sets of ϕ were exactly the level sets of ϕ. Key step 4 is the co-area formula, [34, Theorem 3.2.12], which requires only that the function whose level sets define the foliation be Lipschitz – an assumption which is satisfied by the normal density function.[1] □

In order to extend this to more general densities, both univariate and multivariate, we need to understand what properties of the normal density allowed us to use the four key steps in the proof. Fubini-Tonelli (key step 1) required that $f'(Z)$ is extended integrable. The fundamental theorem of calculus (key step 2) required only that the super-level sets to be a countable union of intervals. In the multivariate setting the variant of the fundamental theorem of calculus which will be relevant is the Gauss-Green theorem, so we will require that (almost all) of the superlevel sets of ϕ *admit a Gauss-Green measure*. Key step 3 required one to relate the boundaries of superlevel sets of ϕ to the corresponding level sets. In the case of the normal distribution this was trivial. In the case of more general densities, which may have extensive flat regions, we establish some geometric results in Section 2.4 which show that level sets are the boundaries of superlevel sets at almost every level for the class of distributions we consider. Finally, key step 4 (the co-area formula) required the

[1]Key step 4 above could have been replaced by a pair of u-substitutions, with $u = \pm\phi^{-1}(r)$, which is the case in most versions of this proof in the literature (e.g. in [25]). In the multivariate version of the result, a substitution will not be adequate.

level sets of π to be level sets of Lipschitz function.

2.3 The General Multivariate Case

For this section we adopt the notation of geometric measure theory, where $\mathcal{L}^n$ denotes the n-dimensional Lebesgue measure, and $\mathcal{H}^n$ denotes the n-dimensional Hausdorff measure. The extended real numbers are denoted by $\overline{\mathbb{R}}$ and endowed with the order topology.

Definition 2.1 (Compositionally Lipschitz function)**.** *$f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is compositionally Lipschitz if there exists a strictly increasing and absolutely continuous function $\rho : \overline{\mathbb{R}} \to \overline{\mathbb{R}}$ with ρ' positive $\mathcal{L}^1$-almost everywhere and $\rho \circ f$ Lipschitz,.*

If we wish to emphasize the ρ used we may say *ρ-compositionally Lipschitz* or *ρ-CL*, and when we wish to emphasize the ρ used and the Lipschitz constant of $\rho \circ f$ we may use the term *(ρ, L)-compositionally Lipschitz* or *(ρ, L)-CL*. Of course, if ρ is the identity function then the function is just Lipschitz. More generally, if ρ^{-1} is L_1-Lipschitz and f is (ρ, L_2)-compositionally Lipschitz then f is also $(L_1 L_2)$-Lipschitz.

Theorem 2.1 (Distributional Integration by Parts, a.k.a. Stein's Lemma)**.** *Suppose that $\pi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a compositionally Lipschitz probability density. Then, for any $f : \mathbb{R}^n \to \mathbb{R}$ which is locally Lipschitz, with $\nabla f(x)$ and $f(x)\nabla \log \pi(x)$ extended integrable functions (w.r.t. $\pi(x)\mathcal{L}^n(dx)$) we have:*

$$\mathop{\mathbb{E}}_{X\sim\pi} f(X)\nabla \log \pi(X) = -\mathop{\mathbb{E}}_{X\sim\pi} \nabla f(X)$$

Remark 2.1 (On the differentiability of f)**.** *By Rademacher's Theorem ([34], Theorem 3.1.6), such f will be differentiable almost everywhere with a measurable gradient, giving meaning to the subsequent expressions. The assumption that f is locally Lipschitz is equivalent to the assumption that f is Lipschitz on compact sets.* $\triangleleft$

Remark 2.2 (Some less smooth choices for π for which the theorem holds)**.** *Our weak assumptions allow us to handle some non-smoothness in π. Two examples are (i) the semi-circle law, $\pi(x) \propto \mathbf{1}_{|x|<2}\sqrt{4-x^2}$, which is not Lipschitz, has compact support, and* *is of significance in random matrix theory, and (ii) unbounded elliptic densities, such as* $\pi(x) \propto |x|^{-1/2} \wedge x^{-2}$. *Neither of these satisfy the conditions of [37, Prop. 3].* $\triangleleft$

Corollary 2.1 (Jacobians and divergences of vector valued functions)**.** *By applying the integration by parts formula for real valued functions coordinate-wise, the analogous formula for Jacobians also holds. If* π *satisfies the conditions above and* $f : \mathbb{R}^n \to \mathbb{R}^m$ *is locally Lipschitz with* Jf *and* $f(X)\nabla \log \pi(X)'$ *coordinate-wise extended integrable, then*

$$\mathop{\mathbb{E}}_{X\sim\pi}[f(X)\nabla \log \pi(X)'] = -\mathop{\mathbb{E}}_{X\sim\pi}[Jf(X)],$$

where Jf *denotes the Jacobian of* f.

If, additionally, $m = n$ *and neither* $\mathop{\mathbb{E}}_{X\sim\pi}[f(X)\nabla \log \pi(X)']$ *nor* $\mathop{\mathbb{E}}_{X\sim\pi}[Jf(X)]$ *have both* $+\infty$ *and* $-\infty$ *on the diagonal, then*

$$\mathop{\mathbb{E}}_{X\sim\pi}[\nabla \log \pi(X)' f(X)] = -\mathop{\mathbb{E}}_{X\sim\pi}[\operatorname{div}(f)(X)].$$

Proof of Distributional Integration by Parts. By assumption, there exists an absolutely continuous function with a.e. positive derivative $\rho : \mathbb{R} \to \mathbb{R}$ such that π is ρ-CL (see Definition 2.1). For each $\epsilon > 0$, let $A_\epsilon = \pi^{-1}((\epsilon, \infty]) \subset K_\epsilon = \pi^{-1}([\epsilon, \infty])$. By definition,

$$\mathop{\mathbb{E}}_{X\sim\pi} \nabla f(X) = \int_{\mathbb{R}^n} \pi(x)\nabla f(x)\ \mathcal{L}^n(dx) = \int_{\mathbb{R}^n} \int_0^{\pi(x)} \mathcal{L}^1(dr)\ \nabla f(x)\ \mathcal{L}^n(dx).$$

(Key step 1): Since $\nabla f(X)$ is assumed to be extended integrable, by Fubini-Tonelli

$$\int_{\mathbb{R}^n} \int_0^{\pi(x)} \mathcal{L}^1(dr)\ \nabla f(x)\ \mathcal{L}^n(dx) = \int_0^\infty \int_{A_r} \nabla f(x)\ \mathcal{L}^n(dx)\ \mathcal{L}^1(dr).$$

(Key step 2): Using a version of the Gauss-Green theorem, Lemma 2.2,

$$\int_0^\infty \int_{A_r} \nabla f(x)\ \mathcal{L}^n(dx)\ \mathcal{L}^1(dr) = \int_0^\infty \int_{\partial^*(A_r)} f(x)\hat{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx)\ \mathcal{L}^1(dr),$$

where $\partial^*(A_r) \subseteq \partial(A_r)$ is the reduced boundary of A_r (see [69, Chapter 15]); and $\hat{\mathbf{n}}(A_r, x)$ is the unit outward-facing measure-theoretic normal vector to A_r at x when A_r has locally-finite perimeter and $x \in \partial^*(A_r)$; and $\hat{\mathbf{n}}(A_r, x)$ is 0 otherwise. Note that $\mathcal{L}^1$-almost every superlevel set of a Lipschitz function has locally finite perimeter [69, Example 13.3]. Since π is ρ-CL, A_r has locally finite perimeter for $\mathcal{L}^1$-almost every $r > 0$.

(Key step 3): If $r \notin E_\pi = \{s > 0 : \partial(\pi^{-1}((s, \infty])) \neq \pi^{-1}(\{s\})\}$, then $\partial(A_r) = \pi^{-1}(\{r\})$. From Lemma 2.4, $\mathcal{L}^1(E_\pi) = 0$. Therefore, for $\mathcal{L}^1$-almost every $r > 0$,

$$\int_{\partial^*(A_r)} f(x)\hat{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx) = \int_{\pi^{-1}(\{r\})} f(x)\tilde{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx),$$

where $\tilde{\mathbf{n}}(A_r, x)$ is the unit outward-facing normal vector to A_r at x when $r \notin E_\pi$, A_r has locally finite perimeter, and $x \in \partial^*(A_r)$, and is 0 otherwise. Therefore

$$\int_0^\infty \int_{\partial^*(A_r)} f(x)\hat{\mathbf{n}}(A_r, x)\mathcal{H}^{n-1}(dx)\mathcal{L}^1(dr) = \int_0^\infty \int_{\pi^{-1}(\{r\})} f(x)\tilde{\mathbf{n}}(A_r, x)\mathcal{H}^{n-1}(dx)\mathcal{L}^1(dr).$$

Changing from π-coordinates to $\rho \circ \pi$ coordinates, $s = \rho(r)$, so that our level sets are taken with respect to a Lipschitz function (and hence we can later apply the co-area formula):

$$\begin{aligned}&\int_0^\infty \int_{\pi^{-1}(\{r\})} f(x)\tilde{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx)\ \mathcal{L}^1(dr) \\ &\quad = \int_{\rho(0)}^{\rho(\infty)} \int_{(\rho\circ\pi)^{-1}(\{s\})} \frac{f(x)}{\rho'(\pi(x))}\tilde{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx)\ \mathcal{L}^1(ds)\end{aligned}$$

Let $J_1[\rho \circ \pi]$ denote the 1×1 Jacobian of $\rho \circ \pi$ [34, Definition 3.2.1]. From [34] we have the formula $J_1\pi(x) = \|\wedge_1 \nabla\rho \circ \pi(x)\| = \rho'(\pi(x))\,\|\nabla\pi(x)\|$. For $r \notin E_\pi$, at every point, $x \in \pi^{-1}(\{r\})$, where π is differentiable and $\nabla\pi(x) \neq 0$, from [69, Theorem 15.9], we have $x \in \partial^* A_r$ and $\tilde{\mathbf{n}}(A_r, x) = -\frac{\nabla\pi(x)}{\|\nabla\pi(x)\|}$. This captures the intuition that when π is differentiable at x and $\nabla\pi(x) \neq 0$ then the negative standardized gradient is the unit outward facing normal to the level set. Moreover, if $\nabla\pi(x) = 0$ and $x \in \partial^*(A_r) \subset \pi^{-1}(\{r\})$, then $\tilde{\mathbf{n}}(A_r, x)\ J_1[\rho \circ \pi(x)] = 0$. Thus $\frac{f(x)}{\rho'\circ\pi(x)}\tilde{\mathbf{n}}(A_r, x)\ J_1[\rho \circ \pi(x)] = -f(x)\nabla\pi(x)$ almost everywhere, and hence is extended integrable (w.r.t $\pi(x)\mathcal{L}^n(dx)$) by assumption.

(Key step 4): Applying the co-area formula [34, Theorem 3.2.12] coordinate-wise, we get:

$$\int_{\rho(0)}^{\rho(\infty)} \int_{(\rho\circ\pi)^{-1}(\{s\})} \frac{f(x)}{\rho'(\pi(x))} \tilde{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx)\ \mathcal{L}^1(ds)$$
$$= \int_{\mathbb{R}^n} \frac{f(x)}{\rho' \circ \pi(x)} \tilde{\mathbf{n}}(A_r, x)\ J_1[\rho \circ \pi(x)]\ \mathcal{L}^n(dx)$$
$$= -\int_{\mathbb{R}^n} f(x)\nabla\pi(x)\mathcal{L}^n(dx)$$
$$= -\mathop{\mathbb{E}}_{X\sim\pi} f(X)\nabla \log \pi(X)$$

□

Lemma 2.2 (Gauss-Green theorem for super level sets of densities.)**.** *If π is a CL prob. density on $\mathbb{R}^n$, then for $\mathcal{L}^1$-a.e. $r > 0$ and any locally Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$*

$$\int_{A_r} \nabla f(x)\ \mathcal{L}^n(dx)\ = \int_{\partial^* A_r} f(x)\hat{\mathbf{n}}(A_r, x)\ \mathcal{H}^{n-1}(dx)$$

where $A_r = \pi^{-1}((r, \infty])$, $\partial^* A_r$ is the reduced boundary of A_r and $\hat{\mathbf{n}}(A_r, x)$ is the unit outward facing normal vector to A_r at point x.

Proof. This is a specialization of the Gauss-Green theorem from geometric measure theory to the problem at hand. The purpose of this Lemma is essentially to verify that existing versions of Gauss-Green can be applied to the collection integrals in question. Let P_π be the set of $r > 0$ such that A_r has locally finite perimeter. Since almost every superlevel set of a Lipschitz function has locally finite perimeter [69, Example 13.3], $L^1(P_\pi^c) = 0$. We consider only $r \in P_\pi$ from now on. Sets of locally finite perimeter admit Gauss-Green measures [69, Proposition 12.1 and Remark 12.2] – the Gauss-Green measure for $E \subset \mathbb{R}^n$ is an $\mathbb{R}^n$-valued Radon measure, μ_E, such that

$$\int_E \nabla g(x)\ \mathcal{L}^n(dx) = \int_{\mathbb{R}^n} g(x)\ \mu_E(dx) \quad \forall g \in C_c^1(\mathbb{R}^n).$$

The Gauss-Green Measure of $E \subset \mathbb{R}^n$ admits the representation $\mu_E = \hat{\mathbf{n}}(E, x)\mathcal{H}^{n-1}|_{\partial^* E}$ where $\partial^* E$ is the reduced boundary of E, and $\hat{\mathbf{n}}(E, x)$ is the (measure-theoretic) outer unit normal to E [69, Chapter 15 and Corollary 16.1]. The definition of μ_E only guarantees a Gauss-Green formula holds for integrands $g \in C_c^1(\mathbb{R}^n)$. This extends to Lipschitz functions with compact support, as outlined in [69, Exercise 12.12], via convolution with smooth bump functions. Recalling that locally Lipschitz functions are Lipschitz on compacts, a locally Lipschitz integrand, f, must have been Lipschitz on $\pi^{-1}([r/2,\infty] \supset A_r$ which is compact by Lemma 2.3. The function f could be extended to a globally Lipschitz function with compact support without changing its value on $\pi^{-1}([r/2,\infty])$. Hence the Gauss-Green formula extends to locally Lipschitz functions on the domains $\{A_r : r \in P_\pi\}$. $\square$

2.4 Geometry of Density Functions

Lemma 2.3 (Compositionally Lipschitz densities have compact superlevel sets)**.** *If π is a (ρ, L)-CL probability density on $\mathbb{R}^n$ then the closed superlevel sets of π,*

$$\left\{K_\epsilon = \pi^{-1}([\epsilon,\infty]) \text{ s.t. } \epsilon > 0\right\},$$

are all compact.

Proof. Since π is continuous, $K_\epsilon = \pi^{-1}([\epsilon,\infty])$ is closed for all $\epsilon > 0$. Suppose, for contradiction, that K_ϵ is not compact for some $\epsilon > 0$. For K_ϵ to fail to be compact, it must be unbounded (since we know it is closed). Let $R \in \left(0, \frac{\rho(\epsilon)-\rho(\epsilon/2)}{L}\right]$. If K_ϵ is unbounded, we may find $\{x_j\}_{j\in\mathbb{N}} \subset K_\epsilon$ such that for $i \neq j$, $\|x_i - x_j\| \geq 3R$. Then $B_R(x_i) \cap B_R(x_j) = \emptyset$ for $i \neq j$. Hence

$$\begin{aligned}
1 = \int \pi(x)\mathcal{L}^n(dx) &\geq \sum_{j\in\mathbb{N}} \int_{B_R(x_j)} \pi(x)\mathcal{L}^n(dx) \\
&\geq \sum_{j\in\mathbb{N}} \int_{B_R(x_j)} \rho^{-1}(\rho\circ\pi(x_j) - L\|x - x_j\|)\mathcal{L}^n(dx) \\
&\geq \sum_{j\in\mathbb{N}} \int_{B_R(x_j)} \rho^{-1}(\rho(\epsilon) - L\|x - x_j\|)\mathcal{L}^n(dx) \\
&\geq \sum_{j\in\mathbb{N}} \int_{B_R(x_j)} \rho^{-1}(\rho(\epsilon/2))\mathcal{L}^n(dx) \\
&= \sum_{j\in\mathbb{N}} \mathrm{Vol}(B_R(0))\ \epsilon/2
\end{aligned}$$

This is a contradiction, since the last term is clearly $+\infty$ $\square$

Lemma 2.4 (Almost every level set of a CL density is the boundary of a superlevel set)**.**

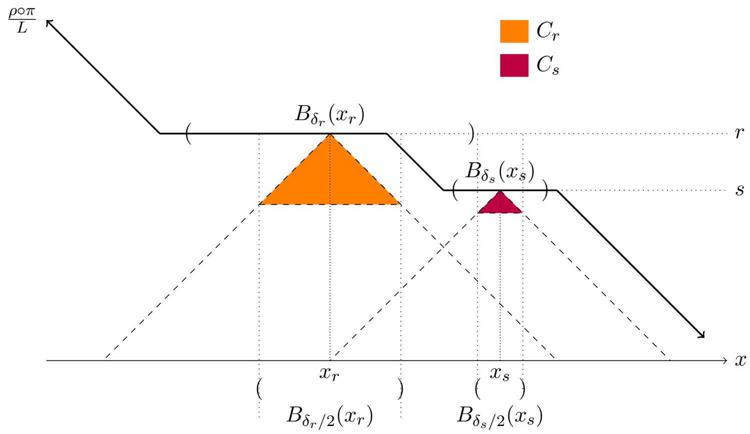


Figure 2.1: Visualization of $C_r \cap C_s = \emptyset$ for $r < s$ in the proof of Lemma 2.4

If π is a (ρ, L)-CL probability density on $\mathbb{R}^n$ then

$$E_\pi = \left\{ r > 0 : \partial(\pi^{-1}((r, \infty])) \neq \pi^{-1}(\{r\}) \right\}$$

is countable (and hence has $\mathcal{L}^1(E_\pi) = 0$).

Proof. Since π is continuous, $A_r = \pi^{-1}((r, \infty])$ is open. Suppose that $x \in \partial(A_r)$. Then $x \notin A_r$ so $\pi(x) \leq r$ and $\pi(x)$ is a limit point of $\pi(A_r) \subseteq (r, \infty]$. Hence $\partial(A_r) \subset \pi^{-1}(\{r\})$. Let $G_\pi = \{(x, y) : x \in \mathbb{R}^n \text{ and } 0 \leq y \leq \pi(x)\}$. Then $\mathcal{L}^{k+1}(G_\pi) = 1$ since π is a probability density. Let $d(x, A) = \inf_{y \in A} \|x - y\|$. Suppose $r \in E_\pi$. Then there exists an $x_r \in \pi^{-1}(\{r\}) \setminus \partial(A_r)$ with $\delta_r = d(x_r, A_r) > 0$. Since $\rho \circ \pi$ is Lipschitz, the ρ-transformed cone

$$C_r = \left\{ (x, y) \text{ s.t. } x \in B_{\delta_r/2}(x_r) \text{ and } \rho^{-1}(\rho(r) - \delta_r L/2) < y < \rho^{-1}(\rho(r) - \|x - x_r\| \, L) \right\}$$

has $C_r \subset G_\pi$ and C_r is open, and hence $\mathcal{L}^{k+1}(C_r) > 0$. We show below that for $s, r \in E_\pi$ with $s < r$ we have $C_s \cap C_r = \emptyset$. This is visualized in Fig. 2.1. Once established, this implies $\sum_{r \in E_\pi} \mathcal{L}^{k+1}(C_r) \leq \mathcal{L}^{k+1}(G_\pi) = 1$, which further implies that E_π is at most countable.

Suppose now that $s, r \in E_\pi$ with $s < r$. Then $\|x_r - x_s\| \geq \delta_s + (\rho(r) - \rho(s))/L$; since the line segment from x_r to x_s must pass trough ∂A_s at some point $x_s^\star \in \partial A_s \subset \pi^{-1}(\{s\})$, and the portion of the segment from x_s to $x_s^\star$ is at least δ_s in length, while the remaining portion from ∂A_s to x_r is bounded using the Lipschitz property of $\rho \circ \pi$, since $\rho(r) - \rho(s) = \rho(\pi(x_r)) - \rho(\pi(x_s^\star)) \leq L \|x_s^\star - x_r\|$.

Suppose, now, that there exists a pair $(x, y) \in C_r \cap C_s$. Then

$$\rho(r) - \|x - x_r\| L > \rho(y) > \rho(s) - \delta_s L/2 \ ,$$

which implies that

$$\begin{aligned} 0 &< \rho(r) - \rho(s) - L\|x - x_r\| + L\delta_s/2 \\ &\le L\|x_r - x_s\| - L\|x - x_r\| - L\delta_s/2 \\ &\le L(\|x_r - x\| + \|x - x_s\|) - L\|x - x_r\| - L\delta_s/2 \\ &\le L\|x - x_s\| - L\delta_s/2 \\ &< 0 \ , \end{aligned}$$

which is a contradiction. □

2.5 Properties of Grad-Log-Lipschitz Densities

A probability density, π, with $\nabla \log \pi$ Lipschitz will be referred to as a grad-log-Lipschitz density.

Lemma 2.5 (Grad-Log-Lipschitz Densities are Tangentially Minorized by Gaussians)**.** *If π is a probability density on $\mathbb{R}^n$, and $\nabla \log \pi$ is L-Lipschitz, then for any $x, x_0 \in \mathbb{R}^n$.*

$$\begin{aligned} \pi(x) &\ge \pi(x_0) e^{\left\|\nabla \log \pi(x_0)\right\|^2/2L} \exp\Big(-\Big\|x - x_0 - \frac{\nabla \log \pi(x_0)}{L}\Big\|^2 \frac{L}{2}\Big) \\ &\ge \pi(x_0) \exp\Big(-\Big\|x - x_0 - \frac{\nabla \log \pi(x_0)}{L}\Big\|^2 \frac{L}{2}\Big). \end{aligned}$$

Proof. Since $\nabla \log \pi$ is L-Lipschitz,

$$\log \pi(x) - \log \pi(x_0) - (x - x_0)' \nabla \log \pi(x_0) \ge -L\|x - x_0\|^2 / 2$$

The result follows by completing the square and exponentiating. □

Lemma 2.6 (Grad-Log-Lipschitz Densities are Bounded Above)**.** *If π is a probability den-*

sity on $\mathbb{R}^n$, and $\nabla \log \pi$ is L-Lipschitz then:

$$\pi(x) \leq \left(\frac{L}{2\pi}\right)^{n/2} e^{-\|\nabla \log \pi(x)\|^2/2L} \leq \left(\frac{L}{2\pi}\right)^{n/2}$$

Proof. Using Lemma 2.5

$$\begin{aligned}
1 &= \int \pi(y)\mathcal{L}^n(dy) \\
&\geq \pi(x)e^{\|\nabla \log \pi(x)\|^2/2L} \int \exp\Big(-\Big\|y - x - \frac{\nabla \log \pi(x)}{2L}\Big\|^2 \frac{L}{2}\Big)\mathcal{L}^n(dy) \\
&= \pi(x)e^{\|\nabla \log \pi(x)\|^2/2L} \int \exp\Big(-\|y\|^2 \frac{L}{2}\Big)\mathcal{L}^n(dy) \\
&= (2\pi/L)^{n/2} e^{\|\nabla \log \pi(x)\|^2/2L}\pi(x).
\end{aligned}$$

□

Lemma 2.7 (Grad-Log-Lipschitz Densities are Lipschitz)**.** *If π is a probability density on $\mathbb{R}^n$, and $\nabla \log \pi$ is L-Lipschitz then π is $\sqrt{L}e^{-1/2}\left(\frac{L}{2\pi}\right)^{n/2}$-Lipschitz.*

Proof. Applying Lemma 2.6, for any $x \in \mathbb{R}^n$

$$\|\nabla\pi(x)\| = \pi(x)\,\|\nabla \log \pi(x)\| \leq \left(\frac{L}{2\pi}\right)^{\frac{n}{2}} e^{-\frac{\|\nabla \log \pi(x)\|^2}{2L}}\,\|\nabla \log \pi(x)\| \leq \left(\frac{L}{2\pi}\right)^{\frac{n}{2}} \sup_{s\geq 0}\left(se^{-\frac{s^2}{2L}}\right).$$

Now, $\frac{d}{ds}\left(s\ e^{-s^2/2L}\right) = \left(e^{-s^2/2L}(1 - s^2/L)\right)$ so the maximum of $s\ e^{-s^2/2L}$ over $s \geq 0$ occurs at $s = \sqrt{L}$ (since the derivative is positive to left and negative to the right of this value) and the maximum value is $\sqrt{L}e^{-1/2}$. Hence $\|\nabla\pi(x)\| \leq \sqrt{L}e^{-1/2}\left(\frac{L}{2\pi}\right)^{n/2}$. □

Corollary 2.2. *Suppose that π is a grad-log-Lipschitz probability density on $\mathbb{R}^n$. Then, for any $f : \mathbb{R}^n \to \mathbb{R}$ which is locally Lipschitz, with $\nabla f(X)$ and $f(x)\nabla \log \pi(x)$ extended integrable (w.r.t. $\pi(x)dx$) we have:*

$$\mathop{\mathbb{E}}_{X\sim\pi} f(X)\nabla \log \pi(X) = -\mathop{\mathbb{E}}_{X\sim\pi} \nabla f(X)$$

Similar formulas for the Jacobian and divergence also hold.

Proof. This is just the combination Theorem 2.1 and Lemma 2.7. □

Lemma 2.8. *Suppose that π is a grad-log-Lipschitz probability density on $\mathbb{R}^n$.*

$$\mathop{\mathbb{E}}_{X\sim\pi}[\nabla \log \pi(X)] = 0, \quad \textit{and} \quad \mathop{\mathrm{Var}}_{X\sim\pi}[\nabla \log \pi(X)] = -\mathop{\mathbb{E}}_{X\sim\pi}[\nabla^2 \log \pi(X)].$$

Proof. Since $\nabla \log \pi(X)'\nabla \log \pi(X)$ is non-negative, then it is extended integrable. Since

$$[\mathrm{div}(\nabla \log \pi)](X) = \mathrm{Tr}(\nabla^2 \log \pi(X)) \in [-nL, nL]$$

is bounded, $\mathrm{Tr}(\nabla^2 \log \pi(X))$ is also integrable. Hence Corollary 2.2 gives us that

$$\mathop{\mathbb{E}}_{X\sim\pi}[\nabla \log \pi(X)'\nabla \log \pi(X)] = -\mathop{\mathbb{E}}_{X\sim\pi}[\mathrm{Tr}(\nabla^2 \log \pi)] \leq nL.$$

Now, we have that

$$\mathbb{E}\,\|\nabla \log \pi(X)\| \leq \sqrt{\mathbb{E}[\nabla \log \pi(X)'\nabla \log \pi(X)]} \leq \sqrt{nL},$$

so that $\nabla \log \pi(X)$ is integrable. Hence Corollary 2.2 gives us that $\mathbb{E}\nabla \log \pi = -\mathbb{E}\nabla 1 = 0$. Next,

$$\begin{aligned}\|\nabla \log \pi \nabla \log \pi'\|_F &= \sqrt{\mathrm{Tr}(\nabla \log \pi \nabla \log \pi' \nabla \log \pi \nabla \log \pi')} \\ &= \sqrt{\mathrm{Tr}(\nabla \log \pi' \nabla \log \pi \nabla \log \pi' \nabla \log \pi)} = \nabla \log \pi' \nabla \log \pi.\end{aligned}$$

Hence, $\nabla \log \pi(X)\nabla \log \pi(X)'$ is integrable. Moreover, $\nabla^2 \log \pi(X)$ is bounded, and hence is also integrable. Therefore the Jacobian version Corollary 2.2 gives us that

$$\mathop{\mathbb{E}}_{X\sim\pi}[\nabla \log \pi(X)\nabla \log \pi(X)'] = -\mathop{\mathbb{E}}_{X\sim\pi}[\nabla^2 \log \pi].$$

$\square$

Theorem 2.2 (If π is a Grad-Log-Lipschitz Density then $[\nabla \log \pi]_\sharp \pi$ is Sub-Gaussian)**.** *Let π be a probability density on $\mathbb{R}^n$ such that $\nabla \log \pi$ is L-Lipschitz. If $X \sim \pi$ then $\nabla \log \pi(X)$*

is sub-Gaussian with proxy-variance L:

$$\psi(t) := \mathop{\mathbb{E}}_{X\sim\pi} \exp(\langle t,\ \nabla \log \pi(X)\rangle) \leq \exp(L \|t\|^2 /2)$$

Proof. We first prove the result with a sub-optimal, dimension dependent, sub-Gaussian constant. We then refine the result to the form stated in the theorem.

Since $\nabla \log \pi$ is L-Lipschitz, then we must have that $\|\nabla^2 \log \pi(x)\| \leq L$ (and hence $|\Delta \log \pi(x)| \leq nL$ as well) for all $x \in \mathbb{R}^n$. Let μ_n denote the nth moment of $\|\nabla \log \pi(X)\|$. As in the proof of Lemma 2.8,

$$\mathop{\mathbb{E}}_{X\sim\pi} \left[\|\nabla \log \pi(X)\|^2\right] = \mathop{\mathbb{E}}_{X\sim\pi} \left[\nabla \log \pi(X)'\nabla \log \pi(X)\right] = -\mathop{\mathbb{E}}_{X\sim\pi} \left[\Delta \log \pi(X)\right] \leq nL$$

For $r \geq 2$,

$$\mu_{2r} = \mathop{\mathbb{E}}_{X\sim\pi} \left[\|\nabla \log \pi(X)\|^{2r}\right] = \mathop{\mathbb{E}}_{X\sim\pi} \left[(\nabla \log \pi(X)'\nabla \log \pi(X)) \|\nabla \log \pi(X)\|^{2r-2}\right].$$

Note that

$$(\nabla \log \pi(X)'\nabla \log \pi(X)) \|\nabla \log \pi(X)\|^{2r-2}$$

is non-negative, so it must be extended integrable.

We need to check that $\operatorname{div}\left(\nabla \log \pi(X) \|\nabla \log \pi(X)\|^{2(r-1)}\right)$ is integrable as well in order to apply Corollary 2.2. Note that

$$\begin{aligned}
&\operatorname{div}\left(\nabla \log \pi(X) \|\nabla \log \pi(X)\|^{2(r-1)}\right) \\
&\quad = \nabla \log \pi(X)'\ 2(r-1) \|\nabla \log \pi(X)\|^{2(r-2)} \nabla^2 \log \pi(X)\nabla \log \pi(X) \\
&\qquad + (\Delta \log \pi(X)) \|\nabla \log \pi(X)\|^{2(r-1)}.
\end{aligned}$$

Thus

$$\left|\operatorname{div}\left(\nabla \log \pi(X) \|\nabla \log \pi(X)\|^{2(r-1)}\right)\right| \leq L(2(r-1)+n) \|\nabla \log \pi(X)\|^{2(r-1)}.$$

Now, if μ_{2r-2} is finite then $\operatorname{div}(\nabla \log \pi(X)\|\nabla \log \pi(X)\|^{2(r-1)})$ is absolutely integrable.

Hence, using (the divergence version of) Corollary 2.2,

$$\begin{aligned}
\mu_{2r} &= -\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla\log\pi(X)'\ 2(r-1)\left\|\nabla\log\pi(X)\right\|^{2(r-2)}\nabla^2\log\pi(X)\nabla\log\pi(X)\right] \\
&\quad -\mathop{\mathbb{E}}_{X\sim\pi}\left[(\Delta\log\pi(X))\left\|\nabla\log\pi(X)\right\|^{2(r-1)}\right] \\
&\le L(2(r-1)+n)\mathop{\mathbb{E}}_{X\sim\pi}\left[\left\|\nabla\log\pi(X)\right\|^{2(r-1)}\right] \\
&= L(2(r-1)+n)\mu_{2(r-1)} \\
&\le (n+2)Lr\mu_{2(r-1)}
\end{aligned}$$

By induction, we find that $\mu_{2r} \le r![(n+2)L]^r$. Hence, from [15, Theorem 2.1], we get that $\|\nabla\log\pi(X)\|$ is sub-Gaussian with proxy variance $4(n+2)L$. Thus we know that $\nabla\log\pi(X)$ must be a sub-Gaussian vector, with proxy variance no larger than $4(n+2)L$.

Now that we know that $\nabla\log\pi(X)$ is sub-Gaussian, we know that its moment generating function is entire. This allows us us to refine our analysis to get a dimension-free sub-Gaussian constant. Fix $t\in\mathbb{R}^n$. The moment generating function of $\nabla\log\pi(X)$ is finite everywhere and is given by:

$$\psi(t) := \mathop{\mathbb{E}}_{X\sim\pi}\left[\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right]$$

Then,

$$\begin{aligned}
\nabla_t\psi(t) &= \nabla_t\mathop{\mathbb{E}}_{X\sim\pi}\left[\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right] \\
&= \mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla_t\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right] = \mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla\log\pi(X)\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right]
\end{aligned}$$

By Cauchy-Schwartz,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\pi}\left[\|\nabla\log\pi(X)\|\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right] \\
&\qquad\le\sqrt{\mathop{\mathbb{E}}_{X\sim\pi}\left[\|\nabla\log\pi(X)\|^2\right]\mathop{\mathbb{E}}_{X\sim\pi}\left[\exp(\langle 2t,\ \nabla\log\pi(X)\rangle)\right]} < \infty,
\end{aligned}$$

hence $\nabla\log\pi(X)\exp(\langle t,\ \nabla\log\pi(X)\rangle)$ is absolutely integrable. Moreover,

$$\left\|(\nabla^2\log\pi(X)t)\exp(\langle t,\ \nabla\log\pi(X)\rangle)\right\| \le L\,\|t\|\exp(\langle t,\ \nabla\log\pi(X)\rangle),$$

so $(\nabla^2 \log \pi(X)t) \exp(\langle t,\ \nabla \log \pi(X)\rangle)$ must also be absolutely integrable. Therefore, using (the Jacobian version of) Corollary 2.2,

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla \log \pi(X) \exp(\langle t,\ \nabla \log \pi(X)\rangle)\right] &= -\mathop{\mathbb{E}}_{X\sim\pi}\left[(\nabla^2 \log \pi(X)t) \exp(\langle t,\ \nabla \log \pi(X)\rangle)\right] \\ &\in L\psi(t)B_t\end{aligned}$$

where B_t is the ball of radius $\|t\|$ centred at the origin. This gives us a differential inequality which is easily solved:

$$\nabla_t \log \psi(t) \in LB_t \implies \log \psi(t) \leq L\left\|t\right\|^2/2 \implies \psi(t) \leq \exp(L\left\|t\right\|^2/2)$$

□

Remark 2.3. *Consequently all the moments of* $\nabla \log \pi$ *exist. Moreover, since* $\|\nabla^2 \log \pi\| \leq L$*, all the moments of* $\nabla^2 \log \pi$ *must exist as well. This means that the assumptions in Roberts et al. [94], Neal and Roberts [81], Bédard [11], etc. that* $\mathbb{E}\left(\frac{\pi'}{\pi}\right)^8 < \infty$ *(or similar moment conditions) and* $\mathbb{E}\left(\frac{\pi''}{\pi}\right)^4 < \infty$ *are redundant once* $\frac{\pi'}{\pi}$ *is assumed to be Lipshitz.* ◁

Chapter 3

Optimal Shaping and Scaling of the Random Walk Metropolis Algorithm

3.1 Introduction

Markov Chain Monte Carlo (MCMC) algorithms are a common tool for estimating expectations with respect to a arbitrary probability measures (the “target”). These methods operate by defining a Markov Chain whose stationary distribution is the target, and whose dynamics are easily computable. Running this Markov chain forwards in time yields a dependent sequence of samples which can be used to estimate expectations. Performance of such algorithms are typically measured based on how quickly empirical expectations will converge to their target values. Among the simplest of such algorithms is Random-Walk Metropolis (RWM), which proposes IID increments (from a “proposal distribution”) which are either accepted or rejected with probabilities tuned so that the stationary measure matches the target distribution. Proposals which land in areas with low target density are likely to be rejected, while those that land in areas with higher density are likely to be accepted.

The choice of proposal distribution is they key tuning parameter in the design and application of RWM algorithms and has a decisive impact on the performance of the algorithm, especially in a high dimensional setting. A typical choice is to use a mean-zero Gaussian proposal, yet among this class one is still required to select the variance-covariance matrix of the proposal. Proposing steps which are too large in any particular direction will lead to poor performance due to frequent rejection, as the proposed point will typically have low target density. Proposing steps which are too small in any particular direction will lead to poor performance, since it will take many steps to move a meaningful distance in any direction. A step size and orientation which is "just right" (not too big, and also not too small, in each direction) is required for good performance. This leads us to consider the optimal shaping and scaling for Gaussian proposals for the RWM algorithm.

The seminal paper of [94] introduced techniques for analyzing the optimal scaling problem in the limit, as the dimension of the target tends to infinity, for independent and identically distributed targets (IID targets). The key insight was that, under appropriate rescaling, the random paths of any single component converge in law to the random path of a diffusion process weakly in the Skorohod topology, and that the speed of the limiting diffusion can be optimized using elementary techniques. Since that work, there has been a reasonable amount of attention placed on extending their results to other MCMC algorithms, as well as to more general targets.

3.1.1 Contributions

In this work we derive the scaling limit of RWM with block-independent targets with possibly complex dependence structures within blocks, and anisotropic proposals. We show that the random path of a full dependent block converges in law to the path of a multivariate anisotropic diffusion. We also show that the entire random path in $\mathbb{R}^{\mathbb{N}}$ converges in law to an infinite product of block-independent multivariate anisotropic diffusions.

Using this scaling limit, we aim address both the optimal scaling and the shaping of the proposal under convergence of the joint process. We find that the optimal scaling for a fixed proposal covariance shaping is the same as given by Roberts et al. [94], to tune the acceptance rate to be approximately 0.234. *Thus, this recommendation is independent of whether the anisotropy of proposal covariance aligns in any particular was with the shape* *of the target distribution.* We also provide a variational characterization of the optimal shaping matrix. The optimal proposal anisotropy matrix (called the *shaping* matrix in the sequel) is given by

$$\operatorname*{arg\,max}_{\Lambda \succ 0} \inf_{\substack{f \in \mathcal{D}(G) \\ \operatorname*{Var}_{X\sim\pi}[f(X)]=1}} \frac{\mathbb{E}_{X\sim\pi}\left[\nabla f(X)' \, \Lambda \, \nabla f(X)\right]}{\mathbb{E}_{X\sim\pi}\left[\nabla \log \pi(X)' \, \Lambda \, \nabla \log \pi(X)\right]}, \tag{3.1}$$

where π is the target distribution, $\mathcal{D}(G)$ is the domain of the infinitesimal generator of the limiting diffusion process of a single block, and $\Lambda \succ 0$ means that Λ ranges over symmetric strictly positive definite matrices. We show that when the blocks are rotation-scalings of independent and identical components, that this can be solved yielding the recommendation from Roberts and Rosenthal [97], to tune RWM so that the covariance of the proposals is proportional to the covariance of the target distribution. More generally, we show that this recommendation optimizes the instantaneous autocorrelation of linear functions in the the scaling limit. Finally, we provide conditions under which high-dimensional dependence in the target distribution will cause RWM performance under optimal shaping and scaling to deteriorate relative to an IID target. This supports the intuition that RWM performance is worse under complex dependence structures.

3.1.2 Outline of this chapter

A summary of prior work is given in Section 3.1.3. Section 3.1.4 sets up the weak convergence and optimal scaling/shaping problems, and also provides notation and definitions used through out this chapter. That subsection also provides a list of consequences of the assumption on the target distribution used to prove weak convergence (that, when π is the target density, $\nabla \log \pi$ is Lipschitz), many of which were demonstrated in Chapter 2.

Section 3.2 includes the main contributions of this work. In particular, Section 3.2.1 states the weak convergence results for the finite dimensional and the infinite dimensional processes. Section 3.2.2 provides the optimal scaling of the RWM proposals for a fixed shaping. Section 3.2.3 formulates the optimal shaping problem in terms of the spectral gap of the generator, and provides a variational characterization of the optimal shaping matrix. Section 3.2.4 presents the optimal shaping problem in terms of the spectral gap of the generator for certain special target distributions for which it is analytically tractable. Section 3.2.5 presents the optimal shaping in terms of short term autocorrelations for more general target distributions, and demonstrates that this alternative objective upper bounds the spectral gap, providing "speed limits" on the performance of RWM algorithms. Lastly among the key results, Section 3.2.6 provides a discursive analysis of the implications of the derived speed limits upon the performance decay of RWM in scenarios of high-dimensional dependence, relative to the independent target case.

3.1.3 Prior work

The seminal paper using scaling limits to address the optimal scaling problem in MCMC is that of Roberts et al. [94]. They consider IID targets of the form $\pi^{\otimes d}$ as $d \to \infty$, where π is a density on $\mathbb{R}^1$ with $D \log \pi$ Lipschitz continuous[1] They provide a scaling limit for the RWM algorithm with spherical Gaussian proposals, which establishes weak convergence of the first component's path process to that of a univariate Langevin diffusion, and derives an optimal scaling criteria of accepting $\approx 23.4\%$. The paper has additional regularity assumptions of smoothness (that the density is twice continuously differentiable) and moment conditions (that $\underset{X\sim\pi}{\mathbb{E}}(D \log \pi(X))^8 < \infty$ and $\underset{X\sim\pi}{\mathbb{E}}(\frac{D^2\pi}{\pi})^4 < \infty$).

The subsequent work of Roberts and Rosenthal [98], derives similar optimal scaling results for the Metropolis Adjusted Langevin Algorithm (MALA). They also consider IID targets of the form $\pi^{\otimes d}$ as $d \to \infty$ as well, where π is a density on $\mathbb{R}^1$ with $D \log \pi$ Lipschitz continuous. That work proves weak convergence of the first component's path process to that of a univariate Langevin diffusion, and derives an optimal scaling criteria of accepting $\approx 57.4\%$ of proposals when using MALA. They require additional regularity assumptions of smoothness (that the density is eight times continuously differentiable), a growth assumption (that the first eight derivatives of $\log \pi$ are all bounded by a polynomial) and moment conditions (that all polynomial moments of π are finite; $\forall k \in \mathbb{N}(\mathbb{E}_{X\sim\pi}(X^k) < \infty)$.

The survey paper of Roberts and Rosenthal [97], provides further context to the optimal scaling problem and presents both theoretical and empirical results. In addition to

[1] D denotes the first derivative operator.

summarizing previous work, the paper provides an examination of how the optimal scaling in finite dimensions approaches the infinite dimensional limits derived via diffusion limits. Lastly, and a large inspiration for this work, they consider extensions to independent products which differ only by heterogeneity of scale. This provides the first optimal shaping result we are aware of. They note that "this result does not appear in any of the MCMC scaling literature, so we have sketched a proof which appears in the Appendix," however the sketched proof considers only convergence of a single component and so the impact of optimal scaling and shaping on the mixing properties of the full multidimensional target remains ambiguous. Our present paper builds on their ideas to provide multivariate convergence and optimal scaling and shaping results which apply to the full multidimensional limit.

The work of Neal and Roberts [81] considers modified RWM and MALA where only a fraction of the components are updated at a time. Algorithms of that type are typically more efficient as an update which would have been rejected because of a single "bad proposal" in one component is not going to affect the speed of all dimensions. That paper derives the optimal scaling and update rate simultaneously with the same assumptions as Roberts et al. [94]. Their proof of weak convergence in the Skorohod topology had more exposition, detail and precision than that of previous work, and largely inspired our proof.

The work of Bédard [11] and Bédard and Rosenthal [13], and the Ph.D. thesis of Bédard [12] consider a more extreme version of the scale homogeneity problem for the RWM algorithm. Particularly they address the case that the scaling of various components shrink or grow at disparate rates as the dimension tends to infinity. That collection of work shows that, depending on which scalings are dominant, the limiting law of the first component may be either a univariate RWM process or a Langevin diffusion, and that in certain situations the optimal acceptance rate will be quite different than the 23.4% of the homogeneous or limited inhomogeneity cases. That work also slightly relaxed the assumptions of the original paper of Roberts et al. [94] by reducing the powers in their moment assumptions.

More recently, some authors have considered working with infinite dimensional targets, particularly in the case that the target has a density with respect to the law of a Gaussian process. This includes the work of Mattingly et al. [71] which covers the RWM case and that of Pillai et al. [90] which covers the MALA case. These papers allow for a non-trivial dependence structure, but only under the strong assumption of absolute continuity with respect to an infinite-dimensional Gaussian distribution. They show that the $\approx$ 23.4% and $\approx$ 57.4% optimal acceptance rates for RWM and MALA respectively carry over to infinite dimensional distributions which have densities with respect to the laws of a Gaussian processes. Though these papers allow for a non-trivial dependence structure, they do not consider the optimal shaping problem.

Lastly, Zanella et al. [126] utilize the theory of Dirichlet forms to establish weak convergence of the infinite dimensional limit process for targets of the same form as considered by Roberts et al. [94]. Using the powerful theory of Mosco convergence, they are able to eliminate many of the assumptions needed by Roberts et al. [94]. In particular, that paper requires no additional smoothness or moment assumptions. In fact, they are able to demonstrate convergence of the Markov semigroup with assumptions on $D \log \pi$ which are weaker than Lipschitz continuity, though to ensure weak convergence of the path processes they do require Lipschitz continuity. Hence, the present chapter's assumptions used to demonstrate weak convergence of the infinite dimensional paths are the same as in that work.

3.1.4 Notation and Definitions

Let π be the Lebesgue density of a probability distribution on $\mathbb{R}^k$. For each $d \in \mathbb{N}$, let $\Pi_d = \pi^{\otimes d}$. Then Π_d is the joint density for d independent blocks, each of dimension k, identically distributed according to π. Let $\Lambda \in \mathbb{R}^{k \times k}$ be a symmetric positive definite covariance matrix representing a RWM proposal shaping, and let $l > 0$ representing the RWM proposal scaling. Let $\Lambda_r = I_r \otimes \Lambda$ for $r \in \mathbb{N}$, where I_r is the $r \times r$ identity matrix. The "accelerated, continuous time" stationary RWM process with stationary distribution Π_d, and proposal distribution:

$$\mathrm{N}_d(l^2\Lambda) := \mathrm{N}\left(0, \frac{l^2}{(d-1)}\Lambda_d\right) , \tag{3.2}$$

is the Markov process, $\mathbf{X}_d(t)$ with initial distribution $\mathbf{X}_d(0) \sim \Pi_d$ and infinitesimal generator

$$[\hat{G}_d^{l,\Lambda} f](x) = kd \underset{Z\sim \mathrm{N}_d(l^2\Lambda)}{\mathbb{E}} \left[(f(x+Z) - f(x)) \left(1 \wedge \frac{\Pi_d(x+Z)}{\Pi_d(x)} \right) \right] . \tag{3.3}$$

defined for $f \in \overline{C}(\mathbb{R}^{kd})$, where for a topological space S, $\overline{C}(S)$ denotes the space of bounded continuous functions $S \to \mathbb{R}$.

Note that $I_d \otimes \Lambda$ is the $kd \times kd$ block diagonal matrix with d blocks of size $k \times k$ all equal to Λ:

$$\Lambda_d = I_d \otimes \Lambda = \begin{bmatrix} \Lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & \Lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \Lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \Lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & \Lambda \end{bmatrix} \tag{3.4}$$

Equivalently, $\mathbf{X}_d(t)$ is the pure-jump Markov process with jump intensity in state x given by

$$kd \underset{Z\sim \mathrm{N}_d(l^2\Lambda)}{\mathbb{E}} \left[1 \wedge \frac{\Pi_d(x+Z)}{\Pi_d(x)} \right], \tag{3.5}$$

and jumps leaving state x distributed according the conditional distribution of a proposal given that it is accepted.

Let $\mathbf{X}_d^{(i)}(t)$ be the stochastic process on $\mathbb{R}^k$ consisting of the ith k-dimensional block of $\mathbf{X}_d(t)$. In general, this process is not Markov because the intensity depends on the full state and does not factor. For $i < j$, let $\mathbf{X}_d^{(i):(j)}(t)$ be the stochastic process consisting of the ith, $(i+1)$th,...,jth k-dimensional blocks of $\mathbf{X}_d(t)$, so that $\mathbf{X}_d^{(i):(j)}(t)$ has paths which take values in $\mathbb{R}^{k(j-i+1)}$.

For each $r \in \mathbb{N}$, the anisotropic Langevin diffusion with stationary distribution Π_r, anisotropy matrix Λ, and time-scaling factor $l^2 a_\Lambda(l)$, $\mathbf{X}^r(t)$, is the Markov process with $\mathbf{X}^r(0) \sim \Pi_r$ and infinitesimal generator

$$[G_r^{l,\Lambda} f] = kl^2 a_\Lambda(l) \left(\frac{1}{2} \Lambda_r : (\nabla^2 f) + \frac{1}{2} [\nabla \log \Pi_r]' \Lambda_r (\nabla f) \right), \tag{3.6}$$

for a sufficiently large class of functions f, and where

$$a_\Lambda(l) = 2\Phi\left(-\frac{l\sqrt{\Sigma : \Lambda}}{2}\right) \quad \Sigma = \operatorname*{Var}_{X\sim\pi}(\nabla \log \pi(X)) \quad A : B = \mathrm{Tr}(A'B)\ . \tag{3.7}$$

Equivalently, it is the diffusion process with initial distribution Π_r satisfying the stochastic differential equation (SDE):

$$d\mathbf{X}^r(t) = kl^2 a_\Lambda(l)\Lambda_r[\nabla \log \Pi_r(\mathbf{X}^r(t))]\ dt + \sqrt{2kl^2 a(l)\Lambda_r}\ d\mathbf{B}(t) \tag{3.8}$$

where $\mathbf{B}(t)$ is a standard kr-dimensional Wiener process, and $\sqrt{\Lambda_r}$ is the symmetric positive definite square-root of the symmetric positive definite matrix Λ_r. Thus, $\mathbf{X}^r$ is the same process, in distribution, as r independent copies of the $\mathbf{X}^1$ appended together. Let $\mathcal{D}(G^\Lambda)$ be the domain of the generator G^Λ. Note that $\mathcal{D}(G^\Lambda) = \mathcal{D}(G^\Theta)$ for Λ and Θ both symmetric and strictly positive definite. Thus, without ambiguity, we denote this common domain by $\mathcal{D}(G) := \mathcal{D}(G^{I_k})$.

Later, in the case $r = 1$, we will also compare this to the generator of a similar diffusion, with the same stationary measure and anisotropy matrix, at a standardized speed

$$[G^\Lambda f](x) = k\left(\frac{\Lambda}{\Lambda : \Sigma} : (\nabla^2 f) + [\nabla \log \pi]'\frac{\Lambda}{\Lambda : \Sigma}(\nabla f)\right) \tag{3.9}$$

The choice of time-scaling used for the standardized speed corresponds to $G_1^{l,\Lambda}$ for the optimal choice of l given π and Λ up to universal constants (not dependent on k, π, or Λ), as we show in Corollary 3.1.

We make the following assumption about π throughout this work:

Assumption 3.1. $\nabla \log \pi$ *is L-Lipschitz continuous for some $L > 0$.*

Some geometric and analytic consequences of this assumption are summarized in Section 3.3.

3.2 Results

3.2.1 Weak Convergence in the Skorohod Topology

Theorem 3.1 (Weak convergence of finite dimensional processes in the Skorohod topology)**.** *Under the definitions above, if Assumption 3.1 holds then (for each $r \in \mathbb{N}$) then $\mathbf{X}_d^{(1):(r)}$ converges weakly in the Skorohod topology on $\mathbb{R}^{kr}$ to $\mathbf{X}^r$ as $d \to \infty$.*

The proof of this result is in Section 3.4. By bootstrapping our result on weak convergence of finite dimensional processes we are also able to demonstrate weak convergence of the infinite dimensional process.

Theorem 3.2 (Weak convergence of the infinite dimensional process in the Skorohod topology)**.** *There is a unique (in law) process $\mathbf{X}^{\mathbb{N}}$ taking values in $\mathbb{R}^{\mathbb{N}}$ such that the marginal process of the first kr components has the same law as $\mathbf{X}^r$. Let $\mathbf{Y}_d(t) = (\mathbf{X}_d(t), 0, 0, ...)$, so that $\mathbf{Y}_d(t) \in \mathbb{R}^{\mathbb{N}}$ for each $d \in \mathbb{N},\ t > 0$. If Assumption 3.1 holds then $\mathbf{Y}_d$ converges weakly to $\mathbf{X}^{\mathbb{N}}$ in the Skorohod topology of $\mathbb{R}^{\mathbb{N}}$, where $\mathbb{R}^{\mathbb{N}}$ is endowed with the product topology.*

The proof of this result is in Section 3.6. The processes $\mathbf{Y}_d(t)$ are similar to the processes considered by Zanella et al. [126] to derive their infinite dimensional scaling limit.

3.2.2 Optimal Scaling Under a Fixed Shaping

For the rest of Section 3.2, for simplicity, we assume that $r = 1$ so that we consider only the limiting dynamics of a single block. We are able to do this without loss of generality, and the results carry over to multiple blocks and to the infinite dimensional limit, because of tensorization properties of spectral gaps, as discussed by Bakry et al. [7].

Having shown that the limiting process is a Langevin diffusion, it is natural to try to select the tuning parameters, (l, Λ) such that the limiting diffusion mixes as quickly as possible. For a fixed choice of Λ, if we change l then we only change the time scaling of the process. That is, for different values of l, we are running a diffusion with the dynamics given by G^{Λ} accelerated by a factor of $l^2 a_{\Lambda}(l)(\Lambda : \Sigma)/2$.

Thus we find that the optimal choice of l for a fixed Λ is easy to determine; we need only maximize the time-change factor $l^2 a_{\Lambda}(l)$ in order to make the diffusion move towards stationarity as quickly as possible. As in Roberts et al. [94] we will characterize the optimal scaling both in terms of the value of the scaling factor, l, and in terms of the limiting average acceptance probability for the RWM algorithm. The optimal choice of Λ will prove more challenging to derive as changing Λ does not induce only a time-change on the dynamics of the process.

Lemma 3.1 (Limiting Acceptance Rate)**.** *The limiting acceptance rate for the RWM proposals of $\mathbf{X}_d$ is $a_\Lambda(l)$. That is to say:*

$$\lim_{d\to\infty} \underset{\substack{X\sim\Pi_d \\ Z\sim \mathrm{N}_d(l^2\Lambda)}}{\mathbb{E}} \left[1 \wedge \frac{\Pi_d(X+Z)}{\Pi_d(X)}\right] = a_\Lambda(l) = 2\Phi\left(-\frac{l\sqrt{\Sigma : \Lambda}}{2}\right) \tag{3.10}$$

where $\Sigma = \underset{X\sim\pi}{\mathrm{Var}}[\nabla \log \pi(X)]$ and $(:)$ is the Frobenius inner product.

This result is proved as a step in the proof of Theorem 3.1, in Section 3.4.

Corollary 3.1 (Optimal Scaling of l for fixed Λ)**.** *The optimal scaling over l for a fixed Λ is $l_\Lambda \approx \frac{2.38}{\sqrt{\Sigma:\Lambda}}$. This is the l_Λ which solves $a(l_\Lambda) \approx 0.234$, as in the original Roberts et al. [94] result. The limiting diffusion corresponds to G^Λ sped up (or slowed down) by a factor of is $l_\Lambda^2 a(l_\Lambda)(\Lambda : \Sigma)/2 \approx 0.66$. This acceleration factor is universal; it does not depend on k, π, or Λ.*

The proof of this result may be found in Section 3.7.

3.2.3 Optimal Shaping I: Variational Characterization via Spectral Gaps

For the rest of this section, we work only with G^Λ. This is equivalent to assuming that the optimal scaling is always used for a given choice of shaping: $l = l^\Lambda$, and the universal acceleration factor $4\tilde{h}(\omega_\star) \approx 0.66$ is ignored. We note that, since $\mathbf{X} := \mathbf{X}^1$ is Feller (Lemma 3.4) then the spectrum of G^Λ is a subset of the non-positive real line, and we note that there is an eigenvalue at 0 corresponding to the constant function.

Definition 3.1 (Spectral Gap of an Infinitesimal Generator)**.** *If L is the infinitesimal generator of Markov process with stationary measure π, then the* spectral gap *of G is given* *by*

$$\rho(L) = \sup\left\{\rho > 0 \ s.t.\ \forall f \in \mathcal{D}(L) \left(\underset{X\sim\pi}{\mathrm{Var}}\, f(X) \leq -\frac{1}{\rho} \underset{X\sim\pi}{\mathbb{E}} \left(f(X)\ [Lf](X)\right)\right)\right\} \tag{3.11}$$

If $\rho(L) > 0$ *then we say that* L has a spectral gap*.*

Lemma 3.2 (Co-occurrence of spectral gaps)**.** *Either all of the generators in the set* $\{G^\Lambda \ s.t.\ \Lambda \succ 0\}$ *have a spectral gap or none of them do.*[2]

The proof of this result may be found in Section 3.7.

3.2.3.1 Variational Characterization of Optimal Shaping

When G^Λ has a spectral gap for at least one strictly positive definite Λ, then the ideal choice of shaping is that which maximizes the spectral gap of G^Λ. This would in turn optimize the exponential rate of convergence to stationarity of the diffusion process (see, for example, [7]). Thus, we add the following assumption when needed, in order to ensure that the optimization over Λ can be meaningfully reduced to the optimization of the spectral gap of G^Λ.

Assumption 3.2. G^{I_k} *has a spectral gap.*

In light of Lemma 3.2 this is equivalent to assuming that G^Λ has a spectral gap for at least one $\Lambda \succ 0$, and that G^Λ has a spectral gap for all $\Lambda \succ 0$.

Theorem 3.3 (Variational Characterization of Optimal Shaping)**.** *Under Assumptions 3.1 and 3.2 the optimal shaping matrix is given by*

$$\begin{aligned}
\Lambda^\star &\in \underset{\Lambda\succ 0}{\arg\max} \inf_{\substack{f\in\mathcal{D}(G)\\ \underset{X\sim\pi}{\mathrm{Var}}[f(X)]=1}} \frac{\underset{X\sim\pi}{\mathbb{E}}\left[\nabla f(X)'\ \Lambda\ \nabla f(X)\right]}{\underset{X\sim\pi}{\mathbb{E}}\left[\nabla \log\pi(X)'\ \Lambda\ \nabla\log\pi(X)\right]} \\
&\equiv \underset{\Lambda\succ 0}{\arg\max} \inf_{\substack{f\in\mathcal{D}(G)\\ \underset{X\sim\pi}{\mathrm{Var}}[f(X)]\neq 0}} \frac{\underset{X\sim\pi}{\mathbb{E}}\left[\nabla f(X)'\ \Lambda\ \nabla f(X)\right]}{(\Lambda : \Sigma)\ \underset{X\sim\pi}{\mathrm{Var}}\, f(X)}
\end{aligned} \tag{3.12}$$

[2] Recall that $\Lambda \succ 0$ means that Λ ranges over symmetric strictly positive definite matrices.

Remark 3.1 (The spectral gap assumption is satisfiable)**.** *The curvature-dimension condition of Bakry et al. [7] provides one way to verify assumption (A2). A simple example is that if* $\log \pi$ *is strongly concave, then the condition is satisfied with*

$$\frac{1}{\rho_\Lambda} = \operatorname*{ess\,inf}_{x \in \mathbb{R}^k} \lambda_1(-\nabla^2 \log \pi(x)) > 0$$

for all Λ *(where* λ_1 *is the function which returns the minimal eigenvalue of a matrix. In this case,* $\frac{1}{\rho_\Lambda}$ *is the strong convexity parameter).* ◁

3.2.4 Optimal Shaping II: Optimal Spectral Gaps in Special Cases

As mentioned before, the optimal shaping problem turns out to be a much more difficult than the optimal scaling was. We solve this problem exactly, first when the target is a multivariate normal distribution, and second when the target density is a rotated independent product of a scale family.

For more general target distributions, the problem of optimizing the spectral gap is not so easily approachable as it is not known in general how to directly compute the spectral gap of the generator for a Langevin diffusion process (or even to determine sharp conditions for when there is a spectral gap at all) or how the spectrum transforms under a change in anisotropy. Instead, we optimize a surrogate measure of the process' speed: the rate of decay of autocorrelations of functions of $\mathbf{X}$ near lag-0. As will be discussed in the next section, this is a continuous time analogue of a common heuristic for the short term mixing properties of MCMC algorithms in discrete time, as well as a relaxation of the variational formula for the spectral gap problem which we would strive to solve if we could. This surrogate will also give novel 'speed limits' on the convergence of RWM—upper bounds on the spectral gap of the generator which limit the convergence rate in terms of properties of π.

Theorem 3.4 (Optimal Shaping when $\pi \equiv \mathrm{N}(\mu, \Gamma)$)**.** *When* π *is the density of a* $\mathrm{N}(\mu, \Gamma)$ *distribution, then* $\Sigma := \operatorname*{Var}_{X\sim\pi}(\nabla \log \pi(X)) = \Gamma^{-1}$ *and the spectral gap of* G^Λ *is maximized by taking the shaping matrix to be (proportional to) the covariance of the target distribution;* $\Lambda = \Sigma^{-1} = \Gamma$*. The convergence to stationarity is* $\frac{\mathrm{Tr}(\Sigma)}{k\lambda_1(\Sigma)}$ *times faster when using the optimal* *shaping as opposed to spherical shaping, where* $\lambda_1(\Sigma)$ *is the minimal eigenvalue of* Σ*.*

Proof. Let $\sigma(A)$ denote the spectrum of the operator A.

Without loss of generality, $\mu = 0$. In this case, $\nabla \log \pi(x) = -\Gamma^{-1}x$, so

$$\Sigma = \mathop{\mathbb{E}}_{X\sim\pi}(\Gamma^{-1}XX'\Gamma^{-1}) = \Gamma^{-1} \ . \tag{3.13}$$

Now, we note that:

$$\begin{aligned}[G^\Lambda f](x) &= \frac{k}{\Lambda:\Sigma}\left(\Lambda : \nabla^2 f(x) + (-x'\Gamma^{-1})\Lambda\nabla f(x)\right) \\ &= \frac{k}{\Lambda:\Sigma}\left(\Lambda : \nabla^2 f(x) + x'(-\Sigma\Lambda)\nabla f(x)\right)\end{aligned} \tag{3.14}$$

From Metafune et al. [75] we know that

$$\sigma(G^\Lambda) = \left\{ \sum_{s\in\sigma(B)} sn_s \text{ s.t. } \ n_s \in \mathbb{N}\cup\{0\} \ \forall s\in\sigma(B) \right\} \ , \tag{3.15}$$

where $B = \frac{-k\Sigma\Lambda}{\Lambda:\Sigma}$. Therefore the spectral gap of G^Λ is exactly the smallest eigenvalue of $\frac{k\Sigma\Lambda}{\Lambda:\Sigma}$. Now, letting $A = \Sigma\Lambda$, and letting $\{\lambda_i(A)\}_{i=1}^k$ be the (non-decreasing) eigenvalues of A we can solve:

$$\arg\max_A \frac{\lambda_1(A)}{\sum_{i=1}^k \lambda_i(A)} \tag{3.16}$$

This function is bounded above by $1/k$ since $a_1 \le a_i$ for all $1 \le i \le k$ and the function is equal to $1/k$ if and only if $A = \gamma I$ for some $\gamma \ne 0$. Hence the optimal spectral gap is achieved at $\Sigma\Lambda = I$. Therefore $\Lambda^\star = \Sigma^{-1} = \Gamma$.

We also find that the spectral gap of $G^{\Lambda^\star}$ is 1. On the other hand, the spectral gap of G^I is $\frac{k\lambda_1(\Sigma)}{\mathrm{Tr}(\Sigma)} = \frac{\lambda_1(\Sigma)}{\overline{\lambda}(\Sigma)}$, where $\overline{\lambda}(\Sigma)$ is the average eigenvalue of Σ. Therefore, the convergence is $\frac{\overline{\lambda}(\Sigma)}{\lambda_1(\Sigma)}$ times faster when using the optimal shaping as opposed to spherical shaping. □

Theorem 3.5 (Optimal Shaping when π is a rotated independent product of a scale family)**.** *Suppose that* π_1 *is a probability density on* $\mathbb{R}$*, with* $D[\log\pi_1]$ *Lipschitz, and the one-dimensional generator*

$$G_1 f = \left(D^2[f] + D[\log\pi_1]D[f]\right) \tag{3.17}$$

satisfies assumption (A2) with spectral gap $\rho = \lambda^\star$.

Let $c_i > 0$ *for each* $1 \le i \le k$ *and let* Q *be a unitary transformation, so* $Q' = Q^{-1}$.

Let $\pi(x) = \prod_{i=1}^k c_i \pi_1(c_i e_i' Q x)$, *so that* $X \sim \pi$ *if and only if* $c_i e_i' Q X \overset{iid}{\sim} \pi_1$ *for* $1 \le i \le k$. *(Where* e_i *are the standard basis vectors).*

Then

$$\Sigma := \operatorname*{Var}_{X\sim\pi}(\nabla \log \pi(X)) = \sigma^2 Q' \operatorname{diag}(c_i^2) Q \tag{3.18}$$

where $\sigma^2 = \mathop{\mathbb{E}}_{X_1 \sim \pi_1}\left[D \log \pi_1(X_1)^2\right]$ *and the spectral gap of* G^Λ *is maximized (among those* Λ *of the form* $Q'BQ$ *with* B *diagonal) by* $\Lambda = \Sigma^{-1}$. *The convergence to stationarity is* $\frac{\operatorname{Tr}(\Sigma)}{k\lambda_1(\Sigma)}$ *times faster when using the optimal shaping as opposed to spherical shaping, where* $\lambda_1(\Sigma)$ *is the minimal eigenvalue of* Σ.

Proof. We first compute

$$\nabla \log \pi(x) = \sum_{i=1}^k c_i Q' e_i [D \log \pi_1](c_i e_i' Q x) \tag{3.19}$$

and

$$\begin{aligned}
&\mathop{\mathbb{E}}_{X\sim\pi} \nabla \log \pi(x) \nabla \log \pi(x)' \\
&\quad = \sum_{i=1}^k \sum_{j=1}^k c_i c_j Q' e_i e_j' Q \mathop{\mathbb{E}}_{X\sim\pi}[D \log \pi_1](c_i e_i' Q X)[D\log \pi_1](c_j e_j' Q X) \\
&\quad = \sum_{i=1}^k \sum_{j=1}^k c_i c_j Q' e_i e_j' Q \mathop{\mathbb{E}}_{Y\sim\pi_1^{\otimes k}}[D \log \pi_1](Y_i)[D\log \pi_1](Y_j) \\
&\quad = \sum_{i=1}^k \sum_{j=1}^k c_i c_j Q' e_i e_j' Q \delta_i^j \sigma^2 \\
&\quad = \sigma^2 Q' \operatorname{diag}(c_i^2) Q
\end{aligned} \tag{3.20}$$

Thus $\Sigma = \sigma^2 Q' \operatorname{diag}(c_i^2) Q$.

Suppose that $Q = I$. Consider the generator G_1 as in the theorem statement. G_1 generates a Feller semigroup with stationary measure π_1, and so there is a complete basis (for $L_2(\mathbb{R}, \pi_1)$) of eigenfunctions for H (see section 4.7 of Pavliotis [88]). Let these eigenfunctions be $\{\phi_\alpha\}_{\alpha\in\mathbb{N}\cup\{0\}}$ and the corresponding eigenvalues be $\{\lambda_\alpha\}_{\alpha\in\mathbb{N}\cup\{0\}}$.

Under the assumption that a spectral gap exists for G_1 with tight constant $\lambda^\star$, we may assume that $\lambda_0 = 0$, $\lambda_1 = \lambda^\star$ and $\lambda_\alpha \ge \lambda^\star$ for $\alpha \ge 2$.

Then $S = \left\{\prod_{i=1}^k \phi_{\alpha_i} \circ C_i \text{ s.t. } \alpha_i \in \mathbb{N}\ \forall i\right\}$ is a complete basis for $L_2(\mathbb{R}^k, \pi)$, where $C_i : x \mapsto c_i x_i$. Moreover, S contains only eigenvectors for G^Λ:

$$\begin{aligned}
&G^\Lambda\left[\prod_{i=1}^k \phi_{\alpha_i} \circ C_i\right] \\
&= \frac{k}{\Lambda : \Sigma} \sum_i \Lambda_{ii} c_i^2 \left[(D^2\phi_{\alpha_i} \circ C_i + (D \log \pi_1)(D[\phi_{\alpha_i} \circ C_i])) \prod_{j\neq i} \phi_{\alpha_j} \circ C_j\right] \\
&= \frac{k}{\Lambda : \Sigma} \sum_i \Lambda_{ii} c_i^2 \lambda_{\alpha_i} \left[\prod_{i=1}^k \phi_{\alpha_i} \circ C_i\right]
\end{aligned} \tag{3.21}$$

Therefore, $\sigma(G^\Lambda) = \left\{k\frac{\sum_{i=1}^k \Lambda_{ii} c_i^2 \lambda_{\alpha_i}}{\sum_{i=1}^k \Lambda_{ii} c_i^2} \text{ s.t. } \alpha_i \in \mathbb{N} \cup \{0\}\ \forall 1 \leq i \leq k\right\}$. Hence the spectral gap for G^Λ is the minimal eigenvalue of $\lambda^\star \frac{k\Lambda\Sigma}{\Lambda:\Sigma}$. The optimal spectral gap is thus achieved, as in Theorem 3.4, by $\Lambda = \Sigma^{-1}$.

For general Q, unitary, let $M_Q f = f \circ Q$. Then G^Λ has the same spectrum as $G_Q^\Lambda = M_Q^{-1} G^\Lambda M_Q$ since these are similar operators.

$$\begin{aligned}
[G_Q^\Lambda f](x) &= \frac{k\Lambda}{\Lambda : \Sigma} : \left(Q'\nabla^2 f(x) Q + \nabla \log \pi(Q'x) \nabla f(x)' Q\right) \\
&= \frac{k\Lambda}{\Lambda : \Sigma} : \left(Q'\nabla^2 f(x) Q + Q' \sum_{i=1}^k c_i [D \log \pi_1](c_i x_i) e_i \nabla f(x)' Q\right) \\
&= \frac{kQ\Lambda Q'}{\Lambda : \Sigma} : \left(\nabla^2 f(x) + \sum_{i=1}^k c_i [D \log \pi_1](c_i x_i) e_i \nabla f(x)'\right)
\end{aligned} \tag{3.22}$$

This generator is of the same form as $G^{Q\Lambda Q'}$ when $Q = I$, except with Σ replaced by $Q\Sigma Q'$. Thus the spectrum of this operator is optimized when $Q\Lambda Q' = (Q\Sigma Q')^{-1}$, which occurs exactly when $\Lambda = \Sigma^{-1}$. □

3.2.5 Optimal Shaping III: Decay of Autocorrelations and Speed Limits

In this section we first describe how the generator G^Λ is related to the slope of the autocorrelation of functions of $\mathbf{X}$. Then we consider a relaxation of the spectral gap problem to searching for smaller subspaces of $\mathcal{D}(G^\Lambda)$.

Lemma 3.3 (Relationship between autocorrelation and generators)**.** *For any* $f \in \mathcal{D}(G^\Lambda)$

which is not is not almost everywhere constant. Let $Y(t) = f(\mathbf{X}(t))$. Then

$$\begin{aligned} \frac{d}{dt}\operatorname{Cov}(Y(0), Y(t))\big|_{t=0^+} &= \mathop{\mathbb{E}}_{X\sim\pi}\left[f(X)\ [G^\Lambda f](X)\right] \\ &= -\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla f(X)' \frac{k\Lambda}{\Lambda:\Sigma}\nabla f(X)\right] \end{aligned} \tag{3.23}$$

and

$$\begin{aligned} \frac{d}{dt}\operatorname{Corr}(Y(0), Y(t))\big|_{t=0^+} &= \frac{\mathop{\mathbb{E}}_{X\sim\pi}\left[f(X)\ [G^\Lambda f](X)\right]}{\mathop{\mathrm{Var}}_{X\sim\pi}[f(X)]} \\ &= -\frac{\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla f(X)' \frac{k\Lambda}{\Lambda:\Sigma}\nabla f(X)\right]}{\mathop{\mathrm{Var}}_{X\sim\pi}[f(X)]} \end{aligned} \tag{3.24}$$

Hence, the spectral gap of G^Λ is given by

$$\begin{aligned} \lambda_\star^\Lambda &= \inf_{f\in C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)} \left|\frac{d}{dt}\operatorname{Corr}(Y(0), Y(t))\big|_{t=0^+}\right| \\ &= \inf_{f\in C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)} \frac{k \mathop{\mathbb{E}}_{X\sim\pi}[\nabla f(X)'\Lambda\nabla f(X)]}{\mathop{\mathrm{Var}}_{X\sim\pi}[f(X)]} \end{aligned} \tag{3.25}$$

Proof. From Itô's lemma, for $t > 0$:

$$\begin{aligned} Y(t) - Y(0) &= \int_0^t G^\Lambda f(\mathbf{X}(t))dt + \int_0^t \nabla f(\mathbf{X}(t))' \frac{k\Lambda}{\Lambda:\Sigma}\nabla \log\pi(\mathbf{X}(t))d\mathbf{B}(t) \\ &= \int_0^t G^\Lambda f(\mathbf{X}(t))dt + M_t \end{aligned} \tag{3.26}$$

In this expansion, M_t is a $\{\mathbf{B}(t), \mathbf{X}(0)\}$-martingale starting at 0, and hence is uncorrelated with Y(0) which is $\sigma(\mathbf{X}(0))$-measurable. Moreover, by Fubini's theorem,

$$\mathbb{E}\left[\int_0^t G^\Lambda f(\mathbf{X}(t))dt\right] = \int_0^t \mathbb{E}[G^\Lambda f(\mathbf{X}(t))]dt = 0 \tag{3.27}$$

where the last equality follows by integration by parts. Hence, using Fubini's theorem again:

$$\begin{aligned} \operatorname{Cov}(Y(0), Y(t)) &= \mathbb{E}\left[Y(0)\left(Y(0) + \int_0^t G^\Lambda f(\mathbf{X}(t))dt + M_t\right)\right] \\ &= \operatorname{Cov}(Y(0), Y(0)) + \int_0^t \mathbb{E}[f(\mathbf{X}(0)G^\Lambda f(\mathbf{X}(t))]dt \end{aligned} \tag{3.28}$$

Now, using the fundamental theorem of calculus,

$$\frac{d}{dt}\operatorname{Cov}(Y(0),Y(t))\Big|_{t=0^+} = \mathop{\mathbb{E}}_{X\sim\pi}\left[f(X)\ [G^\Lambda f](X)\right] \tag{3.29}$$

Applying integration by parts, we get:

$$\begin{aligned}\frac{d}{dt}\operatorname{Cov}(Y(0),Y(t))\Big|_{t=0^+} &= \mathop{\mathbb{E}}_{X\sim\pi}\left[f(X)\ [G^\Lambda f](X)\right]\\ &= -\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla f(X)'\frac{k\Lambda}{\Lambda:\Sigma}\nabla f(X)\right]\end{aligned} \tag{3.30}$$

Finally, the statement regarding correlations follows from the definition of correlation in terms of covariance. □

Since $C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)$ is dense in $\mathcal{D}(G^\Lambda)$, if assumption (A2) holds, then the spectral gap of G^Λ is exactly the worst-case (negative) lag-0 slope of the autocorrelation of functions in $C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)$.

Thus, for convenience of solution, one may consider in place of $C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)$ a smaller class of functions, $\mathcal{F}\subsetneq C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)$ over which to solve

$$\max_{\Lambda\succ 0}\min_{f\in\mathcal{F}}\frac{\mathop{\mathbb{E}}_{X\sim\pi}\left[\nabla f(X)'\frac{k\Lambda}{\Lambda:\Sigma}\nabla f(X)\right]}{\mathop{\operatorname{Var}}_{X\sim\pi}[f(X)]} \tag{3.31}$$

The solution to this relaxed problem may be interpreted as optimizing the worst case autocorrelation (in a neighbourhood of lag-0) among functions from $\mathcal{F}$. When $\mathcal{F}$ is itself a subspace of $C^2(\mathbb{R}^k)\cap\mathcal{D}(G^\Lambda)$, then the solution may be interpreted as optimizing the spectral gap of the restricted generator $G^\Lambda\big|_{\mathcal{F}}:\mathcal{F}\to G^\Lambda(\mathcal{F})$ which is also a linear operator (which may be bounded or unbounded, depending on the choice of $\mathcal{F}$). Obviously, the solution to the restricted problem, for some choice of $\mathcal{F}$, is in fact optimizing an upper bound on the spectral gap of G^Λ, not the actual spectral gap of G^Λ.

Suppose that $X\sim\pi$ has finite second moments, so that (for each $v\in\mathbb{R}^k$) if the process $v'\mathbf{X}(t)$ is started from the stationary distribution, it admits a stationary autocorrelation function. A heuristic commonly used to tune MCMC algorithms in discrete time is the lag-1 autocorrelation of each component (X_i), with smaller lag-1 autocorrelation being better. The continuous time analogue of minimizing the discrete-time lag-1 autocorrelation is maximizing the (absolute value of the) slope of the autocorrelation function at lag 0. Rather than considering only each component projection, $\mathcal{F}_0 = \{x \to e_i' x : i \in \{1, ..., k\}\}$, we will consider their span, the subspace $\mathcal{F} = \left\{x \to v'x \ : \ v \in \mathbb{R}^k \setminus \{0\}\right\}$. Solving the optimization problem in Eq. (3.31) with $\mathcal{F}$ will provide a tighter surrogate optimization problem (since $\mathcal{F}_0 \subset \mathcal{F}$, it corresponds to optimizing a better bound on the spectral gap), and the solution will be covariant to linear transformations of $\mathbf{X}(t)$.

Theorem 3.6 ($\Lambda = \mathrm{Var}_{X\sim\pi}(X)$ is optimal in terms of lag-0 rate of decay of autocorrelations of linear functions of $\mathbf{X}$)**.** *Suppose that π admits second moments. Let $\Gamma = \underset{X\sim\pi}{\mathrm{Var}}(X)$. If $\mathbf{X}$ has generator G^Λ, then $\Lambda = \Gamma$ maximizes the worst case (over $f \in \mathcal{F} = \left\{x \to v'x \ : \ v \in \mathbb{R}^k \setminus \{0\}\right\}$) rate at which the autocorrelation of $f(\mathbf{X})$ decays in a neighbourhood of lag-0. Thus, in terms of short-run autocorrelations of linear functions of $\mathbf{X}$, the optimal shaping matrix for RWM proposals is the covariance matrix of the target distribution.*

Proof. From Lemma 3.3, for $v \in \mathbb{R} \setminus \{0\}$ and for $t > 0$:

$$\frac{d}{dt}\mathrm{Corr}(v'\mathbf{X}(0), v'\mathbf{X}(0))\big|_{t=0^+} = -\frac{v'\frac{k\Lambda}{\Lambda:\Sigma}v}{v'\Gamma v} \tag{3.32}$$

Hence, we need to solve:

$$\max_{\Lambda\succ 0}\ \min_{v\in\mathbb{R}^k\setminus\{0\}} \frac{v'\frac{k\Lambda}{\Lambda:\Sigma}v}{v'\Gamma v} \tag{3.33}$$

Substituting $w = \Gamma^{1/2}v$, we can instead solve:

$$\max_{\Lambda\succ 0}\ \min_{w\in\mathbb{R}^k\setminus\{0\}} \frac{w'\Gamma^{-1/2}k\Lambda\Gamma^{-1/2}w}{(\Lambda:\Sigma)(w'w)} \equiv \max_{\Lambda\succ 0} \frac{\lambda_1\left(\Gamma^{-1/2}k\Lambda\Gamma^{-1/2}\right)}{\Lambda:\Sigma} \tag{3.34}$$

where $\lambda_i(A)$ returns the ith smallest eigenvalue of A. Equivalently, we can solve:

$$\min_{\Lambda}\left(\lambda_k(\Gamma^{1/2}\Lambda^{-1}\Gamma^{1/2})(\Lambda:\Sigma)\right) \tag{3.35}$$

Substituting $\Theta = \Gamma^{1/2}\Lambda^{-1}\Gamma^{1/2}$, we can solve instead:

$$\min_{\Theta}\left(\lambda_k(\Theta)(\Theta^{-1} : (\Gamma^{1/2}\Sigma\Gamma^{1/2}))\right) \tag{3.36}$$

We will solve this optimization problem by lower bounding the objective function. It will be obvious that $\Theta = I$ achieves the lower bound, and so we will have $\Lambda = \Gamma$ is optimal. The lower bound is given by:

$$\begin{aligned}\left(\lambda_k(\Theta)(\Theta^{-1} : (\Gamma^{1/2}\Sigma\Gamma^{1/2}))\right) &\geq \lambda_k(\Theta)\lambda_1(\Theta^{-1})\mathrm{Tr}(\Gamma^{1/2}\Sigma\Gamma^{1/2}) \\ &= \mathrm{Tr}(\Gamma^{1/2}\Sigma\Gamma^{1/2})\end{aligned} \tag{3.37}$$

$\square$

This result shows that the rate of convergence of RWM is fundamentally limited by $\frac{k}{\Gamma:\Sigma}$ in the sense that, no matter the choice of proposal shaping matrix, the spectral gap of the generator will be bounded by $\frac{k}{\Gamma:\Sigma}$. When $\Gamma^{-1} = \Sigma$ this leads to the same speed limit as is witnessed by the $\pi \equiv \mathrm{N}$ case. When Γ^{-1} and Σ are very different, this would demonstrate that RWM will be very inefficient no matter how it is tuned.

This can also be used to say, for example, that the rate of convergence of the limiting diffusion, when the proposals are spherical (Λ= I), cannot be faster that $\frac{k\lambda_1(\Gamma^{-1})}{\mathrm{Tr}(\Sigma)}$. On the other hand, using proposals of $\Lambda = \Gamma$, the convergence rate could plausibly be as fast as $\frac{k}{\Sigma:\Gamma}$. Thus, we are somewhat justified to be more optimistic regarding the performance of the shaping $\Lambda = \Gamma$ than in that of the shaping $\Lambda = I$, but we cannot not provided any formal guarantee that the worst case rate of convergence for an arbitrary expected value is actually faster. The very short run performance, however has actually been optimized (by construction), justifying the intuition that optimizing the lag-1 autocorrelation is a “greedy”, suboptimal (but possibly reasonable) approximate solution to the optimal shaping problem.

One may also attempt to address the autocorrelations for non-linear functions. A particular function of interest is the log-density, $\log \pi(\mathbf{X})$, which is the only example we will consider here. Corollary 3.2, which follows directly from Lemma 3.3, shows us that for all choices of Λ (when the optimal scaling is used) $\log \pi$ has the same rate of decay of the autocorrelation at lag-0. This may be interpreted as saying that the speed of a RWM algorithm is fundamentally limited by the variance of the log-density, uniformly over all possible proposal shaping matrices. That is to say, when $k^{-1} \underset{X\sim\pi}{\mathrm{Var}}[\log \pi(X)]$ is very large, then RWM will be inefficient no matter how it is tuned.

Corollary 3.2. *If $\mathbf{X}(t)$ has generator G^{Λ} then the rate of change of the autocorrelation of $\log \pi(\mathbf{X}(t))$ at lag-0 is $\frac{k}{\underset{X\sim\pi}{\mathrm{Var}}[\log \pi(X)]}$ uniformly in Λ.*

Proof. This follows directly from Lemma 3.3 applied to $\log \pi(X)$. □

3.2.6 High Dimensional Dependence Asymptotics

We now consider the implications of the speed limits derived above on the performance decay for targets with high-dimensional dependence. This section is intentionally less mathematically rigorous than the rest of the work, with the intention of motivating future research in the area of optimal scaling for high-dimensional targets with non-trivial dependence structures.

We attempt to use the two "speed limits" derived in the previous section to characterize some regimes in which RWM will perform poorly. Consider a sequence of densities of varying dimension $\{\pi_k\}_{k\in\mathbb{N}}$ with π_k a density on $\mathbb{R}^k$. We consider $\Pi_{k,d} = \pi_k^{\otimes d}$.

Having chosen to scale time by dk rather than by d the acceleration required to get the weak limit in Theorem 3.1 is comparable across targets with equal total dimension, and hence the limiting diffusions should be comparable as well (at least in terms of their spectral gaps and rates of convergence).

If $\underset{X\sim\pi_k}{\mathrm{Var}} \log \pi_k(X) \notin O(k)$ then RWM performance drops off as $k \to \infty$, so dependence structures for which lead to this behaviour are expected to work poorly using RWM (much worse than for an IID target, with a spectral gap tending to 0), no matter how they are tuned. In the case of IID targets (where $\pi_k = \pi_1^{\otimes k}$) we have $\underset{X\sim\pi_k}{\mathrm{Var}} \log \pi_k(X) = k \underset{X\sim\pi_1}{\mathrm{Var}} \log \pi_1(X) \in \Theta(k)$. A similar result will hold for rotations and scalings of IID targets), showing that properly shaped proposals will yield commensurate performance for such targets.

The same story holds true if $\Gamma : \Sigma = \underset{X\sim\pi}{\mathrm{Var}}[X] : \underset{X\sim\pi}{\mathrm{Var}}[\nabla \log \pi] \notin O(k)$. Again, in the case of IID targets (where $\pi_k = \pi_1^{\otimes k}$, $\Gamma = \mathrm{diag}(\gamma)$, $\Sigma = \mathrm{diag}(\sigma)$) we have $\Gamma : \Sigma = k\gamma\sigma \in \Theta(k)$. A similar result will hold for rotations and scalings of IID targets), showing that properly shaped proposals will yield commensurate performance for such targets.

While these criteria may not be the sharpest possible, the recipe of (i) deriving a diffusion limit, (ii) deriving upper bounds on the spectral gap using formulas such as Eq. (3.31), and (iii) considering the asymptotics of the upper bound on the spectral gap as the dependence structure tends to infinity can be useful in developing our understanding of the behaviour of MCMC methods for dependent targets (which is an under explored topic in the literature).

3.3 Consequences of Assumption 3.1

Proposition 3.1 (Summary of consequences of Assumption 3.1 on π)**.** *The assumption that $\nabla \log \pi$ is L-Lipschitz continuous implies all of the following:*

(a) $\nabla \log \pi$ is differentiable (Lebesgue-)almost everywhere, and $\|\nabla^2 \log \pi\| \leq L$ where it exist (by Rademacher's theorem, see Federer [34]).

(b) The tails of π are at least as heavy of those of a Gaussian distribution. In fact it can be bounded below by a tangent Gaussian curve with variance-covariance matrix $\frac{1}{L}I$ at each point. (Lemma 2.5). This further implies that $\mathrm{Support}(\pi) = \mathbb{R}^k$*.*

(c) π is uniformly bounded above (Lemma 2.6).

(d) π is Lipschitz (Lemma 2.7).

(e) π has a broadly applicable integration by parts formula (Corollary 2.2): For any $f : \mathbb{R}^k \to \mathbb{R}$ which is locally Lipschitz, with $\nabla f(X)$ and $f(x)\nabla \log \pi(x)$ integrable (w.r.t. $\pi(x)dx$) we have

$$\mathop{\mathbb{E}}_{X\sim\pi} f(X)\nabla \log \pi(X) = -\mathop{\mathbb{E}}_{X\sim\pi} \nabla f(X) \ . \tag{3.38}$$

Similar formulas also hold for Jacobians of locally Lipschitz functions $f : \mathbb{R}^k \to \mathbb{R}^m$, and for divergences of locally Lipschitz functions $f : \mathbb{R}^k \to \mathbb{R}^k$.

(f) The following identities hold (Lemma 2.8):

$$\mathop{\mathbb{E}}_{X\sim\pi}[\nabla \log \pi(X)] = 0 \quad \textit{and} \quad \mathop{\mathrm{Var}}_{X\sim\pi}[\nabla \log \pi(X)] = -\mathop{\mathbb{E}}_{X\sim\pi}[\nabla^2 \log \pi(X)] \ . \tag{3.39}$$

(g) If $X \sim \pi$ then $\nabla \log \pi(X)$ is sub-Gaussian with proxy-variance L (Theorem 2.2). Hence all polynomial moments of $\nabla \log \pi$ and $\nabla^2 \log \pi$ are finite (Remark 2.3).

3.4 Proof of Theorem 3.1

This section utilizes well-established results on infinitessimal generators and Markov semigroups in order to prove our weak convergence result. For some introductory details to these topics see Section 4.6.3. Further suggested reading includes Ethier and Kurtz [33] and Kallenberg [53].

3.4.1 Definitions

We will make consistent use of the results listed in Proposition 3.1 which hold under our assumptions. Let $\hat{G}_d^{l,\Lambda}$ be the generator for a pure jump process with homogeneous jump intensity kd, and jump distribution given by the Random Walk Metropolis transition kernel with normal increments of mean 0 and variance $l^2 I_d \otimes \Lambda/(d-1)$, where Λ is symmetric and strictly positive definite. The generator is explicitly given by

$$\hat{G}_d^{l,\Lambda} f(x) = kd\, \mathbb{E}_{Z\sim \mathrm{N}_d(l^2\Lambda)} \left[(f(x+Z) - f(x)) \left(1 \wedge \frac{\Pi_d(x+Z)}{\Pi_d(x)} \right) \right] \ . \tag{3.40}$$

Then $\hat{G}_d^{l,\Lambda}$ is a bounded linear operator on $\hat{C}(\mathbb{R}^{kd})$, so we can take its domain to be the Banach space $\hat{C}(\mathbb{R}^{kd})$. Let $G_d^{l,\Lambda}$ be the restriction of $\hat{G}_d^{l,\Lambda}$ to functions of the form $f(x_{1:kd}) = f_1(x_{1:rk})$ which act only on the first rk components.

Let $G^{l,\Lambda}$ be the generator of an anisotropic Langevin diffusion,

$$G^{l,\Lambda} f = kl^2 a(l) \frac{1}{2} \Big([I_r \otimes \Lambda] : \nabla^2 f + (\nabla \log \pi^{\otimes r})' [I_r \otimes \Lambda] (\nabla f) \Big) \tag{3.41}$$

where $A : B = \mathrm{Tr}(A'B)$, and where $a(l) = 2\Phi\left(\frac{-l\sqrt{J}}{2}\right)$, and

$$\begin{aligned}\Sigma &= \underset{X\sim\pi}{\mathrm{Var}}[\nabla \log \pi(X)] = \underset{X\sim\pi}{\mathbb{E}} \nabla \log \pi(X) \nabla \log \pi(X)' \\ J &= \underset{X\sim\pi}{\mathbb{E}} \left[[\nabla \log \pi(X)]' \Lambda [\nabla \log \pi(X)]\right] = \Lambda : \Sigma\end{aligned} \tag{3.42}$$

We take the domain of $G^{l,\Lambda}$ to be $\mathcal{D}(G^{l,\Lambda}) = C_c^\infty(\mathbb{R}^{rk}) \cap L_2(\mathbb{R}^{rk}, \pi)$.

3.4.2 A General Convergence Theorem

Our goal is to show that if $\mathbf{X}_d$ has generator $\hat{G}_d^{l,\Lambda}$ and $\mathbf{X}^r$ has generator $G^{l,\Lambda}$ then the stochastic process of the first rk components of $\mathbf{X}_d$, $\mathbf{X}_d^{(1:r)}$, converges weakly to $\mathbf{X}^r$ in the Skorohod topology: $\mathbf{X}_d^{(1:r)} \Rightarrow \mathbf{X}^r$. The following result, paraphrased and specialized from Ethier and Kurtz [33], establishes sufficient conditions for this convergence to hold.

Proposition 3.2 (Convergence Theorem from Ethier and Kurtz [33])**.** *Suppose that:*

(i) *$\mathbf{X}_d$ is a Markov process in E_d with cadlag sample paths and with single-valued full generator $\hat{G}_d$, and $\mathbf{X}_d^{(1:r)} = \rho_d(\mathbf{X}_d)$ where $\rho_d : E_d \to E$ is measurable.*

(ii) *G is single-valued and its closure generates a Feller semigroup on E corresponding to the Markov process $\mathbf{X}^r$.*

(iii) *The initial distribution of $\mathbf{X}_d^{(1:r)}$ converges weakly to the initial distribution of $\mathbf{X}^r$;*

$$\mathbf{X}_d^{(1:r)}(0) \rightsquigarrow \mathbf{X}^r(0)\ . \tag{3.43}$$

(iv) *$\overline{\mathcal{D}(G)}$ contains an algebra which strongly separates points,*

(v) *For each $f \in \mathcal{D}(G)$, and each $T > 0$, there is a sequence of functions $f_d \in \mathcal{D}(\hat{G}_d)$, and a sequence of sets $F_d \subset E_d$ such that $\sup_d \|f_d\| < \infty$, and:*

$$\lim_{d\to\infty} \mathbb{P}(\mathbf{X}_d \in F_d \quad \forall 0 \le t \le T) = 1 \tag{3.44}$$

$$\lim_{d\to\infty} \sup_{\mathbf{x}_d \in F_d} |f(\rho_d(\mathbf{x}_d)) - f_d(\mathbf{x}_d)| = 0 \tag{3.45}$$

$$\lim_{d\to\infty} \sup_{\mathbf{x}_d \in F_d} \left| [Gf](\rho_d(\mathbf{x}_d)) - [\hat{G}_d f_d](\mathbf{x}_d) \right| = 0 \tag{3.46}$$

Then $\mathbf{X}_d^{(1:r)}$ *converges weakly in the Skorohod topology to* $\mathbf{X}^r$*:* $\mathbf{X}_d^{(1:r)} \Rightarrow \mathbf{X}^r$*..*

Proof. This is a restatement of Ethier and Kurtz [33] Chapter 4, corollary 8.7. where we have simplified and specialised some of the stated assumptions. In particular, we use that (cadlag $\implies$ progressive), and we assume that all generators involved are single valued, and that $\mathbf{X}^r$ is Feller which implies its generator must generate a strongly continuous contraction semigroup. □

Remark 3.2. *Because* $\mathbf{X}^r$ *is assumed to be a Feller process, it has a cadlag modification. Thus without loss of generality,* $\mathbf{X}^r$ *may be assumed to be cadlag.* ◁

Thus, taking $E = \mathbb{R}^{rk}$, and $E_d = \mathbb{R}^{kd}$, and $\rho_d(x_{1:kd}) = x_{1:rk}$, and $\hat{G}_d = \hat{G}_d^{l,\Lambda}$ and $G = G^{l,\Lambda}$ as defined above, we need only verify the five premises of Proposition 3.2 in order to establish Theorem 3.1. The first four are relatively straight forward, while the fifth will be much more involved.

Lemma 3.4 (Verifying Premises (i)-(iv) of Proposition 3.2)**.** *Under the definitions above, and the assumption that* $\nabla \log \pi$ *is L-Lipschitz we have that premises (i)-(iv) of Proposition 3.2 hold.*

Proof. (i) Since $\hat{G}_d^{l,\Lambda}$ is a bounded linear operator it must be single valued, and since the domain is the Banach space $\hat{C}(\mathbb{R}^{rkd})$ it must be its own closure. Since it generates a pure jump Markov process (with homogeneous intensity and the RWM transition function), the sample paths of $\mathbf{X}_d$ must be cadlag.

(ii) Since $\nabla \log \pi$ is assumed to be Lischitz, then by Ethier and Kurtz [33] [chapter 8, theorem 2.5], the closure of

$$\left\{ (f, G^{l,\Lambda} f) : f \in C_c^\infty(\mathbb{R}^{rk}) \right\} \tag{3.47}$$

is single valued and generates a Feller semigroup on $\hat{C}(\mathbb{R}^{rk})$.

(iii) This is trivially satisfied because of the assumption that $\mathbf{X}(0) \sim \pi$ and $\mathbf{X}_d(0) \sim \Pi_d = \pi^{\otimes d}$.

(iv) In our case, $\hat{C}(\mathbb{R}^{rk}) \supseteq \overline{\mathcal{D}(G^{l,\Lambda})} \supseteq C_c^\infty(\mathbb{R}^{rk})$. We verify that the algebra $C_c^\infty(\mathbb{R}^{rk})$ strongly separates points. Fix $x \in \mathbb{R}^k$ and $\delta > 0$. Consider the location-scale bump function:

$$f_{x,\delta}(y) = \exp\left(-\frac{1}{1-\frac{\|y-x\|^2}{\delta^2}}\right)\mathbb{1}_{\|y-x\|<\delta} \ . \tag{3.48}$$

This function is in $C_c^\infty(\mathbb{R}^{rk})$, and for $\|x-y\| > \delta$ we have

$$|f(y) - f(x)| \geq 1/e \ . \tag{3.49}$$

□

Therefore, to prove Theorem 3.1 we need only verify premise (v) of Proposition 3.2. This is done in the next subsection.

3.4.3 Verifying Premise (v) of Proposition 3.2

This premise is more complicated to verify. We first construct the sequence of "large sets", $\{F_d\}_{d\in\mathbb{N}}$ and verify Eq. (3.44). Then, we verify "uniform convergence of generator evaluations on large sets", Eq. (3.46) for $f \in C_c^\infty$ which we have taken to be the domain of $G^{l,\Lambda}$, using $f_d = f \circ \rho_d$.

The structure of this section closely follows that of the weak convergence proof in Neal and Roberts [81]. In addition to proving a scaling limit for RWM in the multivariate setting, we make two additional notable changes to the structure of the proof relative to [81]. First, we control the size of $\left\|\nabla \log \pi^{\otimes r}(x^{(1:r)})\right\|$ on our "large set" by including $F_{d,4}$, which we need for our proof. It also seems necessary and missing from their proof, since they need to control a term of the form

$$\lim_{d\to\infty} \sup_{x\in F_d} \left| d\nabla \log \pi^{\otimes r}(x^{(1:r)})' \left(\mathop{\mathbb{E}}_{Z^{(1:r)}} \left[Z^{(1:r)} \left(h(x^{(1:r)} + Z^{(1:r)}) - h(x^{(1:r)}) \right) \right] - [I_r \otimes \Lambda] \nabla h(x^{(1:r)}) \right) \right| = 0 \tag{3.50}$$

while they appear to only show the equivalent of

$$\lim_{d\to\infty} \sup_{x\in F_d} \left| kd \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[Z^{(1:r)} \left(h(x^{(1:r)} + Z^{(1:r)}) - h(x^{(1:r)}) \right) \right] - kl^2 \Lambda \nabla h(x^{(1:r)}) \right|, = 0 \tag{3.51}$$

which is not sufficient , since $\nabla \log \pi$ may be unbounded and $\mathbb{E}_{Z^{(1:r)}} h(x^{(1:r)} + Z^{(1:r)})$ is not compactly supported even if h is. Secondly, [81] implicitly assumes that the 3rd order partial derivatives of π exist and are uniformly bounded. This is needed in their proof to control the 3rd order remainder of a 2nd order Taylor expansion. We circumvent this by including $F_{d,3}$ below in our "large set", allowing us to control the relevant error using the convergence of an integrated finite difference to the corresponding derivative. The use of dominated convergence to control the approximation error on this set was inspired by Lalancette [60], though we have modified the technique to also not require continuous second derivatives. Thirdly, we correct an apparent error in the proof of [81] from taking a second order Taylor expansion of $x \mapsto 1 \wedge \exp(x)$, which is not valid since it's first derivative is discontinuous. That the final result was correct regardless of the error is serendipitous, and perhaps reflects that the authors had insight into what the limit should be before beginning the derivation.

Remark 3.3. *Taking $f_d = f \circ \rho_d$, we have that Eq. (3.45) is trivially satisfied.* $\triangleleft$

3.4.3.1 Large Sets

Suppose that $d > r$. The behaviour on the initial segment is irrelevant for the limit.

Let

$$F_d = F_0^d \cap F_{1,d} \cap F_{2,d} \cap F_{3,d} \cap F_{4,d} \tag{3.52}$$

where:

$$\begin{aligned}
F_0 &= \left\{ x \in \mathbb{R}^k : \nabla \log \pi \text{ is differentiable at } x \right\}, \\
F_{1,d} &= \left\{ x \in \mathbb{R}^{kd} : \left| R_d(x^{(r+1:d)}) - J \right| < d^{-1/8} + \frac{J(r-1)}{(d-1)} \right\}, \\
F_{2,d} &= \left\{ x \in \mathbb{R}^{kd} : \left| S_d(x^{(r+1:d)}) - J \right| < d^{-1/8} + \frac{J(r-1)}{(d-1)} \right\}, \\
F_{3,d} &= \left\{ x \in \mathbb{R}^{kd} : U_d(x^{(r+1:d)}) \le \theta(d) + Ll^2 K_\Lambda \sqrt{\frac{\log d}{d}} \right\}, \\
F_{4,d} &= \left\{ x \in \mathbb{R}^{kd} : \left\| \nabla \log \pi^{\otimes r}(x^{(1:r)}) \right\| \le 2\sqrt{krL \log d} \right\} ,
\end{aligned} \tag{3.53}$$

and

$$
\begin{aligned}
R_d(x^{(r+1:d)}) &= \frac{1}{d-1}\sum_{i=r+1}^{d}(\nabla \log \pi(x^{(i)})'\Lambda(\nabla \log \pi(x^{(i)})) \\
S_d(x^{(r+1:d)}) &= \frac{-1}{d-1}\sum_{i=r+1}^{d}\Lambda : (\nabla^2 \log \pi(x^{(i)})) \\
U_d(x^{(r+1:d)}) &= \mathop{\mathbb{E}}_{Z\sim \mathrm{N}_d(l^2\Lambda)}\left|\sum_{i=r+1}^{d}\left(\log\left(\tfrac{\pi(x^{(i)}+Z^{(i)})}{\pi(x^{(i)})}\right)-Z^{(i)\prime}\nabla \log \pi(x^{(i)})-Z^{(i)\prime}\tfrac{\nabla^2 \log \pi(x^{(i)})}{2}Z^{(i)}\right)\right| \\
K_\Lambda &= \sqrt{2\left\|\Lambda\right\|_F^2+\mathrm{Tr}(\Lambda)^2} \\
\theta(d) &= l^2 K_\Lambda \mathop{\mathbb{E}}_{X\sim\pi}\sqrt{\mathop{\mathbb{E}}_{Z\sim \mathrm{N}\left(0,\frac{l^2\Lambda}{d-1}\right)}\frac{1}{\left\|Z\right\|^4}\left|\log\frac{\pi(X+Z)}{\pi(X)}-\nabla \log \pi(X)Z-Z'\frac{\nabla^2 \log \pi(X)}{2}Z\right|^2}\ .
\end{aligned}
\tag{3.54}
$$

Lemma 3.5. *Under assumption (A1)* $\lim_{d\to\infty}\theta(d)=0$

Proof. For $w\neq 0$, $z=w/\sqrt{d-1}$, and $x\in F_0$, using the fundamental theorem of calculus:

$$
\begin{aligned}
&\frac{1}{\left\|z\right\|^2}\left(\log\frac{\pi(x+z)}{\pi(x)}-z'\nabla \log \pi(x)-z'\frac{\nabla^2 \log \pi(x)}{2}z\right) \\
&=\frac{1}{\left\|w\right\|^2}\left(\sqrt{d-1}\int_0^1\left(\nabla \log \pi\left(x+\frac{hw}{\sqrt{d-1}}\right)-\nabla \log \pi(x)\right)w\ dh - w'\frac{\nabla^2 \log \pi(x)}{2}w\right) \\
&=\frac{1}{\left\|w\right\|^2}\left(\int_0^1\frac{\nabla \log \pi\left(x+\frac{hw}{\sqrt{d-1}}\right)-\nabla \log \pi(x)}{1/\sqrt{d-1}}w\ dh - w'\frac{\nabla^2 \log \pi(x)}{2}w\right).
\end{aligned}
\tag{3.55}
$$

As $d\to\infty$ the integrand converges pointwise to

$$
w'\nabla^2 \log \pi(x)wh \tag{3.56}
$$

from the differentiability of $\nabla \log \pi$ at x. Also, from the Lipschitz property of $\nabla \log \pi$, the integrand is bounded by $Lh\left\|w\right\|^2$ for all $d\geq 2$ and all $h\in[0,1]$. In fact, we have the following bound which will also be useful later:

$$\begin{aligned}
&\left| \int_0^1 \frac{\nabla \log \pi \left(x + \frac{hw}{\sqrt{d-1}} \right) - \nabla \log \pi(x)}{1/\sqrt{d-1}} w dh - w' \frac{\nabla^2 \log \pi(x)}{2} w \right| \\
&\quad \leq \left| \int_0^1 \frac{\nabla \log \pi \left(x + \frac{hw}{\sqrt{d-1}} \right) - \nabla \log \pi(x)}{1/\sqrt{d-1}} w dh \right| + \left| w' \frac{\nabla^2 \log \pi(x)}{2} w \right| \\
&\quad \leq \int_0^1 L \left\| w \right\|^2 h \; dh + \frac{L}{2} \left\| w \right\|^2 \\
&\quad = L \left\| w \right\|^2
\end{aligned} \tag{3.57}$$

Therefore, by the bounded convergence theorem, for $w \neq 0$ and $x \in F_0$

$$\begin{aligned}
&\lim_{d\to\infty} \int_0^1 \frac{\nabla \log \pi \left(x + \frac{hw}{\sqrt{d-1}} \right) - \nabla \log \pi(x)}{1/\sqrt{d-1}} w dh \\
&\quad = \int_0^1 {}' \nabla^2 \log \pi(x) w h \; dh \\
&\quad = w' \frac{\nabla^2 \log \pi(x)}{2} w.
\end{aligned} \tag{3.58}$$

Therefore, for $x \in F_0$, and $w \neq 0$,

$$\lim_{d\to\infty} \frac{d-1}{\left\| w \right\|^2} \left(\log \frac{\pi(x + \frac{w}{\sqrt{d-1}})}{\pi(x)} - \frac{w'}{\sqrt{d-1}} \nabla \log \pi(x) - \frac{w}{\sqrt{d-1}}' \frac{\nabla^2 \log \pi(x)}{2} \frac{w}{\sqrt{d-1}} \right) = 0 \tag{3.59}$$

Now, using the bound in Eq. (3.57) again to upper bound the integrand by L^2, by the dominated convergence theorem, we have (with $W \sim \mathrm{N}(0, l^2 \Lambda)$):

$$\lim_{d\to\infty} \mathbb{E} \frac{(d-1)^2}{\left\| W \right\|^4} \left| \left(\log \frac{\pi(x + \frac{W}{\sqrt{d-1}})}{\pi(x)} - \frac{W' \nabla \log \pi(x)}{\sqrt{d-1}} - \frac{W}{\sqrt{d-1}}' \frac{\nabla^2 \log \pi(x)}{2} \frac{W}{\sqrt{d-1}} \right) \right|^2 = 0 \; . \tag{3.60}$$

As a function of x, this is uniformly bounded by L^2 on F_0, so applying the dominated convergence theorem a third time, since F_0 has measure 1 under π (since π is absolutely continuous with respect to the Lebesgue measure) we have:

$$\theta(d) = l^2 K_\Lambda \mathop{\mathbb{E}}_{X\sim\pi} \sqrt{\mathop{\mathbb{E}}_{Z\sim \mathrm{N}\left(0,\frac{l^2\Lambda}{d-1}\right)} \frac{1}{\|Z\|^4} \left| \log \frac{\pi(X+Z)}{\pi(X)} - Z'\nabla \log \pi(X) - Z' \frac{\nabla^2 \log \pi(X)}{2} Z \right|^2} \\ \to 0 \tag{3.61}$$

□

Lemma 3.6. *If $\mathbf{X}_d$ is the pure jump process with generator $G_d^{l,\Lambda}$, and if $\mathbf{X}_d(0) \sim \Pi_d$, then for any $T > 0$*

$$\mathbb{P}(\mathbf{X}_d(t) \in F_d \quad \forall 0 \le t \le T) \to 1 \tag{3.62}$$

Proof. Since the number of possible jumps times of the process $\mathbf{X}$ in the interval $[0, T]$, N_T, is distributed as $N_T \sim \text{Poisson}(Tkd)$, we have

$$\begin{aligned} &\mathbb{P}(\text{not } \ (\mathbf{X}_d(t) \in F_d \quad \forall 0 \le t \le T)) \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\exists t\in[0,T]:\ \mathbf{X}_d(t)\notin F_d} | N_T\right]\right] \\ &\le \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}_{X_0^{(1:d)}\notin F_d} + \sum_{\substack{t>0 \\ N_t\neq N_{t-}}} \mathbf{1}_{\mathbf{X}_d(t)\notin F_d} \quad \Big| N_T\Big]\Big] \\ &= \mathbb{E}\left[(N_T+1) \mathop{\mathbb{P}}_{X_d\sim\Pi_d}(X_d \notin F_d)\right] \\ &= (Tkd+1) \mathop{\mathbb{P}}_{X_d\sim\Pi_d}(X_d \notin F_d) \end{aligned} \tag{3.63}$$

Thus it is sufficient to show that $\mathop{\mathbb{P}}_{X_d\sim\Pi_d}(X_d \notin F_d) \in o(1/d)$. Applying subadditivity:

$$\begin{aligned} \mathop{\mathbb{P}}_{X_d\sim\Pi_d}(X_d \notin F_d) \le & \mathop{\mathbb{P}}_{X_d\sim\Pi_d}\left(\left|R_d(X_d^{(r+1:d)}) - J\right| > d^{-1/8} + \frac{J(r-1)}{(d-1)}\right) \\ &+ \mathop{\mathbb{P}}_{X_d\sim\Pi_d}\left(\left|S_d(X^{(r+1:d)}) - J\right| > d^{-1/8} + \frac{J(r-1)}{(d-1)}\right) \\ &+ \mathop{\mathbb{P}}_{X_d\sim\Pi_d}\left(U_d(X^{(r+1:d)}) > \theta(d) + Ll^2 K_\Lambda \sqrt{\frac{\log d}{d}}\right) \\ &+ \mathop{\mathbb{P}}_{X_d\sim\Pi_d}\left(\left\|\nabla \log \pi^{\otimes r}(X^{(1:r)})\right\| > 2\sqrt{Lkr \log d}\right) , \end{aligned} \tag{3.64}$$

so it is sufficient to show that each of these four terms is individually $o(1/d)$.

It is obvious that $\mathbb{E}_{X_d\sim\Pi_d} R_d(X_d^{(r+1:d)}) = J\frac{d-r}{d-1}$. From distributional integration by parts (Theorem 2.1) we also have $\mathbb{E}_{X\sim\Pi_d} S_d(X^{(r+1:d)}) = J\frac{d-r}{d-1}$.

For the first term, since $\nabla \log \pi(Y)$ is subgaussian for $Y \sim \pi$ (see Theorem 2.2), and hence has all of its polynomial moments, applying Markov's inequality gives:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{X\sim\Pi_d}\left(\left|R_d(X^{(r+1:d)}) - J\right| > d^{-1/8} + \frac{J(r-1)}{(d-1)}\right) \\
&\quad\leq \mathop{\mathbb{E}}_{X\sim\Pi_d}\left(\left|R_d(X^{(r+1:d)}) - \mathbb{E}_{X_d\sim\Pi_d} R_d(X_d^{(r+1:d)})\right|^4\right) d^{1/2} \\
&\quad= d^{1/2}\frac{\mathop{\mathbb{E}}_{Y\sim\pi}\left(A(Y)^4\right) + 3(d-2)\mathop{\mathbb{E}}_{Y\sim\pi}\left(A(Y)^2\right)^2}{(d-1)^3} \\
&\quad\leq \frac{6}{(d-1)^{3/2}}\|\Lambda\|^4 M_1
\end{aligned} \tag{3.65}$$

for $M_1 < \infty$ sufficiently large, where $A(Y) = (\nabla \log \pi(Y)'\Lambda(\nabla \log \pi(Y)) - J$.

For the second term, again since $\nabla \log \pi(Y)$ has all of its polynomial moments for $Y \sim \pi$:

$$\begin{aligned}
&\mathop{\mathbb{P}}_{X\sim\Pi_d}\left(\left|S_d(X^{(r+1:d)}) - J\right| > d^{-1/8} + \frac{J(r-1)}{(d-1)}\right) \\
&\quad\leq \mathop{\mathbb{E}}_{X\sim\Pi_d}\left(\left|S_d(X^{(r+1:d)}) - \mathbb{E}_{X\sim\Pi_d} S_d(X^{(r+1:d)})\right|^4\right) d^{1/2} \\
&\quad= d^{1/2}\frac{\mathop{\mathbb{E}}_{Y\sim\pi}\left(B(Y)^4\right) + 3(d-2)\mathop{\mathbb{E}}_{Y\sim\pi}\left(B(Y)^2\right)^2}{(d-1)^3} \\
&\quad\leq \frac{6}{(d-1)^{3/2}}\|\Lambda\|_F^4 M_2
\end{aligned} \tag{3.66}$$

for $M_2 < \infty$ sufficiently large, where $B(Y) = \Lambda : (\nabla^2 \log \pi(Y)) - J$.

For the third term (letting $W \sim \mathrm{N}\left(0, l^2\Lambda\right)$):

$$
\begin{aligned}
&U_d(x^{(r+1:d)}) \\
&= \mathop{\mathbb{E}}_{Z\sim \mathrm{N}_d(l^2\Lambda)} \left| \sum_{i=r+1}^{d} \left(\log \tfrac{\pi(x^{(i)}+Z^{(i)})}{\pi(x^{(i)})} - \nabla \log \pi(x^{(i)})' Z^{(i)} - Z^{(i)\prime} \tfrac{\nabla^2 \log \pi(x^{(i)})}{2} Z^{(i)} \right) \right| \\
&\leq \mathop{\mathbb{E}}_{Z\sim \mathrm{N}_d(l^2\Lambda)} \sum_{i=r+1}^{d} \left| \left(\log \tfrac{\pi(x^{(i)}+Z^{(i)})}{\pi(x^{(i)})} - \nabla \log \pi(x^{(i)})' Z^{(i)} - Z^{(i)\prime} \tfrac{\nabla^2 \log \pi(x^{(i)})}{2} Z^{(i)} \right) \right| \\
&= \frac{1}{d-1} \sum_{i=r+1}^{d} \mathbb{E}_W \left| \int_0^1 \tfrac{\nabla \log \pi\left(x^{(i)} + \frac{hW}{\sqrt{d-1}}\right) - \nabla \log \pi(x^{(i)})}{1/\sqrt{d-1}} W dh - W' \tfrac{\nabla^2 \log \pi(x^{(i)})}{2} W \right| \\
&\leq \frac{\left(\mathbb{E}_W \|W\|^4\right)^{\frac{1}{2}}}{d-1} \sum_{i=r+1}^{d} \left(\mathbb{E}_W \frac{1}{\|W\|^4} \left| \int_0^1 \tfrac{\nabla \log \pi\left(x^{(i)} + \frac{hW}{\sqrt{d-1}}\right) - \nabla \log \pi(x^{(i)})}{1/\sqrt{d-1}} W dh - W' \tfrac{\nabla^2 \log \pi(x^{(i)})}{2} W \right|^2 \right)^{\frac{1}{2}}
\end{aligned}
\tag{3.67}
$$

Using Isserlis' theorem (Isserlis [46], equation (39)* therein)

$$
\begin{aligned}
\mathop{\mathbb{E}}_{W\sim \mathrm{N}(0,l^2\Lambda)} \|W\|^4 &= \mathop{\mathbb{E}}_{W\sim \mathrm{N}(0,l^2\Lambda)} \left(\sum_{i=1}^{k} W_i^2 \right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} \mathop{\mathbb{E}}_{W\sim \mathrm{N}(0,l^2\Lambda)} [W_i^2 W_j^2] \\
&= \sum_{i=1}^{k} \sum_{j=1}^{k} l^4 (\Lambda_{ii}\Lambda_{jj} + 2\Lambda_{ij}^2) = l^4 \left(2 \|\Lambda\|_F^2 + \mathrm{Tr}(\Lambda)^2 \right)
\end{aligned}
\tag{3.68}
$$

Thus:

$$
\begin{aligned}
&U_d(x^{(r+1:d)}) \leq V_d(x^{(r+1:d)}) \\
&\quad := \frac{l^2 K_\Lambda}{(d-1)} \sum_{i=r+1}^{d} \left(\mathbb{E} \left| \frac{1}{\|W\|^2} \Big[\int_0^1 \frac{\nabla \log \pi\left(x^{(i)} + \frac{hW}{\sqrt{d-1}}\right) - \nabla \log \pi(x^{(i)})}{1/\sqrt{d-1}} W dh - W' \frac{\nabla^2 \log \pi(x^{(i)})}{2} W \Big] \right|^2 \right)^{\frac{1}{2}}
\end{aligned}
$$

and so:

$$
\mathbb{E}_{X\sim \Pi_d} V_d(X) = \theta(d) \frac{d-r}{d-1} \leq \theta(d) \tag{3.69}
$$

Moreover, from Eq. (3.57), for $x \in F_0$:

$$
\left(\mathbb{E} \left| \frac{1}{\|W\|^2} \Big[\int_0^1 \frac{\nabla \log \pi\left(x^{(i)} + \frac{hW}{\sqrt{d-1}}\right) - \nabla \log \pi(x^{(i)})}{1/\sqrt{d-1}} W dh - W' \frac{\nabla^2 \log \pi(x^{(i)})}{2} W \Big] \right|^2 \right)^{\frac{1}{2}} \leq L \tag{3.70}
$$

Thus, by Hoeffding's Inequality (Boucheron et al. [15], theorem 2.8 therein):

$$
\begin{aligned}
&\mathbb{P}_{X_d\sim\Pi_d}\left(U_d(X_d^{(r+1:d)}) > \theta(d) + l^2 L K_\Lambda \sqrt{\frac{\log d}{d}}\right)\\
&\quad\le \mathbb{P}_{X_d\sim\Pi_d}\left(V_d(X_d^{(r+1:d)}) > \theta(d) + l^2 L K_\Lambda \sqrt{\frac{\log d}{d}}\right)\\
&\quad\le \frac{1}{d^2}
\end{aligned}
\tag{3.71}
$$

For the fourth (and last) term, since, $\nabla \log \pi(X)$ is subgaussian with proxy variance L for $X \sim \pi$ (see Theorem 2.2), then

$$
\begin{aligned}
\mathbb{E}_{X\sim\pi}[\exp(s\,\|\nabla \log \pi(X)\|)] &\le \mathbb{E}_{X\sim\pi}[\exp(s\sqrt{k}\max_{j\le k}|\partial_j \log \pi(X)|)]\\
&= \mathbb{E}_{X\sim\pi}[\max_{j\le k}\exp(s\sqrt{k}\,|\partial_j \log \pi(X)|)]\\
&\le \mathbb{E}_{X\sim\pi}[\sum_{j\le k}\exp(s\sqrt{k}\partial_j \log \pi(X)) + \exp(-s\sqrt{k}\partial_j \log \pi(X))]\\
&\le 2k\exp(s^2 kL/2),
\end{aligned}
\tag{3.72}
$$

so

$$
\begin{aligned}
\mathbb{P}_{X\sim\pi}(\|\nabla \log \pi(X)\| > t) &\le \inf_{s>0} e^{-st}\,\mathbb{E}_{X\sim\pi}[\exp(s\,\|\nabla \log \pi(X)\|)]\\
&\le 2k \inf_{s>0} e^{-st+s^2kL/2}\\
&= 2ke^{-\frac{t^2}{2kL}}.
\end{aligned}
\tag{3.73}
$$

Now, for $\left(\left\|\nabla \log \pi^{\otimes r}(X_d^{(1:r)})\right\| > 2\sqrt{rkL\log d}\right)$ to occur, at least one block, indexed by $j \in \{1, ..., r\}$, must have $\left(\left\|\nabla \log \pi(X_d^{(j)})\right\| > 2\sqrt{kL\log d}\right)$. Thus,

$$
\mathbb{P}_{X_d\sim\Pi_d}\left(\left\|\nabla \log \pi^{\otimes r}(X_d^{(1:r)})\right\| > 2\sqrt{rkL\log d}\right) \le \frac{2kr}{d^2}
\tag{3.74}
$$

Thus:

$$
1 - \mathbb{P}(\mathbf{X}_d(t) \in F_d \quad \forall 0 \le t \le T) \le (Tkd+1)\,\mathbb{P}_{X_d\sim\Pi_d}(X_d \notin F_d) \to 0
\tag{3.75}
$$

$\square$

3.4.3.2 Uniform Convergence of Generator Evaluations on Large Sets

For each $h \in C(\mathbb{R}^k)$ let $h_d = h \circ \rho_d$. For the remainder of this section, $Z \sim \mathrm{N}\left(0, \frac{l^2}{(d-1)} I_d \otimes \Lambda\right)$ unless stated otherwise, and $Z^{(1)} \sim \mathrm{N}_d\left(l^2\Lambda\right)$ is the first k component block of Z.

We introduce an intermediate object, $\tilde{G}_d^{l,\Lambda}$ on $\{h \circ \rho_d : h \in C_c^\infty\}$, which resembles, but is not, a generator for a diffusion process. Take $\tilde{G}_d^{l,\Lambda}$ given by:

$$\begin{aligned}
&\tilde{G}_d^{l,\Lambda} h_d(x) \\
&= \frac{kl^2}{2}\mathbb{E}_Z[1 \wedge e^{B_d(x,Z^{(r+1:d)})}]\ [I_r \otimes \Lambda] : \nabla^2 h(x^{(1:r)}) \\
&\qquad + kl^2\mathbb{E}_Z[1 \wedge e^{C_d(x,Z)}; C_d(x,Z) < 0]\ (\nabla \log \pi^{\otimes r}(x^{(1:r)}))'[I_r \otimes \Lambda](\nabla h(x^{(1:r)}))
\end{aligned} \tag{3.76}$$

where

$$\begin{aligned}
\epsilon(x,z) &= \log\frac{\pi(x+z)}{\pi(x)}\ , \\
B_d(x, z^{(r+1:d)}) &= \sum_{i=r+1}^{d} \epsilon(x^{(i)}, z^{(i)})\ , \\
\mathcal{E}(x, z^{(1:r)}) &= \sum_{j=1}^{r} \epsilon(x^{(j)}, z^{(j)})\ , \\
C_d(x,z) &= \mathcal{E}(x, z^{(1:r)}) + B_d(x, z^{(r+1:d)})\ .
\end{aligned} \tag{3.77}$$

We will show that for any $h \in C_c^\infty(\mathbb{R}^{rk})$ we have both:

$$\lim_{d\to\infty}\sup_{x\in F_d}\left|\hat{G}_d^{l,\Lambda} h_d(x) - \tilde{G}_d^{l,\Lambda} h_d(x)\right| = 0 \tag{3.78}$$

which is verified in Lemma 3.7, and

$$\lim_{d\to\infty}\sup_{x\in F_d}\left|\tilde{G}_d^{l,\Lambda} h_d(x) - G^{l,\Lambda} h(x)\right| = 0 \tag{3.79}$$

which is verified through Lemma 3.8.

Then, since $G^{l,\Lambda} h(x^{(1:r)}) = [G^{l,\Lambda} h] \circ \rho_d(x)$, we will have verified Eq. (3.46).

Lemma 3.7 ($\tilde{G}_d^{l,\Lambda}$ is close to $\hat{G}_d^{l,\Lambda}$)**.**

$$\lim_{d\to\infty} \sup_{x\in F_d} \left|\hat{G}_d^{l,\Lambda} h_d(x) - \tilde{G}_d^{l,\Lambda} h_d(x)\right| = 0 \tag{3.80}$$

Proof. We will make use of the following shorthands in the proof: $B_d = B_d(x, Z^{(r+1:d)})$, $\varepsilon = \mathcal{E}(z^{(1:r)}, x)$, $\mathcal{E} = \mathcal{E}(Z^{(1:r)}, x)$, $c_d = \varepsilon + B_d$, and $C_d = \mathcal{E} + B_d$.

Recall that:

$$\hat{G}_d^{l,\Lambda} h_d(x) = kd \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\left(h(x^{(1:r)} + Z^{(1:r)}) - h(x^{(1:r)}) \right) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(1 \wedge \prod_{i=1}^{d} \frac{\pi(x^{(i)} + Z^{(i)})}{\pi(x^{(i)})} \right) \right] . \tag{3.81}$$

Let

$$\begin{aligned} \gamma(x, z) &= \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(1 \wedge e^{\mathcal{E}(x,z) + B_d(x, Z^{(r+1:d)})} \right) \\ &= \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(1 \wedge e^{c_d} \right) \ . \end{aligned} \tag{3.82}$$

We can compute the first derivative of γ by differentiating under the integral, since the integrand in question is weakly differentiable. We cannot compute the second derivative in the same way, since the integrand is not twice differentiable. In contrast to [81], we will avoid needing to take a second derivative of γ in the proof, which will allow us to circumvent this issue. The first derivative is given by:

$$\begin{aligned} &\nabla_z \gamma(x, z^{(1:r)}) \\ &= \nabla \log \pi^{\otimes r}(x^{(1:r)} + z^{(1:r)}) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(e^{\mathcal{E}(x, z^{(1:r)}) + B_d(x, Z^{(r+1:d)})}; \mathcal{E}(x, z^{(1:r)}) + B_d(x, Z^{(r+1:d)}) < 0 \right) \\ &= \nabla \log \pi^{\otimes r}(x^{(1:r)} + z^{(1:r)}) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(e^{c_d}; c_d < 0 \right) \ . \end{aligned} \tag{3.83}$$

Now, since $h \in C_c^\infty(\mathbb{R}^{rk})$, there is a $[0,1]$-valued random variable H with:

$$\begin{aligned}
&\hat{G}_d^{l,\Lambda} h_d(x) \\
&= kd \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\left(\nabla h(x^{(1:r)})' Z^{(1:r)} + \frac{\nabla^2 h(x^{(1:r)}) : [Z^{(1:r)}]^{\otimes 2}}{2} + \frac{\nabla^3 h(x^{(1:r)} + HZ^{(1:r)}) : [Z^{(1:r)}]^{\otimes 3}}{6} \right) \gamma(x, Z^{(1:r)}) \right] \\
&= kd \nabla h(x^{(1:r)})' \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[Z^{(1:r)\prime} \gamma(x, Z^{(1:r)}) \right] \\
&\quad + kd \frac{\nabla^2 h(x^{(1:r)})}{2} : \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[[Z^{(1:r)}]^{\otimes 2} \gamma(x, Z^{(1:r)}) \right] \\
&\quad + kd \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\frac{\nabla^3 h(x^{(1:r)} + HZ^{(1:r)}) : [Z^{(1:r)}]^{\otimes 3}}{6} \gamma(x, Z^{(1:r)}) \right]
\end{aligned} \tag{3.84}$$

First, using Stein's lemma,

$$\begin{aligned}
&kd \nabla h(x^{(1:r)})' \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[Z^{(1:r)\prime} \gamma(x, Z^{(1:r)}) \right] \\
&= kl^2 \frac{d}{d-1} \nabla h(x^{(1:r)})' \Lambda \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\nabla_z \gamma(x, Z^{(1:r)}) \right] \\
&= kl^2 \frac{d}{d-1} \nabla h(x^{(1:r)})' \Lambda \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\nabla \log \pi^{\otimes r}(x^{(1:r)} + z^{(1:r)}) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(e^{C_d}; C_d < 0 \right) \right]
\end{aligned}$$

Thus,

$$\begin{aligned}
&\sup_{x \in F_d} \left| kd \nabla h(x^{(1:r)})' \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[Z^{(1:r)\prime} \gamma(Z^{(1:r)}, x) \right] - kl^2 \nabla h(x^{(1:r)})' \Lambda \nabla \log \pi^{\otimes r}(x^{(1:r)}) \mathbb{E}_Z \left[e^{C_d}; C_d < 0 \right] \right| \\
&= \sup_{x \in F_d} \frac{kl^2 d}{d-1} \left| \nabla h(x^{(1:r)})' \Lambda \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\left(\nabla \log \pi^{\otimes r}(x^{(1:r)} + Z^{(1:r)}) - \nabla \log \pi^{\otimes r}(x^{(1:r)}) \right) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(e^{C_d}; C_d < 0 \right) \right] \right| \\
&\quad + \frac{kl^2}{d-1} \left| \nabla h(x^{(1:r)})' \Lambda \mathop{\mathbb{E}}_{Z^{(1:r)}} \left[\nabla \log \pi^{\otimes r}(x^{(1:r)}) \mathop{\mathbb{E}}_{Z^{(r+1:d)}} \left(e^{C_d}; C_d < 0 \right) \right] \right| \\
&\leq \sup_{x \in F_d} \frac{kl^2}{d-1} \left\| \nabla h(x^{(1:r)}) \right\| \|\Lambda\| \left(dL \mathop{\mathbb{E}}_{Z^{(1:r)}} [\|Z^{(1:r)}\|] + \|\nabla \log \pi^{\otimes r}(x^{(1:r)})\| \right) \\
&\leq \sup_{x \in F_d} \frac{kl^2}{d-1} \| \|\nabla h\| \|_\infty \|\Lambda\| \left(dL \sqrt{rl^2 \operatorname{Tr}(\Lambda)/(d-1)} + 2\sqrt{krL \log d} \right) \\
&\in O(d^{-1/2})
\end{aligned}$$

where, in the last inequality we used the bound on $\|\nabla \log \pi^{\otimes r}(x^{(1:r)})\|$ for $x \in F_{4,d}$.

Second, for $W \sim \mathrm{N}(0, l^2 I_r \otimes \Lambda)$

$$
\begin{aligned}
&\sup_{x\in F_d}\left|kd\frac{\nabla^2 h(x^{(1:r)})}{2} : \mathop{\mathbb{E}}_{Z^{(1:r)}}\left[[Z^{(1:r)}]^{\otimes 2}\gamma(Z^{(1:r)},x)\right] - \frac{kl^2}{2}\mathbb{E}_Z[1\wedge e^{B_d}]\ [I_r\otimes\Lambda] : \nabla^2 h(x^{(1:r)})\right| \\
&\le \sup_{x\in F_d}\left|kd\frac{\nabla^2 h(x^{(1:r)})}{2} : \mathop{\mathbb{E}}_{Z^{(1:r)}}\left[[Z^{(1:r)}]^{\otimes 2}\left(\gamma(Z^{(1:r)},x) - \mathop{\mathbb{E}}_{Z^{(r+1:d)}}[1\wedge e^{B_d}]\right)\right]\right| \\
&\quad + \frac{k\nabla^2 h(x^{(1:r)})}{2(d-1)}\mathbb{E}_Z[1\wedge e^{B_d}] \\
&\le \sup_{x\in F_d}\left|kd\frac{\nabla^2 h(x^{(1:r)})}{2} : \mathbb{E}_Z\left[[Z^{(1:r)}]^{\otimes 2}\left(\mathop{\mathbb{E}}_{Z^{(r+1:d)}}[1\wedge e^{\mathcal{E}+B_d} - 1\wedge e^{B_d}]\right)\right]\right| \\
&\quad + \frac{k\left\|\left\|\nabla^2 h\right\|_F\right\|_\infty \|\Lambda\|_F}{2(d-1)}\mathbb{E}_Z[1\wedge e^{B_d(x,Z^{(r+1:d)})}] \\
&\le \sup_{x\in F_d} kd\frac{\left\|\left\|\nabla^2 h\right\|\right\|_\infty}{2}\mathbb{E}_Z\left[\left\|Z^{(1:r)}\right\|^2\left|\mathcal{E}(Z^{(1:r)},x)\right|\right] + \frac{k\left\|\left\|\nabla^2 h\right\|_F\right\|_\infty \|\Lambda\|_F}{2(d-1)} \\
&\le \sup_{x\in F_d} kd\frac{\left\|\left\|\nabla^2 h\right\|\right\|_\infty}{2}\mathbb{E}_Z\left[\left\|Z^{(1:r)}\right\|^2\left(\left\|\nabla\log\pi^{\otimes r}(x^{(1:r)})\right\|\left\|Z^{(1:r)}\right\| + \left\|Z^{(1:r)}\right\|^2 L/2\right)\right] \\
&\quad + \frac{k\left\|\left\|\nabla^2 h\right\|_F\right\|_\infty \|\Lambda\|_F}{2(d-1)} \\
&\le \sup_{x\in F_d} k\frac{\left\|\left\|\nabla^2 h\right\|\right\|_\infty}{2}\mathbb{E}\left[\frac{d\sqrt{\log d}}{(d-1)^{3/2}}\|W\|^3\, 2\sqrt{krL} + \frac{d}{(d-1)^2}\|W\|^4 L/2\right] \\
&\quad + \frac{k\left\|\left\|\nabla^2 h\right\|_F\right\|_\infty \|\Lambda\|_F}{2(d-1)} \\
&\in O(\sqrt{\log(d)/d})
\end{aligned}
\tag{3.85}
$$

Third, for $W \sim \mathrm{N}(0, l^2 I_r \otimes \Lambda)$,

$$
\begin{aligned}
&\sup_{x\in F_d}\left|kd\mathop{\mathbb{E}}_{Z^{(1:r)}}\left[\frac{\nabla^3 h(x^{(1:r)} + HZ^{(1:r)}) : [Z^{(1:r)}]^{\otimes 3}}{6}\gamma(Z^{(1:r)},x)\right]\right| \\
&\le \sup_{x\in F_d} kd\left\|\left\|\nabla^3 h\right\|\right\|_\infty \mathop{\mathbb{E}}_{Z^{(1:r)}}\left[\left\|Z^{(1:r)}\right\|^3\right] \\
&\le \sup_{x\in F_d}\frac{kd}{(d-1)^{3/2}}\left\|\left\|\nabla^3 h\right\|\right\|_\infty \mathbb{E}\left[\|W\|^3\right] \\
&\in O(d^{-1/2})
\end{aligned}
\tag{3.86}
$$

Thus:

$$
\sup_{x\in F_d}\left|\hat{G}_d h_d - \tilde{G}_d h_d\right| \in O(\sqrt{\log(d)/d}) \tag{3.87}
$$

□

Lemma 3.8 ($\tilde{G}_d^{l,\Lambda}$ is close to $G^{l,\Lambda}$)**.**

$$\lim_{d\to\infty}\sup_{x\in F_d}\left|2\Phi(-l\sqrt{I}/2)-\mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[1\wedge e^{B_d(x,Z^{(r+1:d)})}\right]\right|=0 \tag{3.88}$$

and:

$$\lim_{d\to\infty}\sup_{x\in F_d}\left|\Phi(-l\sqrt{I}/2)-\mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[e^{C_d(x,Z)};C_d(x,Z)<0\right]\right|=0 \tag{3.89}$$

and hence:

$$\lim_{d\to\infty}\sup_{x\in F_d}\left|\tilde{G}_d^{l,\Lambda}h_d(x)-G^{l,\Lambda}h(x^{(1:r+1)})\right|=0 \tag{3.90}$$

Proof. Let

$$A_d(x,Z^{(r+1:d)})=\sum_{i=r+1}^{d}\left[\nabla\log\pi(x^{(i)})'Z^{(i)}-\frac{l^2}{2(d-1)}\nabla\log\pi(x^{(i)})'\Lambda\nabla\log\pi(x^{(i)})\right] \tag{3.91}$$

and let

$$W_d(x^{(1:d)})=\frac{1}{2}\sum_{i=r+1}^{d}\left[Z^{(i)\prime}[\nabla^2\log\pi(x^{(i)})]Z^{(i)}+\tfrac{l^2}{(d-1)}(\nabla\log\pi(x^{(i)}))'\Lambda(\nabla\log\pi(x^{(i)}))\right] \tag{3.92}$$

Thus, since $y\mapsto 1\wedge e^y$ is 1-Lipschitz,

$$\begin{aligned}
&\sup_{x\in F_d}\left|\mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[1\wedge e^{A_d(x,Z^{(r+1:d)})}\right]-\mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[1\wedge e^{B_d(x,Z^{(r+1:d)})}\right]\right|\\
&\le\sup_{x\in F_d}\mathbb{E}\left|W_d(x)\right|+U_d(x)\\
&\le\theta(d)+\sqrt{2Ll^2K_\Lambda\frac{\log d}{d}}+\sup_{x\in F_d}\mathbb{E}\left|W_d(x)\right|\ .
\end{aligned} \tag{3.93}$$

where U_d was defined in the Section 3.4.3.1.

Let $\phi_d=\theta(d)+\sqrt{2Ll^2K_\Lambda\frac{\log d}{d}}+\sup_{x\in F_d}\mathbb{E}\left|W_d(x)\right|$. By Lemmas 3.5 and 3.9, $\phi_d\to 0$.

For the second result, let

$$q_d(x,Z)=\left(e^{A_d(x,Z^{(r+1:d)})};A_d(x,Z^{(r+1:d)})<0\right)-\left(e^{C_d(x,Z)};C_d(x,Z)<0\right) \tag{3.94}$$

and let $\delta_d=\delta_{1,d}+\delta_{2,d}$ where $\delta_{1,d}=\sqrt{\phi_d}$ and $\delta_{2,d}=(d-1)^{-1/4}$. For simplicity in the rest of the proof, we abbreviate $q_d(x,Z)$,$q_{1,d}(x,Z)$, $q_{2,d}(x,Z)$, and $A_d(x,Z^{(r+1:d)})$, as q_d, $q_{1,d}$, $q_{2,d}$, and A_d respectively.

$$\mathbb{E}_Z |q_d| \leq \delta_d \mathbb{P}(|q_d| \leq \delta_d) + \mathbb{P}(|q_d| > \delta_d) \tag{3.95}$$

The first term is $O(\delta_d)$, uniformly in x, so its $\lim_{d\to\infty} \sup_{x\in F_d}$ is 0.

The second term can be bounded as:

$$\begin{aligned}
\mathbb{P}_Z(|q_d| > \delta_d) &= \mathbb{P}_Z(|q_d| > \delta_d;\ A_d(x,Z) < 0;\ C_d < 0) \\
&\quad + \mathbb{P}_Z(|q_d| > \delta_d;\ A_d \geq 0;\ C_d < 0) \\
&\quad + \mathbb{P}_Z(|q_d| > \delta_d;\ A_d < 0;\ C_d \geq 0) \\
&\leq \mathbb{P}_Z(|A_d - C_d| > \delta_d;\ A_d < 0;\ C_d < 0) \\
&\quad + \mathbb{P}_Z(A_d \geq 0;\ C_d < 0) + \mathbb{P}_Z(A_d < 0;\ C_d \geq 0) \\
&\leq \mathbb{P}_Z(|A_d - C_d| > \delta_d;\ A_d < 0;\ C_d < 0) \\
&\quad + \mathbb{P}_Z(|A_d - C_d| > \delta_d;\ A_d \geq 0;\ C_d < 0) + \mathbb{P}(|A_d - C_d| > \delta_d;\ A_d < 0;\ C_d \geq 0) \\
&\quad + \mathbb{P}_Z(|A_d - C_d| \leq \delta_d;\ A_d \geq 0;\ C_d < 0) + \mathbb{P}_Z(|A_d - C_d| \leq \delta_d;\ A_d < 0;\ C_d \geq 0) \\
&\leq \mathbb{P}_Z(|A_d - C_d| > \delta_d) + \mathbb{P}_Z(-\delta_d \leq A_d \leq \delta_d) \\
&\leq \mathbb{P}_Z(|A_d - B_d| > \delta_{1,d}) + \mathbb{P}_Z(|B_d - C_d| > \delta_{2,d}) + \mathbb{P}_Z(-\delta_d \leq A_d \leq \delta_d)
\end{aligned} \tag{3.96}$$

First, by Markov's Inequality, uniformly in $x \in F_d$

$$\mathbb{P}_Z(|A_d - B_d| > \delta_{1,d}) \leq \frac{1}{\delta_{1,d}} \phi_d \leq \sqrt{\phi_d} \tag{3.97}$$

Second, for each $x \in F_d$

$$
\begin{aligned}
&\mathbb{P}_Z(|B_d - C_d| > \delta_{2,d}) \\
&= \mathbb{P}_Z(|\mathcal{E}| > \delta_{2,d}) \\
&= \mathbb{P}_Z\left(\left|\left(\nabla \log \pi^{\otimes r}(x^{(1:r)})Z^{(1:r)} + \int_0^1 (1-h)Z^{(1:r)\prime}\nabla^2 \log \pi^{\otimes r}(x^{(i)} + hZ^{(1:r)})Z^{(1:r)}\ dh\right)\right| > \delta_{2,d}\right) \\
&\le \mathbb{P}_Z\left(\left\|\nabla \log^{\otimes r} \pi(x^{(1:r)})\right\| \left\|Z^{(1:r)}\right\| + \frac{L}{2}\left\|Z^{(1:r)}\right\|^2 > \delta_{2,d}\right) \\
&\le \mathbb{P}_Z\left(\left\|\nabla \log \pi^{\otimes r}(x^{(1:r)})\right\| \left\|Z^{(1:r)}\right\| > \delta_{2,d}/2\right) + \mathbb{P}_Z\left(\frac{L}{2}\left\|Z^{(1:r)}\right\|^2 > \delta_{2,d}/2\right) \\
&\le \mathbb{P}_Z\left(2\sqrt{krL\log d}\left\|Z^{(1:r)}\right\| > \delta_{2,d}/2\right) + \mathbb{P}_Z\left(\frac{L}{2}\left\|Z^{(1:r)}\right\|^2 > \delta_{2,d}/2\right) \\
&\le \underset{W\sim \mathrm{N}(0,I_r\otimes\Lambda)}{\mathbb{P}}\left(\|W\|^2 > \frac{\delta_{2,d}^2(d-1)}{16krL\log d}\right) + \underset{W\sim \mathrm{N}(0,I_r\otimes\Lambda)}{\mathbb{P}}\left(\|W\|^2 > (d-1)L\delta_{2,d}\right) \\
&\le \underset{W\sim \mathrm{N}(0,I_r\otimes\Lambda)}{\mathbb{P}}\left(\|W\|^2 > \frac{\sqrt{d-1}}{16krL\log d}\right) + \underset{W\sim \mathrm{N}(0,I_r\otimes\Lambda)}{\mathbb{P}}\left(\|W\|^2 > (d-1)^{3/4}L\right) \\
&\le r\mathrm{Tr}(\Lambda)\left(\frac{16krL\log d}{\sqrt{d-1}} + \frac{1}{(d-1)^{3/4}L}\right)
\end{aligned}
\tag{3.98}
$$

Third, since $A_d \sim \mathrm{N}(-l^2R_d/2, l^2R_d)$, and since $|R_d - J| \le d^{-1/8} + J\frac{r-1}{d-1}$ on F_d, and since $J > 0$, we have:

$$
\begin{aligned}
\sup_{x\in F_d} \mathbb{P}(-\delta_d < A_d < \delta_d) &= \sup_{x\in F_d} \Phi(l\sqrt{R_d}/2 + \delta_d/\sqrt{lR_d}) - \Phi(\sqrt{lR_d}/2 - \delta_d/\sqrt{lR_d}) \\
&\le \sup_{x\in F_d} \delta_d\sqrt{\frac{2}{\pi l R_d}} \\
&\le \delta_d\sqrt{\frac{2}{\pi l(J - d^{-1/8} - J\frac{r-1}{d-1})}}
\end{aligned}
\tag{3.99}
$$

Thus, $\lim_{d\to\infty}\sup_{x\in F_d}\mathbb{P}(-\delta_d < A_d < \delta_d) = 0$.

Now, since $A_d \sim \mathrm{N}(-l^2R_d/2, l^2R_d)$, by Proposition 3.3:

$$
\mathbb{E}[1 \wedge e^{A_d}] = 2\Phi(-l\sqrt{R_d}/2) \tag{3.100}
$$

Thus, because $J > 0$, we have:

$$\begin{aligned}
&\lim_{d\to\infty}\sup_{x\in F_d}\left|\mathbb{E}[1\wedge e^{A_d}] - 2\Phi(-l\sqrt{J}/2)\right| \\
&= \lim_{d\to\infty}\sup_{x\in F_d}\left|2\Phi(-l\sqrt{R_d}/2) - 2\Phi(-l\sqrt{J}/2)\right| \\
&\leq \lim_{d\to\infty}\sup_{x\in F_d}\sqrt{\frac{l}{2\pi}}\left|\sqrt{R_d}-\sqrt{J}\right| \\
&\leq \sqrt{\frac{l}{2\pi}}\lim_{d\to\infty}\sup_{x\in F_d}\frac{|R_d - J|}{2\sqrt{\min(R_d, J)}} \\
&\leq \sqrt{\frac{l}{2\pi}}\lim_{d\to\infty}\sup_{x\in F_d}\frac{\left(d^{-1/8} + J\frac{r-1}{d-1}\right)}{2\sqrt{J - d^{-1/8} - J\frac{r-1}{d-1}}} \\
&= 0
\end{aligned} \tag{3.101}$$

Analogously, for the truncated expectation:

$$\begin{aligned}
&\lim_{d\to\infty}\sup_{x\in F_d}\left|\mathbb{E}[1\wedge e^{A_d}; A_d < 0] - \Phi(-l\sqrt{J}/2)\right| \\
&= \lim_{d\to\infty}\sup_{x\in F_d}\left|\Phi(-l\sqrt{R_d}/2) - \Phi(-l\sqrt{J}/2)\right| \\
&= 0
\end{aligned} \tag{3.102}$$

Finally, since h is smooth with compact support, then both of the functions $\left\|\nabla^2 h(x)\right\|_F$ and $\|\nabla \log \pi^{\otimes r}(x)\| \, \|\nabla h(x)\|$ are continuous with compact support, and hence they are both uniformly bounded over $x \in \mathbb{R}^k$. Let $M_h < \infty$ be a uniform bound on both. Then:

$$\begin{aligned}
&\lim_{d\to\infty}\sup_{x\in F_d}\left|\tilde{G}_d^{l,\Lambda} h_d(x) - G^{l,\Lambda} h(x)\right| \\
&\quad\leq \lim_{d\to\infty}\sup_{x\in F_d}\frac{kl^2}{2}\left|2\Phi(-l\sqrt{I}/2) - \mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[1\wedge e^{B_d}\right]\right| \|\Lambda\|_F \left\|\nabla^2 h(x)\right\|_F \\
&\qquad\qquad + kl^2\left|\Phi(-l\sqrt{I}/2) - \mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[e^{C_d}; C_d < 0\right]\right| \|\nabla \log \pi^{\otimes r}(x)\| \, \|\Lambda\| \, \|\nabla h(x)\| \\
&\quad\leq M_h \lim_{d\to\infty}\sup_{x\in F_d}\frac{kl^2}{2}\left|2\Phi(-l\sqrt{I}/2) - \mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[1\wedge e^{B_d}\right]\right| + kl^2\left|\Phi(-l\sqrt{I}/2) - \mathop{\mathbb{E}}_{Z^{(r+1:d)}}\left[e^{C_d}; C_d < 0\right]\right| \\
&\quad= 0
\end{aligned} \tag{3.103}$$

$\square$

3.5 Additional Lemmas for the Proof of Theorem 3.1

Proposition 3.3 (Acceptance Moments (Roberts et al. [94], proposition 2.4))**.** *If* $W \sim \mathrm{N}(\mu, \sigma^2)$ *then*

$$\mathbb{E}[1 \wedge e^W] = \Phi(\mu/\sigma) + e^{\mu+\sigma^2/2}\Phi(-\sigma - \mu/\sigma) \tag{3.104}$$

and

$$\mathbb{E}[e^W; W < 0] = e^{\mu+\sigma^2/2}\Phi(-\sigma - \mu/\sigma) \tag{3.105}$$

Lemma 3.9. *Let*

$$\begin{aligned} W_d(x) &= \frac{1}{2}\sum_{i=r+1}^{d}\Big[Z^{(i)\prime}[\nabla^2 \log \pi(x^{(i)})]Z^{(i)} + \tfrac{l^2}{(d-1)}(\nabla \log \pi(x^{(i)}))'\Lambda(\nabla \log \pi(x^{(i)}))\Big] \\ &= \frac{1}{2}\sum_{i=r+1}^{d}\Big[[\nabla^2 \log \pi(x^{(i)})]{:}[Z^{(i)}Z^{(i)\prime}] + \tfrac{l^2}{(d-1)}(\nabla \log \pi(x^{(i)}))'\Lambda(\nabla \log \pi(x^{(i)}))\Big] \end{aligned} \tag{3.106}$$

where $Z^{(i)} \overset{iid}{\sim} \mathrm{N}\left(0, \frac{l^2\Lambda}{(d-1)}\right)$.

Then $\lim_{d\to\infty} \sup_{x^{(1:d)}\in F_d} \mathbb{E}\,|W_d(x)| = 0$

Proof. Using Isserlis' theorem (Isserlis [46], equation (39)*):

$$\begin{aligned} &\mathbb{E}\left[([\nabla^2 \log \pi(x^{(i)})] : [Z^{(i)}Z^{(i)\prime}])^2\right] \\ &= \mathbb{E}\left[\left(\sum_{\alpha,\beta} \frac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta} Z^{(i)}_\alpha Z^{(i)}_\beta\right)^2\right] \\ &= \sum_{\alpha,\beta,\gamma,\delta} \frac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta}\frac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\gamma \partial x^{(i)}_\delta}\mathbb{E}\left[Z^{(i)}_\alpha Z^{(i)}_\beta Z^{(i)}_\gamma Z^{(i)}_\delta\right] \\ &= \frac{l^4}{(d-1)^2}\sum_{\alpha,\beta,\gamma,\delta} \tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta}\tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\gamma \partial x^{(i)}_\delta}\left(\Lambda_{\alpha\beta}\Lambda_{\gamma\delta} + \Lambda_{\alpha\gamma}\Lambda_{\beta\delta} + \Lambda_{\alpha\delta}\Lambda_{\beta\gamma}\right) \\ &= \frac{l^4}{(d-1)^2}\left[\left(\sum_{\alpha,\beta} \tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta}\Lambda_{\alpha\beta}\right)^2 + 2\sum_{\alpha,\beta,\gamma,\delta} \tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta}\tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\gamma \partial x^{(i)}_\delta}\Lambda_{\alpha\gamma}\Lambda_{\beta\delta}\right] \\ &= \frac{l^4}{(d-1)^2}\left[[\Lambda : \nabla^2 \log \pi(x^{(i)})]^2 + 2\sum_{\alpha,\beta,\gamma,\delta} \tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\alpha \partial x^{(i)}_\beta}\tfrac{\partial^2 \log \pi(x^{(i)})}{\partial x^{(i)}_\gamma \partial x^{(i)}_\delta}\Lambda_{\alpha\gamma}\Lambda_{\beta\delta}\right] \end{aligned} \tag{3.107}$$

where all Greek subscripts range over $\{1, ..., k\}$.

Thus we have:

$$
\begin{aligned}
&[\mathbb{E}\,|W_d(x)|]^2 \\
&\leq \mathbb{E}\left[|W_d(x)|^2\right] \\
&= \frac{l^4}{4(d-1)^2}\left(\sum_{i=r+1}^{d}\left([\Lambda{:}\nabla^2\log\pi(x^{(i)})]+(\nabla\log\pi(x^{(i)}))'\Lambda(\nabla\log\pi(x^{(i)}))\right)\right)^2 \\
&\quad + \frac{2l^4}{4(d-1)^2}\sum_{i=r+1}^{d}\sum_{\alpha,\beta,\gamma,\delta}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\alpha^{(i)}\partial x_\beta^{(i)}}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\gamma^{(i)}\partial x_\delta^{(i)}}\Lambda_{\alpha\gamma}\Lambda_{\beta\delta}
\end{aligned}
\tag{3.108}
$$

For $x^{(1:d)} \in F_d$ we have:

$$
\frac{1}{2(d-1)}\left|\sum_{i=r+1}^{d}\left([\Lambda{:}\nabla^2\log\pi(x^{(i)})]+(\nabla\log\pi(x^{(i)}))'\Lambda(\nabla\log\pi(x^{(i)}))\right)\right| \leq d^{-1/8} \tag{3.109}
$$

Since $\nabla \log \pi$ is Lipschitz, the second order partials of $\log \pi$ are essentially bounded, hence $\sum_{\alpha,\beta,\gamma,\delta}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\alpha^{(i)}\partial x_\beta^{(i)}}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\gamma^{(i)}\partial x_\delta^{(i)}}\Lambda_{\alpha\gamma}\Lambda_{\beta\delta}$ is essentially bounded. Thus:

$$
\sup_{x\in\mathbb{R}^{kd}}\frac{2l^2}{4(d-1)^2}\sum_{i=r+1}^{d}\sum_{\alpha,\beta,\gamma,\delta}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\alpha^{(i)}\partial x_\beta^{(i)}}\frac{\partial^2\log\pi(x^{(i)})}{\partial x_\gamma^{(i)}\partial x_\delta^{(i)}}\Lambda_{\alpha\gamma}\Lambda_{\beta\delta} \in O(1/d) \tag{3.110}
$$

Combining these two limits we get that $\lim_{d\to\infty}\sup_{x\in F_d}\mathbb{E}\,|W_d(x)| = 0$. □

3.6 Proof of Theorem 3.2

To prove convergence of $\mathbf{Y}_d$ to $\mathbf{X}^{\mathbb{N}}$ in the Skorohod topology (of $\mathbb{R}^{\mathbb{N}}$ with the product topology) we need the following lemma:

Lemma 3.10. *If $\mathbb{R}^{\mathbb{N}}$ is equipped with the metric*

$$
r(x,y) = \sum_{i\geq 1} 2^{-i}(|x_i - y_i| \wedge 1), \tag{3.111}
$$

which happens to generate the product topology, then

$$
M = \left\{ f \circ \rho_j \ \ s.t.\ j \in \mathbb{N}, f \in \overline{C}(\mathbb{R}^j)\right\} \tag{3.112}
$$

strongly separates points (where $\rho_j : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^j$ is the projection map onto the first j components).

Proof of Lemma 3.10. Fix $1 > \delta > 0$ and $x \in \mathbb{R}^{\mathbb{N}}$, and let $m_\delta = \lceil -\log_2(\delta) \rceil$. Let

$$h_{x,\delta}(z) = \frac{2}{\delta}\left(\frac{\delta}{2} - \sum_{i=1}^{m_\delta+1} 2^{-i}(|x_i - z_i| \wedge 1)\right)_+ \quad . \tag{3.113}$$

Notice that $h_{x,\delta} \in M$. Obviously $h_{x,\delta}(x) = 1$.

Suppose $y \in \mathbb{R}^{\mathbb{N}}$ such that $r(x,y) \geq \delta$; since

$$\sum_{i=m_\delta+2}^{\infty} 2^{-i}(|x_i - y_i| \wedge 1) \leq 2^{-(m_\delta+1)} \leq \delta/2 \ , \tag{3.114}$$

then

$$\sum_{i=1}^{m_\delta+1} 2^{-i}(|x_i - y_i| \wedge 1) \geq \delta/2 \tag{3.115}$$

and hence $h_{x,\delta}(y) = 0$. □

Proof of Theorem 3.2. By the Kolmogorov extension theorem (see for example [114], section 2.4 therein) applied to the sequence $\mathbf{X}^r$ over $r \in \mathbb{N}$, there is a unique (in probability law) process $\mathbf{X}^{\mathbb{N}}$ taking values in $\mathbb{R}^{\mathbb{N}}$ such that the marginal process of the first kr components has the same distribution as $\mathbf{X}^r$.

From Lemma 3.10, M (as defined above) strongly separates points. Consider any finite subset of M, say $\{h_1, ..., h_n\}$. Then without loss of generality there exists an $m \in \mathbb{N}$ with and a set of functions $\{f_1, ..., f_n\} \subset \overline{C}(\mathbb{R}^{mk})$ with $h_i = f_i \circ \rho_{mk}$ for all $i \in \{1, ..., n\}$.

If E is a metric space, and $f : E \to \mathbb{R}$ then define its "lift" onto $D_E[0,\infty)$ as $\tilde{f} : (t \mapsto X(t)) \mapsto (t \mapsto f(X(t))$, so that $\tilde{f} : D_E[0,\infty) \to D_{\mathbb{R}}[0,\infty)$. If f is continuous in the topology on E then $\tilde{f}$ must be continuous in the Skorohod topology on $D_E[0,\infty)$. This fact is proven by Jakubowski [49] (theorem 4.3 therein)[3].

Now, since all of the finite dimensional processes of $\mathbf{Y}_d$ converge weakly in the Skorohod topology, and since the lift of a continuous function on $\mathbb{R}^{km}$ to $D_{\mathbb{R}^{km}}[0,\infty)$ is continuous

[3]In fact, Jakubowski [49] tells us the stronger result, that the Skorohod topology on $D_E(0,\infty]$ is the coarsest topology for which the lifts of all continuous functions are continuous.

then, by the continuous mapping theorem,

$$(h_1, ..., h_n)(\mathbf{Y}_d) \Rightarrow (h_1, ..., h_n)(\mathbf{X}^{\mathbb{N}}) \ . \tag{3.116}$$

By Ethier and Kurtz [33] (corollary 9.2 therein) this is sufficient to ensure that $\mathbf{Y}_d$ converges weakly in the Skorohod topology to $\mathbf{X}^{\mathbb{N}}$.

Moreover, since the product topology on $\mathbb{R}^{\mathbb{N}}$ is generated by a collection of compatible pseudometrics, Jakubowski [49, Theorem 1.3] tells us that the Skorohod topology on $\mathbb{R}^{\mathbb{N}}$ defined using the metric r, only depends on the product topology of $\mathbb{R}^{\mathbb{N}}$; it does not depend on the choice of metric. □

3.7 Proofs of Scaling and Shaping Results

Proof of Corollary 3.1. For fixed Λ, $G^{l,\Lambda} = l^2 a_\Lambda(l)(\Lambda : \Sigma)G^\Lambda/2$. Thus $G^{l,\Lambda}$ corresponds to G^Λ accelerated (decelerated) by a factor of $l^2 a_\Lambda(l)(\Lambda : \Sigma)/2$. Hence, to maximize the speed of $G^{l,\Lambda}$ over the choice of l we need only maximize $h(l) := l^2 a_\Lambda(l) = 2l^2\Phi\left(-\frac{l\sqrt{\Sigma:\Lambda}}{2}\right)$ over l.

This is equivalent to the original optimization from Roberts et al. [94]. Notice that:

$$h(l) = \frac{8}{\Sigma : \Lambda}\left(\frac{l\sqrt{\Sigma : \Lambda}}{2}\right)^2 \Phi\left(-\frac{l\sqrt{\Sigma : \Lambda}}{2}\right) \tag{3.117}$$

Taking $\omega = \frac{l\sqrt{\Sigma:\Lambda}}{2}$ we can maximize instead:

$$\tilde{h}(\omega) = \omega^2\Phi(-\omega) \tag{3.118}$$

This may be done numerically to get $\omega_\star \approx 1.1906$, $\tilde{h}(\omega_\star) \approx 0.165717$. Then solving for l yields $l_\Lambda = \frac{2\omega_\star}{\sqrt{\Sigma:\Lambda}} = \frac{\approx 2.3812}{\sqrt{\Sigma:\Lambda}}$ and $h(l_\Lambda) = \frac{8\tilde{h}(\omega_\star)}{\Sigma:\Lambda} = \frac{\approx 1.32574}{\Sigma:\Lambda}$.

Hence $G^{l_\Lambda,\Lambda} = 4\tilde{h}(\omega_\star)G^\Lambda = (\approx 0.66)G^\Lambda$ □

Proof of Lemma 3.2. If G^Λ has a spectral gap for at least one strictly positive definite shaping matrix, Λ, then it has a spectral gap for all strictly positive definite shaping matrices. This is true because, for f in a core of $\mathcal{D}(G^\Lambda) = \mathcal{D}(G^\Theta)$, we can use the Dirichlet form corresponding to the generator (see, for example, [7]) to write

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim\pi}\Big(f(X)\ [G^{\Lambda} f](X)\Big) &= \mathop{\mathbb{E}}_{X\sim\pi}\left(\nabla f(X)' \frac{\Lambda}{\Lambda:\Sigma} \nabla f(X)\right) \\
&\leq \frac{\frac{\lambda_{\max}(\Lambda)}{\Lambda:\Sigma}}{\frac{\lambda_{\min}(\Theta)}{\Theta:\Sigma}} \mathop{\mathbb{E}}_{X\sim\pi}\left(\nabla f(X)' \frac{\Theta}{\Theta:\Sigma} \nabla f(X)\right) \\
&= \frac{\lambda_{\max}(\Lambda)}{\lambda_{\min}(\Theta)}\ \frac{\Theta:\Sigma}{\Lambda:\Sigma} \mathop{\mathbb{E}}_{X\sim\pi}\Big(f(X)\ [G^{\Theta} f](X)\Big),
\end{aligned} \tag{3.119}$$

implying that

$$\rho(G^{\Lambda})\frac{\lambda_{\max}(\Theta)}{\lambda_{\min}(\Lambda)}\ \frac{\Lambda:\Sigma}{\Theta:\Sigma} \geq \rho(G^{\Theta}) \geq \rho(G^{\Lambda})\frac{\lambda_{\min}(\Lambda)}{\lambda_{\max}(\Theta)}\ \frac{\Theta:\Sigma}{\Lambda:\Sigma}. \tag{3.120}$$

□

Chapter 4

Statistical Inference with Stochastic Gradient Methods

4.1 Introduction

Stochastic gradient algorithms were originally proposed as optimization methods by Robbins and Monro [93], and have become the unequivocal standard for large scale optimization problems in statistics and machine learning. The success of stochastic gradient methods is due to the fact that improvements in computational complexity from subsampling outweigh the accuracy loss from stochastic approximation for empirical objectives, and thus stochastic optimization methods scale more favourably with the sample size and model complexity than their deterministic counterparts. Moreover theoretical results for stochastic optimization demonstrate that it can match the accuracy of deterministic methods. In contrast, classical and exact gradient-based MCMC methods cannot directly benefit from subsampling in the same way as optimization, due to the need to accept/reject using the full-sample likelihood in the Metropolis-Hastings adjustment. Thus, apparently, one must either sacrifice the speed gains from subsampling, or lose accuracy relative to non-stochastic-gradient methods for sampling.

Regardless of the loss in accuracy, the need for faster sampling methods has lead researchers to use approximate, unadjusted MCMC methods, based on discretizations of continuous time stochastic processes (e.g., [29, 30, 31, 68]), and their stochastic gradient counterparts [121]. While these methods may be asymptotically exact when run with decreasing step-sizes, this does not directly characterize their finite time behaviour, and the use of decreasing step-sizes means that time we wait for the next effective sample is ever increasing. A practical way around this is to accept that some approximation error will persist, and to work with fixed step-sizes. When working with fixed step-sizes, then, it is important to understand in what way and to what degree the use of stochastic gradients affects the samples we generate, and how to tune our stochastic gradient algorithms to provide accurate uncertainty quantification.

This chapter addresses questions of accuracy, tuning, and robustness of stochastic gradient algorithms with fixed step-sizes for approximate sampling by examining the scaling limits of stochastic gradient algorithms as the sample size tends to infinity. We show that the sample paths of stochastic gradient algorithms with fixed step-sizes converge weakly in the Skorohod topology in probability to the sample paths of an Ornstein–Uhlenbeck process under relatively mild statistical conditions. We then use the properties of the limiting process (e.g., stationary law, mixing time, etc.) to characterize the corresponding properties of the stochastic gradient algorithm. The scaling limit result we prove provides rigorous justification for the similar scaling limit proposed heuristically by Mandt et al. [70], and a more complete theory characterizing such limits by examining a wider range of tuning parameters and their relative scalings. The present work is further motivated by an empirical finding: that properly tuned stochastic gradient methods can yield approximate uncertainty quantification that is asymptotically more robust to model misspecification. More specifically, we show that the tuning parameters of stochastic gradient algorithms can be chosen so that the stationary distributions of the limiting process matches either the asymptotics of the posterior, the bagged posterior, or a local asymptotic fiducial distribution for the maximal likelihood estimator.

4.1.1 Formalism

Let $\mathbf{X}^{(n)} = (X_i)_{i=1}^n \in \mathcal{X}^n$ denote a dataset with observations X_i independently and identically distributed (i.i.d.) from an unknown distribution P. Consider the potential

$\mathcal{U}^{(n)}(\theta) := \log \pi^{(0)}(\theta) + \sum_{i=1}^{n} \ell(\theta; X_i)$ where ℓ typically either represents a log-likelihood, or a negative loss function; and $\pi^{(0)}$ typically represents either a (possibly improper) prior density that is everywhere positive on Θ, or $\log \pi^{(0)}(\theta)$ is a regularizer. The potential $\mathcal{U}^{(n)}(\theta)$ can be viewed either in a frequentist setting as the complete log-likelihood with regularizer $\log \pi^{(0)}(\theta)$ or in a Bayesian setting as the log of the joint model density.

In the frequentist case, perhaps the most popular estimator for the (locally) optimal population parameter $\theta_\star$ satisfying $\mathbb{E}\{\nabla \ell\,(\theta_\star; X_1)\} = 0$, is the M-estimator $\widehat{\theta}^{(n)}$ satisfying the first-order optimality condition $\nabla \mathcal{U}^{(n)}(\widehat{\theta}^{(n)}) = 0$. In the Bayesian case, the quantity of interest is (usually) an expectation with respect to the posterior density $\pi^{(n)}(\theta) \propto \exp\{-\mathcal{U}^{(n)}(\theta)\}$ of a function $f : \Theta \to \mathbb{R}^\ell$, which we denote $\pi^{(n)}(f)$. In either case, when n is large relative to computational cost of evaluating $\ell(\theta; X_i)$, classical optimization methods (e.g., Newton–Raphson) for approximating $\widehat{\theta}^{(n)}$ and sampling methods for estimating $\pi^{(n)}(f)$ (e.g., Metropolis-Hastings methods) become computationally prohibitive.

Stochastic gradient algorithms have been widely adopted for both optimization and sampling as a means to reducing computational cost of each iteration of an iterative method, improving scalability. For a stochastic gradient algorithm, to generate a sequence of iterates $\theta_1^{(n)}, \ldots, \theta_k^{(n)}, \ldots \in \Theta$, rather than computing exact gradients of $n^{-1}\mathcal{U}^{(n)}$ using the full dataset, at iteration k a small batch of subsampled data is instead used to compute an unbiased gradient estimate

$$\hat{G}_k^{(n)} := \frac{1}{n} \nabla \log \pi^{(0)} \left(\theta_k^{(n)} \right) + \frac{1}{b^{(n)}} \sum_{j=1}^{b^{(n)}} \nabla \ell \left(\theta_k^{(n)};\ X_{I_k^{(n)}(j)} \right), \tag{4.1}$$

where $(I_k^{(n)})_{k \in \mathbb{N}} \in ([n]^b)^{\mathbb{N}}$ are an independent and identically distributed (i.i.d.) sequence of uniform random samples from $\{1, \ldots, n\}$ of size $b^{(n)}$, which are formed either with or without replacement.[1]

Most analyses of stochastic gradient optimization procedures focus on the optimality gap, while analyses of stochastic gradient sampling procedures focus on how well the standard posterior is approximated. In practice stochastic gradient algorithms appear to be

[1] "With replacement" means that $(I_k^{(n)})_{k \in \mathbb{N}} \overset{\text{iid}}{\sim} \mathrm{Unif}([n]^b)$ and "without replacement" means that $(I_k^{(n)})_{k \in \mathbb{N}} \overset{\text{iid}}{\sim} \mathrm{Unif}(\{I \in [n]^b \text{ s.t. } (j_1 \neq j_2 \implies I(j_1) \neq I(j_2))\})$.

successful even when used with tuning parameter combinations (e.g., large step-size and small batch size) insufficient to result in accurate approximations according to the standard theory. The lack of an explanatory theory has forced users to rely on heuristic and problem-specific approaches to setting tuning parameters. The aim of the present work is to take a step toward filling this gap, allowing users to understand how the choice of tuning parameters affect the statistical properties of the algorithms. Our approach is motivated by the hypothesis that, while the variability introduced by subsampling is usually viewed as detrimental, it is plausible—in light of the success of methods like the bootstrap—that it could also be beneficial. For example, an immediate consequence of Polyak and Juditsky [92] is that averaging the iterates of stochastic gradient descent (SGD), whose one-step updates are

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h_k^{(n)}}{2}\hat{G}_k^{(n)}, \tag{4.2}$$

can provide automatic optimal uncertainty quantification in maximum likelihood estimation. More precisely, when $h_k \propto k^{-\varsigma}$ for $\varsigma \in (0,1)$, the iterate average $\bar{\theta}_k^{(n)} := \frac{1}{k}\sum_{k'=1}^{k}\theta_{k'}^{(n)}$ satisfies

$$\lim_{n\to\infty}\lim_{k\to\infty} k\,\mathrm{Cov}(\bar{\theta}_k^{(n)}) = \mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} = \lim_{n\to\infty} n\,\mathrm{Cov}(\widehat{\theta}^{(n)}), \tag{4.3}$$

where $\mathcal{I}_\star := \mathbb{E}\left\{\nabla_\theta \ell(\theta_\star; X) \otimes \nabla_\theta \ell(\theta_\star; X)\right\}$ is the first-order Fisher information matrix and $\mathcal{J}_\star := -\mathbb{E}\{\nabla_\theta^{\otimes 2}\ell(\theta_\star; X)\}$ is the second-order Fisher information matrix.

We develop conceptually similar results that characterize the large-sample behavior of iterates of a large class of preconditioned stochastic gradient algorithms with fixed step-size $h_k^{(n)} = h^{(n)}$ depending on the sample size but not the iteration number. The fixed–step-size setting proves to be practically relevant because convergence to a near-optimum is rapid and robust to the precise step-size choice [27, 78]. Moreover, for sampling the fixed step-size setting leads to better mixing time behaviour as the number of iterations until the next approximately independent sample will be static, unlike in the decreasing-step-size regime where the number of iterations until the next approximately independent sample is increasing without bound.

4.1.2 Scope of the present work

Our analysis covers both optimization algorithms—including stochastic and deterministic gradient descent with and without momentum, SGD with Nesterov acceleration—and unadjusted Markov chain Monte Carlo (MCMC) algorithms—(preconditioned) over/underdamped Langevin dynamics with and without stochastic gradients—where the fixed–step-size setting is the most relevant since the iterates form a time-homogeneous Markov chain which typically has a well-defined stationary distribution and mixes after a fixed number of iterations. We characterize the behaviour not just of individual iterates, but also of the iterates jointly, which enables a unified analysis of stationarity, mixing, and iterate averaging properties of stochastic gradient algorithms. Specifically, we show that, near a local optimum, the iterates converge (weakly) to sample paths of an Ornstein–Uhlenbeck (OU) process in probability as the sample size tends to infinity and, jointly, the constant step-size decreases to 0. Under an additional regularity condition, we also establish that the global stationary distribution of the limiting OU process exists, implying a Bernstein–Von Mises-like theorem. Depending on the choice of step-size, preconditioner, batch size, and method used, the stationary covariance of the limiting OU process can be tuned to equal the sandwich covariance, the asymptotic posterior covariance $\mathcal{J}_\star^{-1}$, or a linear combination of the two. For example, consider stochastic gradient Langevin dynamics [SGLD; 121], a popular stochastic gradient MCMC algorithm that has one-step update given by

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h_k}{2}\hat{G}_k^{(n)} + \sqrt{\frac{h_k}{\beta}}\,\xi_k, \tag{4.4}$$

where $\xi_k \sim \mathrm{N}_d(0, I)$ is standard Gaussian noise and $\beta \in (0, \infty]$ is the inverse temperature, which is usually taken to be n.[2] We show that SGD and SGLD then have mixing times of order n when tuned to have asymptotic covariance matching the asymptotic posterior covariance, the asymptotic covariance of the MLE (the sandwich), or a mixture thereof (as would be obtained from the asymptotics of the *bagged posterior* [20, 45]). The latter tunings depart from targeting the posterior distribution, but still leads to plausible—and potentially better—uncertainty quantification.

[2]We take β^{-1} to mean 0 when $\beta = +\infty$, in which case we recover SGD from Eq. (4.2).

Overall, our results (1) demonstrate that stochastic gradient algorithms can provide computationally efficient, statistically robust asymptotic uncertainty quantification, particularly in the case of model misspecification, and (2) provide practical guidance to users of these algorithms for both optimization and sampling. Our theory is supported by a number of experiments, using both real and simulated data.

The assumptions required by our analysis are substantially weaker than the previous results that have characterized scaling limits of specific stochastic gradient algorithms. We allow the batch size used to compute the stochastic gradient to depend on the dataset size and allow the batches to be sampled with or without replacement. We only require the local maximizer to converge in probability and we do not assume the model is correctly specified. At the same time, our results are stronger than previous analyses since we characterize both the sample paths of the iterates and the complete stationary distribution; not just, e.g., first and second moments. As such, our results can be viewed a generalization and formalization of the heuristic arguments of Mandt et al. [70], and open the way for further generalizations to situations where heuristics provide minimal insight such as infinite-dimensional models and models where the number of parameters scales with the sample size. Our results also complement those of Kushner and Yin [59], who provides the basis for the assumptions in [70], and who do in fact establish weak convergence of stochastic gradient algorithms to OU process in a large number of settings. Notably, they do not cover the case that the objective function is itself stochastic (in particular arising as the random likelihood function based on an IID sample of size n), or the joint scaling of the objective function with the tuning parameters of the algorithm required to obtain asymptotic statistical results.

4.1.3 Asymptotic distributions and misspecification

In order to interpret our scaling limits in the context of asymptotic uncertainty quantification, it is important that we identify the relevant asymptotic distributions for Bayesian and frequentist inference under misspecification.

The Bernstein-von Mises theorem tells us that the posterior distribution of the parameter θ is asymptotically normal, centred at the MLE, with covariance $\mathcal{J}_\star^{-1}$. Similarly, the (local) maximum likelihood estimator $\widehat{\theta}^{(n)}$ is itself asymptotically normal, centred at the true parameter $\theta_\star$, with covariance equal to the "sandwich" covariance matrix, $\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}$. If the model is well-specified (i.e., $P = Q_\theta$ for some $\theta \in \Theta$), then $\mathcal{J}_\star = \mathcal{I}_\star$, and so $\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} = \mathcal{J}_\star^{-1}$. However, if the model is misspecified (i.e., $P \neq Q_\theta$ for any $\theta \in \Theta$), then the sandwich may differ from $\mathcal{J}_\star^{-1}$ [43, 123]. In this case, posterior credible sets are not asymptotically well-calibrated frequentist confidence sets [58].

The question of how to account for misspecification in the Bayesian setting has been addressed in a number of ways. The asymptotic normality of the posterior under misspecification was identified by Chen [24] and Bunke and Milhaud [22]. Shalizi [107] presents sufficient conditions for posterior convergence when the model hypotheses are wrong and the data have complex dependencies. Kleijn and van der Vaart [58] prove the Bernstein-Von-Mises theorem under misspecification. Bühlmann and van de Geer [21] investigate the robustness of asymptotic high-dimensional inference for misspecified linear models.

There are also several modified Bayesian approaches proposed in the literature. Royall and Tsou [101] show that the posterior based on the adjusted (profile) likelihood function [111] can be robust asymptotically. Müller [79] shows that Bayesian inference about the pseudo-true parameter under squared error has lower frequentist risk asymptotically when the posterior is substituted by an artificial normal posterior centred at the MLE with sandwich covariance matrix. Grünwald and van Ommen [38] study the use of power likelihood to improve robustness to misspecification and propose a method for choosing the power term. Bissiri et al. [14] suggest a general framework for Bayesian inference using a loss function rather than the traditional likelihood function. Recently, Huggins and Miller [45] study the use of bagging technique on the Bayesian posterior and develop the asymptotic theory. It is shown that under misspecification, the covariance of the "bagged posterior" is a mixture of $\mathcal{J}_\star^{-1}$ and the sandwich covariance.

4.1.4 Other Related Work

There is extensive work on the viability of stochastic gradient algorithms for approximate inference. Some examples which are relevant in the context of the present work include the following. Dieuleveut et al. [27] analyse the optimization properties of constant step-size stochastic gradient algorithms using tools from the theory of time homogeneous

Markov chains, and proposes numerical extrapolation methods to improve optimization performance. Toulis and Airoldi [117] characterizes the asymptotic first and second moments of the iterates of stochastic gradient algorithms, and the asymptotic normality of iterate averages, but not the full limiting distribution of the path-process, and show that these limits are robust to online tuning of algorithm parameters. Brosse et al. [19] study the asymptotic properties of SGD and SGLD, and find that, with naive tuning parameters, they do not provide an accurate representation of the posterior, while control-variate based methods do, which is consistent with our results. Teh et al. [115] study the consistency, CLT, and asymptotic bias–variance decomposition of SGLD for a sequence of decreasing step-sizes that converge zero. Vollmer et al. [120] characterize the asymptotic bias of constant step-size SGLD explicitly with its dependence on the step-size and the variance of the stochastic gradient, as well as bounds on the finite-time bias, variance and mean squared error (MSE). Mandt et al. [70] study constant learning rate SGD by approximating it with a continuous-time Ornstein–Uhlenbeck process. The conclusions they draw are similar to ours, however that the OU process is a good approximation for SGD in the large-sample scaling limit is taken as an assumption in that work, while we prove that the limit does in fact hold under reasonable conditions. They compute the stationary distribution for a class of SGD algorithms, all of which converge to a Gaussian distribution parameterized by the learning rate, mini-batch size and preconditioning matrix. Tzen et al. [118] study path-wise behaviour of discrete-time Langevin algorithm for non-convex empirical risk minimization through metastability. They show that, for a particular local optimum of the empirical risk, with high probability, either the Langevin trajectory ends up outside the a neighbourhood of this local optimum within a short recurrence time, or it enters this neighbourhood by the recurrence time and stays there until a potentially exponentially long escape time. This states that the Langevin scheme will eventually visit all local minima, but it will take an exponentially long time to transit among them. Yu et al. [125] show that the average of constant learning rate SGD iterates is asymptotically normally distributed around the expected value of their unique invariant distribution, as long as the non-convex and non-smooth objective function satisfies a dissipativity property. Liu et al. [66] study the stationary distribution of discrete-time SGD and its variants in a quadratic loss function and obtain the analytic form for the asymptotic covariance matrix of the model parameters. The asymptotics of their results agree with ours.

Some existing work studies a continuous-time process by assuming it as a model of an iterative algorithm. For example Li et al. [63] study the stochastic modified equations framework for analyzing the dynamics of stochastic gradient algorithms, where the latter is approximated by a class of stochastic differential equations with small noise parameters. Gupta et al. [40] consider recursive stochastic algorithms as approximations of certain contraction operators and view them within the framework of iterated random operators. Sirignano and Spiliopoulos [108, 109] study stochastic gradient descent in continuous time, where the algorithm follows a (noisy) descent direction along a continuous stream of data and the parameter updates occur in continuous time and satisfy a stochastic differential equation.

Weak convergence techniques have become very popular in the theoretical MCMC literature since the seminal paper of Roberts et al. [94]. However, most of analyses have been performed in the asymptotic regime where the parameter dimension $d \to \infty$. The "large-sample regime" where d is fixed and the number of data goes to infinity has been recently studied in [105, 106] for random-walk Metropolis algorithms. To the best of our knowledge, our work is the first work for analyzing stochastic gradient algorithms in the "large-sample regime" using the weak convergence techniques originating from the MCMC optimal scaling literature.

This chapter does not cover several popular topics in recent literature, which we leave as future directions. For example, studying properties of stochastic gradient algorithms for overparameterized models (see e.g. [64, 124, 127]) and for Bayesian deep learning (see e.g. [1, 47, 76, 122]).

4.1.5 Additional notation

Let $\mathcal{M}_{1,+}(\mathcal{X})$ denote the set of probability measures on the observation space $\mathcal{X}$ and suppose that X_i ($i \in \mathbb{N} := \{1, 2, \ldots\}$) are independent and identically distributed (i.i.d.) from $P \in \mathcal{M}_{1,+}(\mathcal{X})$. Let $\mathbf{X}^{(n)} = (X_i)_{i \in [n]}$, where $n \in \mathbb{N}$ is the sample size and $[n] := \{1, \ldots, n\}$. Fix a model $\{Q_\theta : \theta \in \Theta\} \subseteq \mathcal{M}_{1,+}(\mathcal{X})$ for P and unless otherwise noted take the parameter space $\Theta = \mathbb{R}^d$. We will assume that $\{Q_\theta\}_{\theta\in\Theta}$ are absolutely continuous with respect to a common base measure μ on $\mathcal{X}$, and write their densities as $q_\theta := \mathrm{d}Q_\theta/\mathrm{d}\mu$ for each $\theta \in \Theta$, and their log-likelihood functions as $\ell(x;\theta) := \log q_\theta(x)$. For a $\theta_\star \in \Theta$, the first-order and second-order Fisher information matrices at $\theta_\star$ are (respectively) defined by

$$\begin{aligned} \mathcal{I}_\star = \mathcal{I}(\theta_\star) &:= \mathop{\mathbb{E}}_{X\sim P}\left\{\nabla_\theta \ell(\theta_\star; X) \otimes \nabla_\theta \ell(\theta_\star; X)\right\}, \qquad \text{and} \\ \mathcal{J}_\star = \mathcal{J}(\theta_\star) &:= -\mathop{\mathbb{E}}_{X\sim P}\{\nabla_\theta^{\otimes 2} \ell(\theta_\star; X)\}. \end{aligned} \tag{4.5}$$

Let

$$\begin{aligned} \widehat{\mathcal{I}}^{(n)}(\theta) &= \frac{1}{n}\sum_{i\in[n]} [\nabla \ell(\theta;\, X_i)]^{\otimes 2}, \qquad \text{and} \\ \widehat{\mathcal{J}}^{(n)}(\theta) &= \frac{1}{n}\sum_{i\in[n]} [-\nabla^{\otimes 2} \ell(\theta;\, X_i)] \end{aligned}$$

denote the empirical first-and second-order Fisher information matrix functions, respectively.

For $d \in \mathbb{N}$, denote the d-dimensional Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and (positive semi-definite) covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ as $\mathrm{N}_d(\mu, \Sigma)$. For vectors $a, b \in \mathbb{R}^d$, define the outer product $a \otimes b \in \mathbb{R}^{d\times d}$ given by $(a \otimes b)_{ij} = a_i b_j$ and write $a^{\otimes 2} := a \otimes a$. Let $\nabla \otimes \nabla = \nabla^{\otimes 2}$ denote the Hessian operator. For random elements $(\xi_k)_{k\in\mathbb{N}}$ and ξ, we write $\xi_k \rightsquigarrow \xi$ to denote convergence in distribution; that is, $\xi_k \rightsquigarrow \xi$ if and only if for every bounded continuous function f, $\mathbb{E}\{f(\xi_k)\} \to \mathbb{E}\{f(\xi)\}$ as $k \to \infty$. We write $\mathcal{L}(\xi)$ for the distribution (law) of a random element ξ, and $\mathcal{L}^\nu(\xi)$ for the conditional distribution of ξ given another random element ν. For a square matrix M, define the symmetrization operator as $\mathrm{Sym}(M) := (M + M^\top)/2$. A square matrix M is *Hurwitz* (also called *stable*) if every eigenvalue of M has negative real part. For a function $f : \mathcal{A} \to L$ with $\mathcal{A}$ a set and $(L, \|\cdot\|)$ a normed linear space, define $\|f\|_\infty := \sup_{a\in\mathcal{A}} \|f(a)\|$.

4.2 Stochastic gradient algorithms and their scaling limits

In this section we develop a comprehensive framework that accurately predicts the large-sample asymptotics of stochastic gradient algorithms with fixed step sizes for inference and parameter estimation, including in cases where the model is misspecified. Our goal is to make their behaviour as methods for both optimization and sampling as transparent as possible – eliminating the ambiguity in the effects of hand-tuning the various parameters, except insofar as the user must determine the goal of the analysis and account for computing-related constraints.

4.2.1 A stochastic gradient meta-algorithm

We develop our methods and theory in the framework of a *stochastic gradient meta-algorithm*, with one-step update

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h^{(n)}\Gamma}{2}\hat{G}_k^{(n)} + \sqrt{\frac{h^{(n)}\Lambda}{\beta^{(n)}}}\,\xi_k, \tag{4.6}$$

where $\Gamma \in \mathbb{R}^{d\times d}$ is the (not necessarily positive semi-definite) gradient preconditioner, $\Lambda \in \mathbb{R}^{d\times d}$ is the (positive semi-definite) diffusion anisotropy matrix, ξ_k are i.i.d. $\mathrm{N}_d(0, I_d)$, and $\hat{G}_k^{(n)}$ implicitly depends on the batch size $b^{(n)}$ (which in turn may vary with the sample size n).

The meta-algorithm subsumes the SGD and SGLD algorithms discussed above. It also includes momentum-based methods, which can be seen by lifting the parameter space to a *phase space*; the details of this modification for underdamped stochastic Langevin dynamics are given in Section 4.4.1. This meta-algorithm does not include variants of stochastic gradient algorithms where the stochastic gradient is not of the form Eq. (4.1), such as the variance-reduction methods of Baker et al. [6], though we will sketch an extension of our results to that particular example in Section 4.4.2.

4.2.2 Scaling limit of the stochastic gradient meta-algorithm

For each $n \in \mathbb{N}$, let $\widehat{\theta}^{(n)}$ satisfy the first-order optimality condition $\sum_{i=1}^n \nabla\ell(\widehat{\theta}^{(n)}; X_i) = 0$ for the optimization problem of maximizing the likelihood function $\sum_{i=1}^n \ell(\theta; X_i)$. We aim to characterize the behaviour of the sample path of the iterates of Eq. (4.6) in the region about $\widehat{\theta}^{(n)}$. By so doing, we will be able to determine the limiting distribution of the iterate average (for optimization), the asymptotic stationary distribution of the iterates (for optimization and sampling), and the mixing speed (for sampling). Our approach is to obtain a functional central limit theorem by taking the scaling limit of the piecewise-constant, continuous-time process

$$\vartheta_t^{(n)} := w^{(n)} \left(\theta^{(n)}_{\lfloor \alpha^{(n)} t \rfloor} - \widehat{\theta}^{(n)} \right), \tag{4.7}$$

where $w^{(n)} \to \infty$ determines the spatial scaling and $\alpha^{(n)} \to \infty$ determines the temporal scaling.

Since it suffices for practical application, we assume polynomial scaling of all tuning parameters as a function of sample size: $h^{(n)} = c_h n^{-\mathfrak{h}}$ for $\mathfrak{h} > 0$, $b^{(n)} = \lfloor c_b n^{\mathfrak{b}} \rfloor$ for $\mathfrak{b} \geq 0$, and $\beta^{(n)} = c_\beta n^{\mathfrak{t}}$ for $\mathfrak{t} \in \mathbb{R}$. Given these tuning parameters, in order to have a stable and non-trivial[3] limit, we must take the time scaling to be $\alpha^{(n)} = n^{\mathfrak{h}}$ and the spatial scaling to be $w^{(n)} = n^{\mathfrak{w}}$ for $\mathfrak{w} = \min\{\mathfrak{b} + \mathfrak{h}, \mathfrak{t}\}/2$. In this setting we have the following result, under Assumptions 4.1 to 4.5 discussed below.

Theorem 4.1 (Scaling limit of the meta-algorithm)**.** *If Assumptions 4.1 to 4.5 all hold, and there exists $\theta_\star \in \Theta$ such that both $\widehat{\theta}^{(n)} \xrightarrow{p} \theta_\star$ and $\vartheta_0^{(n)} \rightsquigarrow \vartheta_0$, then*

$$(\vartheta_t^{(n)})_{t \in \mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t \in \mathbb{R}_+} \tag{4.8}$$

in the Skorohod topology[4] in probability, where $(\vartheta_t)_{t \in \mathbb{R}}$ is an Ornstein–Uhlenbeck process given by

$$d\vartheta_t = -\frac{1}{2} B \vartheta_t \, dt + \sqrt{A} \, dW_t, \tag{4.9}$$

with W_t a d-dimensional standard Brownian motion, drift matrix $B = c_h \Gamma \mathcal{J}_\star$, positive semi-definite diffusion matrix $A = \mathbb{I}_{[\mathfrak{b}+\mathfrak{h} \leq \mathfrak{t}]} \frac{c_h^2 \overline{c_b}}{4 c_b} \Gamma \mathcal{I}_\star \Gamma^\top + \mathbb{I}_{[\mathfrak{t} \leq \mathfrak{b}+\mathfrak{h}]} \frac{c_h}{c_\beta} \Lambda$, and batch constant

$$\overline{c_b} := \begin{cases} 1 - c_b & \mathfrak{b} = 1 \text{ and “no replacement”} \\ 1 & \text{otherwise.} \end{cases}$$

Assumptions 4.1 to 4.5 are fairly mild given the strength of the result.

Assumption 4.1. *$\nabla \log \pi^{(0)}$ is L_0-Lipschitz, and $\ell(\cdot; x) \in C^2(\Theta)$ for each $x \in \mathcal{X}$.*

[3]By non-trivial here, we mean that the limiting SDE should have both non-zero drift and non-zero diffusion terms if possible.

[4]See Section 4.6.3 for further discussion.

Assumption 4.2. $\mathfrak{h} - \mathfrak{w} - \mathfrak{a}/3 > 0$ *and* $\mathbb{E}\left[\|\nabla \ell(\theta_\star; X_1)\|^{p_2}\right] < \infty$ *for some* $p_2 > \frac{1}{\mathfrak{h}-\mathfrak{w}-\mathfrak{a}/3}$.

Assumption 4.3. *There exists* $q_3 \in [0, \mathfrak{w})$ *such that*

$$\|\widehat{\theta}^{(n)} - \theta_\star\| \in o_p(1/n^{q_3}), \quad \text{and} \quad \mathbb{E}\left[\left\|\nabla^{\otimes 2}\ell(\cdot; X_1)\right\|_\infty^{p_3}\right] < \infty, \tag{4.10}$$

where $p_3 = \frac{1}{\mathfrak{h}+q_3-\mathfrak{w}-\mathfrak{a}/3}$.

Assumption 4.4. *There is a nondecreasing sequence* $(r_{\mathcal{J},n})_{n\in\mathbb{N}}$ *with* $r_{\mathcal{J},n} \to \infty$, *such that*

$$\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\|\widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star)\right\| \xrightarrow{p} 0$$

Assumption 4.5. *There is a nondecreasing sequence* $(r_{\mathcal{I},n})_{n\in\mathbb{N}}$ *with* $r_{\mathcal{I},n} \to \infty$, *such that*

$$\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)} \left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}(\theta_\star)\right\| \xrightarrow{p} 0$$

Assumption 4.1 requires that the likelihood has a minimal number of continuous derivatives, and that the log-prior is smooth.[5] Assumption 4.2 ensures that the gradient value of the log-likelihood at the limiting parameter is not too volatile via a moment condition. Assumption 4.3 ensures that the random likelihood functions from each data sample are sufficiently smooth via a moment condition on the random smoothness parameter. Assumptions 4.4 and 4.5 require convergence of the empirical Fisher information matrices. The assumptions all hold, for example, for generalized linear models with bounded covariates and either Lipschitz inverse-link functions, or suitably constrained parameter domains (see Section 4.4.3 for the extension of the main result to constrained parameter spaces). Several sufficient conditions for each of Assumptions 4.4 and 4.5 are given in Section 4.9.

The proof of Theorem 4.1 is given in Section 4.7. With minor modifications it can be extended to the SGLD fixed point algorithm [e.g., 6], in which case $\overline{c_b} = 0$. For a discussion of this modification see Section 4.4.2.

[5] "Smoothness" here is being used in the optimization theory sense of the word, referring to Lipschitz gradients.

4.2.3 Theoretical implications of the scaling limit

Based on Theorem 4.1, we can to establish the following corollaries which we will further leverage to explain the empirical behaviour of stochastic gradient methods and to make recommendations for how these methods could be best tuned. First, we have a characterization of the marginal and (when it exists) the stationary covariance of the limiting process, including conditions under which simplified forms are possible.

Corollary 4.1. *In the setting of Theorem 4.1, the following hold:*

1. *For any initial parameter ϑ_0, the marginal covariance of the limiting process is*

$$Q_t := \mathrm{Cov}(\vartheta_t|\vartheta_0) = \int_0^t e^{-sB/2} A e^{-sB^\top/2} ds$$

and

$$\mathcal{L}^{\vartheta_0}(\vartheta_t) = \mathrm{N}\left(e^{-sB/2}\vartheta_0, Q_t\right).$$

2. *If $-\Gamma\mathcal{J}_\star$ is* Hurwitz, *then $Q_\infty := \lim_{t\to\infty} Q_t$ exists and the stationary distribution of $(\vartheta_t)_{t\in\mathbb{R}}$ is $\nu = \mathrm{N}_d(0, Q_\infty)$. In this case, Q_∞ solves the equation*

$$\frac{1}{2}BQ_\infty + \frac{1}{2}Q_\infty B^\top = A. \tag{4.11}$$

where, as before,

$$\begin{aligned} B &= c_h \Gamma \mathcal{J}_\star \\ A &= \mathbb{I}_{[\mathfrak{b}+\mathfrak{h}\le\mathfrak{t}]} \frac{c_h^2 \overline{c_b}}{4c_b} \Gamma \mathcal{I}_\star \Gamma^\top + \mathbb{I}_{[\mathfrak{t}\le\mathfrak{b}+\mathfrak{h}]} \frac{c_h}{c_\beta} \Lambda. \end{aligned}$$

The previous corollary leads to conditions for a Bernstein–von Mises-like result for the stationary distributions of the meta-algorithm.

Corollary 4.2 (Bernstein–von Mises-like theorem)**.** *In the setting of Theorem 4.1, if $((\vartheta_t^{(n)})_{t\in\mathbb{R}_+})_{n\in\mathbb{N}}$ has a sub-sequence with uniformly tight stationary measures, and if $-\Gamma\mathcal{J}_\star$ is* Hurwitz, *then the sub-sequence of stationary measures converges weakly to $\mathrm{N}_d(0, Q_\infty)$ in probability.*

4.2.4 Iterate Averages

Let $\bar{\theta}_k^{(n)} = \frac{1}{k}\sum_{j=1}^{k}\theta_j^{(n)}$ be the average of the first k iterations of the algorithm. The accuracy of the iterate average is characterized by its covariance $\bar{Q}_k^{(n)} := \mathrm{Cov}(\bar{\theta}_k^{(n)})$. We can approximate $\bar{Q}_k^{(n)}$ in terms of the covariance of the averaged limiting process, which is defined as $\bar{\vartheta}_t := t^{-1}\int_0^t \vartheta_s\, ds$.

Proposition 4.1. *For a stationary initial parameter $\vartheta_0 \sim \mathrm{N}(0, Q_\infty)$, the covariance of the averaged limiting process is*

$$\bar{Q}_t := \mathrm{Cov}\left(\bar{\vartheta}_t\right) = \frac{4}{t}\mathrm{Sym}\left((c_h\Gamma\mathcal{J}_\star)^{-1}Q_\infty\right) - \frac{8}{t^2}\mathrm{Sym}\left((c_h\Gamma\mathcal{J}_\star)^{-2}\left\{I - e^{-t(c_h\Gamma\mathcal{J}_\star)/2}\right\}Q_\infty\right). \tag{4.12}$$

The proof of this result is in Section 4.10. Using Eq. (4.11), the leading term has the explicit form

$$\frac{4}{t}\mathrm{Sym}\left((c_h\Gamma\mathcal{J}_\star)^{-1}Q_\infty\right) = \frac{1}{t}\left(\mathbb{I}_{[\mathfrak{b}+\mathfrak{h}\le\mathfrak{t}]}\frac{\overline{c_b}}{c_h c_b}\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} + \mathbb{I}_{[\mathfrak{t}\le\mathfrak{b}+\mathfrak{h}]}\frac{4}{c_\beta c_h}\mathcal{J}_\star^{-1}\Gamma^{-1}\Lambda(\Gamma^\top)^{-1}\mathcal{J}_\star^{-1}\right).$$

When either $\mathfrak{b}+\mathfrak{h} < \mathfrak{t}$, or $\mathfrak{b}+\mathfrak{h} = \mathfrak{t}$ and $c_\beta = +\infty$, then this simplifies to

$$\frac{4}{t}\mathrm{Sym}\left((c_h\Gamma\mathcal{J}_\star)^{-1}Q_\infty\right) = \frac{\overline{c_b}}{tc_b}\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}.$$

It is interesting, and perhaps initially surprising, that this is invariant to the choice of preconditioning matrix. Moreover, up to rescaling by $\overline{c_b}/(tc_b)$, this matches the covariance matrix of the asymptotic distribution of the MLE. However, it is perhaps less surprising in light of similar results for SGD with decreasing step-size due to Polyak and Juditsky [92].

4.3 Practical implications of the scaling limit

The characterization of the stochastic gradient meta-algorithm given by Theorem 4.1 and Corollaries 4.1 and 4.2 lets us answer fundamental questions about the large-sample properties of the stochastic gradient meta-algorithm:

1. When and how does mini-batch noise affect the algorithm?

2. When does the meta-algorithm sample from the posterior?

3. What other useful distributions can the meta-algorithm sample from?

4. What is the mixing time of the meta-algorithm?

5. What is the behaviour of the iterate averages?

We address each question in turn.

4.3.1 Effect of mini-batch noise

The mini-batch noise contributes in the large-sample regime when $\mathfrak{h} + \mathfrak{b} \leq \mathfrak{t}$. This exactly corresponds to when the mini-batch noise in a single step is on the same order (=) or dominates (<) the noise from the Gaussian innovations, ξ_k. We can interpret the phase transition as occurring because the variance of the mini-batch gradient scales as $n^{-2\mathfrak{h}-\mathfrak{b}}$ while the variance of update due to the Gaussian innovations scale as $n^{-\mathfrak{h}-\mathfrak{t}}$. The spatial scaling is chosen as $\mathfrak{w} = \min\{\mathfrak{b} + \mathfrak{h}, \mathfrak{t}\}/2$ to ensure that at least one of (a) the mini-batch noise or (b) the Gaussian innovations contribute to the limit, as otherwise the limit would be a gradient flow instead of a OU process, and hence fail to capture the asymptotically dominant local stochastic behaviour around $\widehat{\theta}^{(n)}$.

4.3.2 Sampling from the posterior

In order for the large-sample stationary distribution of Eq. (4.6) to match the Bernstein–von Mises limit of the posterior, we must first enforce that $\mathfrak{w} = 1/2$. Then, there are two ways to ensure that the limiting process has the same distribution as the limiting posterior. First we can chose our hyperparameters so that $\mathfrak{h} + \mathfrak{b} > \mathfrak{t}$. This will require us to set $\mathfrak{t} = 1$ to ensure $\mathfrak{w} = 1/2$. This condition can be interpreted as saying that combinations of mini-batch size and step size must yield mini-batch gradient variances that vanish fast enough to become negligible in the limit. In this case, selecting $\Gamma = \Lambda$ for any positive definite Λ, $c_\beta = 1$, and arbitrary values of c_h, c_b will suffice.

The second way in which we can match the posterior is by trying to precondition the mini-batch gradients so that the contribution of mini-batch noise to the limit is oriented exactly to give the correct variance. This, in turn can be achieved in two ways. First, if $\mathfrak{h} + \mathfrak{b} < \mathfrak{t}$ one may select Γ such that the smatrix Q_∞ that solves

$$\frac{1}{2}\Gamma \mathcal{J}_\star Q_\infty + \frac{1}{2} Q_\infty \mathcal{J}_\star^\top \Gamma^\top = \frac{c_h \overline{c_b}}{4 c_b} \Gamma \mathcal{I}_\star \Gamma^\top \tag{4.13}$$

is $Q_\infty = \mathcal{J}_\star$. As can be verified directly, and is essentially argued in Mandt et al. [70, Corollary 4], taking $\Gamma = \mathcal{I}_\star^{-1}$ and $c_h = \frac{4c_b}{\overline{c_b}}$, the limiting stationary measure will match the limiting posterior. Similarly, if $\mathfrak{h} + \mathfrak{b} = \mathfrak{t}$, then taking $\Lambda = \mathcal{I}_\star^{-1} = \Gamma$ and choosing c_h and c_β jointly so that $\frac{c_h \overline{c_b}}{4 c_b} + \frac{1}{c_\beta} = 1$, then the limiting stationary measure will match the limiting posterior.

In terms of the number of gradient queries per unit time in the scaling limit scales, the second way is more efficient as the query-count scales linearly with the dataset size ($\mathfrak{h} + \mathfrak{b} = 1$), while for the first way it scales super-linearly ($\mathfrak{h} + \mathfrak{b} > 1$).

4.3.3 Alternative uncertainty quantification

When our models are misspecified, however, the posterior distribution may provide less-than-robust uncertainty quantification [45]. In this case, we may want to either match the asymptotic covariance of the MLE, which by definition is robust to model misspecification, or match the asymptotic distribution of the bagged posterior, which combines aspects of both the asymptotics of the posterior and the MLE. Either of these desiderata can be obtained by setting $\Gamma = \Lambda = \mathcal{J}_\star^{-1}$, and any valid $\mathfrak{h} + \mathfrak{b} = 1 = \mathfrak{t}$. With this tuning, for any $v_1, v_2 > 0$, taking $c_h = 4 v_1 c_b$ and $c_\beta = v_2^{-1}$, gives

$$Q_\infty = v_1 \mathcal{J}_\star^{-1} \mathcal{I}_\star \mathcal{J}_\star^{-1} + v_2 \mathcal{J}_\star^{-1}.$$

This matches the asymptotic distribution of the bagged posterior with re-sampling rate v_1 when $v_1 = v_2$ [45]. Moreover, we can obtain any convex combinations of the uncertainty quantification from the posterior and from the asymptotics of the MLE, by taking $v_1 + v_2 = 1$. This would allow one to interpolate between frequentist-like and Bayesian-like forms of inference. We can also obtain the covariance of the MLE when sampling with replacement by using $\mathfrak{h} + \mathfrak{b} = 1$, $c_\beta = +\infty$, preconditioning with $\Gamma = \mathcal{J}_\star^{-1}$, and setting $c_h = 4c_b$.

4.3.4 Mixing time

Let $\hat{\nu}_k^{(n)}(f) := k^{-1}\sum_{k'=1}^{k} f(\theta_{k'}^{(n)})$ denote the Monte Carlo estimate of $\nu^{(n)}(f)$, where $\nu^{(n)}$ is that stationary measure of the stochastic gradient algorithm when the sample-size is n, if it exists. We can use the *mixing time* (or worst-case integrated autocorrelation time) $\tau^{(n)} := \sup_f \inf\{k : \operatorname{Var}_{\hat{\nu}_k^{(n)}}(f)/\operatorname{Var}_{\nu^{(n)}}(f) \leq 1\}$ to characterize the efficiency of MCMC algorithms. For the limiting process, define the "Monte Carlo average" $\hat{\nu}_t(f) := t^{-1}\int_0^t f(\vartheta_s)\,ds$ and the mixing time $\tau := \sup_f \inf\{t : \operatorname{Var}_{\hat{\nu}_t}(f)/\operatorname{Var}_\nu(f) \leq 1\}$. When the limiting process is reversible, standard results (e.g., applying the spectral theorem for self-adjoint operators [102] to the Poincaré inequality [8]) allow us to upper bound τ by the reciprocal of the spectral gap of the limiting process. Since the spectral gap of the Ornstein–Uhlenbeck process is $\lambda_{\min}(B)/2$, where $\lambda_{\min}(B)$ denotes its minimum eigenvalue of B, we may *heuristically* conclude then that the limiting mixing time is $\tau^{(n)} = 2\alpha^{(n)}/\lambda_{\min}(B)$ iterations. This mixing time corresponds to $2\alpha^{(n)}b^{(n)}/\lambda_{\min}(B)$ likelihood evaluations, or equivalently $2\alpha^{(n)}b^{(n)}/\{n\lambda_{\min}(B)\}$ dataset passes. Even when the limiting process is not reversible, the spectral gap is still a useful metric for the large-time rate of mixing of the process, and is given by the same formula, while the integrated autocorrelation time becomes intractable.

This is only a heuristic because, even if the process converge weakly and the stationary distributions converge weakly, it is insufficient to conclude that the mixing times converge. Instead the mixing time of limiting process corresponds to fixing a duration of scaled time for which to run the process, say T, then computing the limit of the covariance of an estimator based on the run up to time T, then letting T tend to infinity. The mixing time of the limit is of more practical relevance for our understanding of the local process since it accurately reflects the time needed for the limiting stationary distribution to provide a good approximation to a sample from the local process. On the other hand the limit of mixing times determines how long it would take to visit other modes if they exist, and would often tend to ∞ with sample size. This can be seen by considering a simple non-identifiable model, for example Gaussian location clustering, for which there would be two identical optimal solutions which differ only by permutations of the clusters. The limit of mixing times corresponds to the time it takes to explore both modes, while the mixing time of the limit corresponds to the time needed to explore the model closer to which the process is started. Even if there was not a second equally good mode, a second suboptimal mode that persists (though shrinking) at all sample sizes, and is moving farther away as the process is re-scaled, could lead to mixing times that do not converge.

In future work, we plan to introduce a more rigorous characterization of the correspondence between limit of mixing times and the mixing time of the limiting process. In particular, Atchadé [5] introduces the ζ-spectral gap, defined as

$$\mathrm{SpecGap}_\zeta := \inf\left\{\frac{\pi[f^2] - \langle f,\ Pf\rangle_{L^2(\pi)}}{\pi[f^2]-\zeta/2} \text{ s.t. } f\in L^2(\pi),\ \pi f = 0,\ \pi[f^2] > \zeta,\ \|f\|_{L^2(\pi)}\right\}. \tag{4.14}$$

We conjecture that for any $\zeta > 0$, under appropriate scaling (corresponding to the time rescaling factor $\alpha^{(n)}$), if the sequence of posterior distributions is tight, then the ζ-spectral gap will converge to that of the OU-process for all $\zeta > 0$. This is supported by the intuitive interpretation of the ζ-spectral gap; that it corresponds to the mixing time of the process within a local region containing most of the probability mass of the stationary distribution. Under the tightness assumption we expect that this is sufficient to rule out the types of pathological behaviour described in the previous paragraph.

4.3.5 Iterate averages

Let $m = kb^{(n)}/n$ denote the number of passes over the dataset (that is, the expected number of times each likelihood term is evaluated) by iteration k.

Corollary 4.3. *Fix a number of passes over the dataset, $m \in \mathbb{R}_+$. Suppose Assumptions 4.1 to 4.5 all hold. If $\mathfrak{b} + \mathfrak{h} \leq \mathfrak{t}$ and $((\vartheta_t^{(n)})_{t\in\mathbb{R}_+})_{n\in\mathbb{N}}$ have initial distributions converging weakly to ν, then the following hold:*

$$
\begin{aligned}
n\,\mathrm{Cov}\left(\bar{\theta}^{(n)}_{\lfloor mn/b^{(n)}\rfloor}\right) \to \frac{4c_b}{c_h m}\Big\{ & \mathbb{I}_{[\mathfrak{b}+\mathfrak{h}\le 1]}\mathrm{Sym}\left(\{\Gamma\mathcal{J}_\star\}^{-1} Q_\infty\right) \\
& - \mathbb{I}_{[\mathfrak{b}+\mathfrak{h}=1]}\frac{8c_b^2}{c_h^2 m^2}\mathrm{Sym}\left([\Gamma\mathcal{J}_\star]^{-2}\left[I - e^{-\frac{c_h m}{2c_b}\Gamma\mathcal{J}_\star}\right] Q_\infty\right)\Big\},
\end{aligned}
\tag{4.15}
$$

If in addition $\mathfrak{b}+\mathfrak{h} < \min(1,\mathfrak{t})$ *or* $c_\beta = +\infty$*, then*

$$
n\,\mathrm{Cov}\left(\bar{\theta}^{(n)}_{\lfloor mn/b^{(n)}\rfloor}\right) \to \frac{\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}}{m}. \tag{4.16}
$$

Proof. For Eq. (4.15), we have

$$
\begin{aligned}
\bar{Q}^{(n)}_k &= \mathrm{Cov}\left(\bar{\theta}^{(n)}_{\lfloor mn/b^{(n)}\rfloor}\right) \approx \frac{1}{(w^{(n)})^2}\,\mathrm{Cov}\left(\bar{\vartheta}_{mn/(b^{(n)}\alpha^{(n)})}\right) \\
&= \frac{4}{m}\frac{\alpha^{(n)}b^{(n)}}{n(w^{(n)})^2}\mathrm{Sym}\left(\{c_h\Gamma\mathcal{J}_\star\}^{-1} Q_\infty\right) \\
&\quad - \frac{8}{m^2}\frac{(\alpha^{(n)}b^{(n)})^2}{(nw^{(n)})^2}\mathrm{Sym}\left(\{c_h\Gamma\mathcal{J}_\star\}^{-2}\left\{I - \exp\left[-\frac{c_h mn}{2b^{(n)}\alpha^{(n)}}\Gamma\mathcal{J}_\star\right] Q_\infty\right\}\right).
\end{aligned}
$$

Now, given $\mathfrak{b}+\mathfrak{h} \le \mathfrak{t}$,

$$
\begin{aligned}
\lim_{n\to\infty} n\bar{Q}^{(n)}_k = & \frac{4c_b}{m}\mathrm{Sym}\left(\{c_h\Gamma\mathcal{J}_\star\}^{-1} Q_\infty\right) \\
& - \mathbb{I}_{[\mathfrak{b}+\mathfrak{h}=1]}\frac{8c_b^2}{m^2}\mathrm{Sym}\left([c_h\Gamma\mathcal{J}_\star]^{-2}\left[I - e^{-\frac{c_h m}{2c_b}\Gamma\mathcal{J}_\star}\right] Q_\infty\right)\Big\}
\end{aligned}
$$

The rest follows by combining this with Proposition 4.1 and the simplifications following it, and by noting that since $\mathfrak{h}+\mathfrak{b} \le 1$ and $\mathfrak{h} > 0$ we must have $\mathfrak{b} < 1$, and hence $\overline{c_b} = 1$. □

We are now positioned to characterize the rate at which Bernstein–von Mises-like limit for the paths of the general stochastic gradient algorithm concentrates, the asymptotic variance of the iterate average, and the mixing speed at stationarity. Observe that a phase change occurs at $\mathfrak{b}+\mathfrak{h} = 1$. When $\mathfrak{b}+\mathfrak{h} = 1$, the rate of concentration for the Bernstein–von Mises-like result is classical, $\mathfrak{w} = 1/2$, and the iterate average has smaller asymptotic variance while the underlying OU process also has a mixing time of order n likelihood evaluations. However, if $\mathfrak{b}+\mathfrak{h} > 1$, the process begins to behave more like a gradient flow and no longer mixes in a constant number of passes over the dataset, so the iterate average would converge more slowly (as measured by number of passes over the dataset) in that regime. If $\mathfrak{b} + \mathfrak{h} < 1$, the mixing time decreases, but is exactly offset by a slower Bernstein–von Mises-like concentration rate relative to when $\mathfrak{b} + \mathfrak{h} = 1$, overall yielding the same rate of concentration for the iterate averages as when $\mathfrak{b} + \mathfrak{h} = 1$.

4.4 Further applications and extensions

In this section we discuss applications and extensions of our scaling limit to more complex, practically relevant stochastic gradient algorithms. In particular, the poor approximation accuracy of SGLD with uninformed tunings has led to the proposal of many alternatives, including [85, 91, 120]. Of particular note are two approaches which are used to reduce the error of both stochastic optimization and sampling. First, momentum-based methods such as (stochastic) heavy ball [39] and underdamped (stochastic gradient) Langevin dynamics [3, 26, 57, 62, 68, 119] aim to improve on SGLD by improving the mixing time of the stochastic process being discretized, typically by moving to a non-reversible process which can in general mix faster than a reversible one. Second, variance reduction methods aim to improve the accuracy of the approximate posterior obtained by improving the stochastic estimates of the gradients used in the update formula at each step. For example [6, 80], does this with a clever choice of control variates. Lastly, in practice, often our parameter spaces are constrained, and we show that this does not affect the scaling limit.

4.4.1 Applications to momentum-based algorithms

Special cases of our results include momentum-based acceleration of SGD, for example, the quasi-hyperbolic momentum algorithm of Ma and Yarats [68], which includes many momentum-based algorithms as special cases, such as momentum algorithm, Nesterov's accelerated gradient, PID control algorithms [3], synthesized Nesterov variants [62], noise-robust momentum [26], triple momentum [119], least-squares acceleration of SGD [57]. See [68, Table 1] for more.

As an example, we show how we can express underdamped stochastic gradient Langevin dynamics in terms of our general stochastic gradient algorithm. We lift the parameter space to a *phase space* given by $\tilde{\Theta} = \Theta \times \mathbb{R}^d$, extend the log-likelihood to the phase space according to $\tilde{\ell}((\theta, \tilde{\theta}); x) = \ell(\theta; x) - \tilde{\theta}^\top M^{-1}\tilde{\theta}/2$, and lift the prior to phase space using the (improper) prior $\tilde{\pi}^{(0)}((\theta, \tilde{\theta})) = \pi^{(0)}(\theta)$. For (stochastic) heavy ball and underdamped (stochastic gradient) Langevin dynamics (cf., e.g., Duncan et al. [29, Eqs. 4 and 5]), the lifted Hamiltonian preconditioner $\tilde{\Gamma}$ and the lifted diffusion matrix $\tilde{\Lambda}$ (⋆) are:

$$\tilde{\Gamma} = \begin{bmatrix} 0 & -I \\ I & \Gamma \end{bmatrix} \quad \text{and} \quad \tilde{\Lambda} = \begin{bmatrix} 0 & 0 \\ 0 & \Gamma \end{bmatrix}.$$

This yields a combined parameter update formula of

$$\begin{aligned} \theta_{k+1}^{(n)} &= \theta_k^{(n)} + \frac{h^{(n)}}{2} M^{-1}\tilde{\theta}_k^{(n)} \\ \tilde{\theta}_{k+1}^{(n)} &= \left(I - \frac{h^{(n)}\Gamma}{2} M^{-1}\right)\tilde{\theta}_k^{(n)} + \frac{h^{(n)}}{2}\hat{G}_k^{(n)} + \sqrt{\frac{h^{(n)}}{\beta^{(n)}}\Gamma}\ \xi_k. \end{aligned} \tag{4.17}$$

4.4.2 Extension to control variates

SGLD Methods with control variates [6, 80] aim improve the reliability of SGLD as an MCMC method to reduce the variance caused by mini-batching by introducing a "zero variance control variate." This control variate is obtained by comparing the mini-batch gradient at the evaluated current parameter to the mini-batch gradient evaluated at the posterior mode (or MLE). Because this modification corresponds to a data-dependent change in the structure of the way stochastic gradients for the potential function are generated, this algorithm does not quite fit into the framework we have analyzed in the present work. However, the methods used herein to derive our scaling limit can be applied with modification to these control variate methods. In Section 4.11 we sketch such a result and its implications.

We find that the scaling limit for SGLD with control variates is nearly the same as without control variates, except that the diffusion term corresponding to mini-batch noise is always 0. This is because the average drift is (by design) not affected by the control variate, the additional Gaussian innovations have the same contribution as before, and the mini-batch noise is now always lower order. Because of this, the spatial scaling can always be chosen so that the noise from Gaussian innovations persists in the limit: that is, taking

$\mathfrak{w} = \mathfrak{t}/2$. Under this scaling, the corresponding limiting process the Ornstein–Uhlenbeck process:

$$d\vartheta_t = -\frac{1}{2}B\vartheta_t\,dt + \sqrt{A}\,dW_t, \tag{4.18}$$

with $B = c_h\Gamma\mathcal{J}_\star$ the drift matrix, $A = \frac{c_h}{c_\beta}\Lambda$ the positive semi-definite diffusion matrix, and W_t a d-dimensional standard Brownian motion.

4.4.3 Extension to constrained parameter spaces

If $\Theta \subsetneq \mathbb{R}^d$, then the iterations given by Eq. (4.6) may exit Θ, resulting in undefined behaviour. The typical way modify these algorithms to handle this case is to impose *boundary dynamics*. The two most common examples of such boundary dynamics for these are *reflecting* and *projecting*. Projecting maps iterates that would exit Θ to the nearest point within Θ. Reflecting, defined when the boundary is sufficiently smooth, treats the dynamics between two iterates as the motion of a particle in constant speed linear motion over a fixed time, and when the particle reaches the boundary it collides elastically and "bounces" off. In either case the new iterate is a measurable function of the previous iterate and the vector between the previous iterate what the new iterate would have been without adjusting for the constraint. Moreover, these conditions both satisfy that the distance between iterates is constrained by what the distance would have been without adjusting for the constraint. In this section we consider boundary dynamics satisfying a generalized version of this property.

Let $P : \Theta \times (\mathbb{R}^d)^3 \to \Theta$ be a measurable function such that:

(i) P is *faithful* to Θ, meaning that if $\mathrm{Conv}(\theta, \theta + \Delta_{\pi(0)} + \Delta_\ell + \Delta_\xi) \subset \Theta$ then

$$P(\theta, \Delta_{\pi(0)}, \Delta_\ell, \Delta_\xi) = \theta + \Delta_{\pi(0)} + \Delta_\ell + \Delta_\xi \tag{4.19}$$

where $\mathrm{Conv}(\theta_1, \theta_2)$ is the line segment from θ_1 to θ_2.

(ii) P is *local*, meaning that there exists $c_P > 0$ such that for all $(\theta, \Delta_{\pi(0)}, \Delta_\ell, \Delta_\xi) \in \Theta \times (\mathbb{R}^d)^3$

$$\|P(\theta, \Delta_{\pi(0)}, \Delta_\ell, \Delta_\xi) - \theta\| \le c_P \left(\|\Delta_{\pi(0)}\| + \|\Delta_\ell\| + \|\Delta_\xi\|\right). \tag{4.20}$$

We will consider the iterative algorithm on Θ given by

$$\theta_{k+1}^{(n)} = P\left(\theta_k^{(n)},\ \frac{h\Gamma}{2n}\nabla \log \pi^{(0)}\left(\theta_k^{(n)}\right),\ \frac{h\Gamma}{2}\frac{1}{b}\sum_{j\in[b]}\nabla \ell\left(\theta_k^{(n)};\ X_{I_k^{(n)}(j)}\right),\ \sqrt{h\beta^{-1}\Lambda}\ \xi_k\right). \tag{4.21}$$

When $\Theta \subsetneq \mathbb{R}^d$ and $\theta_\star \in \text{interior}(\Theta)$ the proof is essentially the same because the boundary dynamics are faithful and local. Intuitively, the assumption that $\vartheta^{(n)}(0) \rightsquigarrow \vartheta(0)$ ensures that the processes we consider all start near $\theta_\star$ and away from the boundary of Θ, and thus the spatial scaling drives the boundary of Θ outside any bounded set. This means that for any compactly supported test function f and any finite time $T > 0$ there is a minimal sample size n_0 large enough that the finite-sample-size process will not witness the boundary condition being activated by time T for sample sizes $n \geq n_0$. For more details see Section 4.12

4.5 Numerical Experiments

In this section we present the results of three experiments using both simulated and real data. We find that the theory we developed is closely reflected in the practical results.

4.5.1 Experiment 1: Gaussian simulation study

In this experiment, we demonstrates the effect of model misspecification. Exact specifications for the experiment are given in Table 4.1. The combination of true distribution and likelihood function was chosen specifically to ensure that $\mathcal{J}_\star \neq \mathcal{I}_\star$, so that the effect of misspecification would be apparent. We run SGD with no preconditioning, with preconditioning by $\mathcal{J}_\star$, and with preconditioning by $\mathcal{I}_\star$, and SGLD with preconditioning by $\mathcal{J}_\star$. We interpret this using our scaling limit with parameters $\mathfrak{w} = 1/2$, $\mathfrak{h} = 1$, $\mathfrak{b} = 0$. This combination of scaling parameters corresponds to the standard statistical local scaling, and a fixed batch size. For SGLD we also use $\mathfrak{t} = 0$ corresponding to a constant tempering. We present the results for this experiment using contour plots for the joint density of the first and last coordinates of the parameter vector. The density for the empirical run of the algorithms is given by a 2D kernel density estimate. The density for the predicted behaviour is given by the stationary distribution of the limiting process. As predicted by our results, preconditioning by $\mathcal{J}_\star$ leads to an empirical distribution for the iterates of the algorithm matching the covariance of the MLE, preconditioning by $\mathcal{I}_\star$ leads to an empirical distribution for the iterates of the algorithm matching the asymptotics of the posterior, and not preconditioning leads to behaviour that matches neither (but is still predictable using our results). Finally preconditioning by $\mathcal{J}_\star$ for SGLD leads to an empirical distribution for the iterates of the algorithm matching the asymptotics of a *bagged posterior*, which is given by a linear combination of the covariance of the MLE and the covariance of the posterior.

	Experiment 1	Experiment 2	Experiment 3
true distribution	$\mathrm{N}_{10}\left(0, \frac{1}{2}I + \frac{1}{2}\mathbf{11}'\right)$	unknown	unknown
log-likelihood $\ell(\cdot;\theta)$	$\sum_{i=1}^{10} \frac{(x_i-\theta_i)^2}{\sqrt{i}}$	$yx^\top\theta - \log(1+e^{x^\top\theta})$	$yx^\top\theta - \exp(x^\top\theta)$
log-prior $\log\pi^{(0)}(\theta)$	0	0	0
sample size n	1000	1000000	150000
batch size b	1	1000	1000
number of steps k	$10000n/b$	$1000n/b$	$1000n/b$
step size (SGD) h	$4b/n$	$4b/n$	$4b/n$
step size (SGLD) h	$2b/n$	b/n	$2b/n$
inv. temp. (SGLD) β	2	1	2

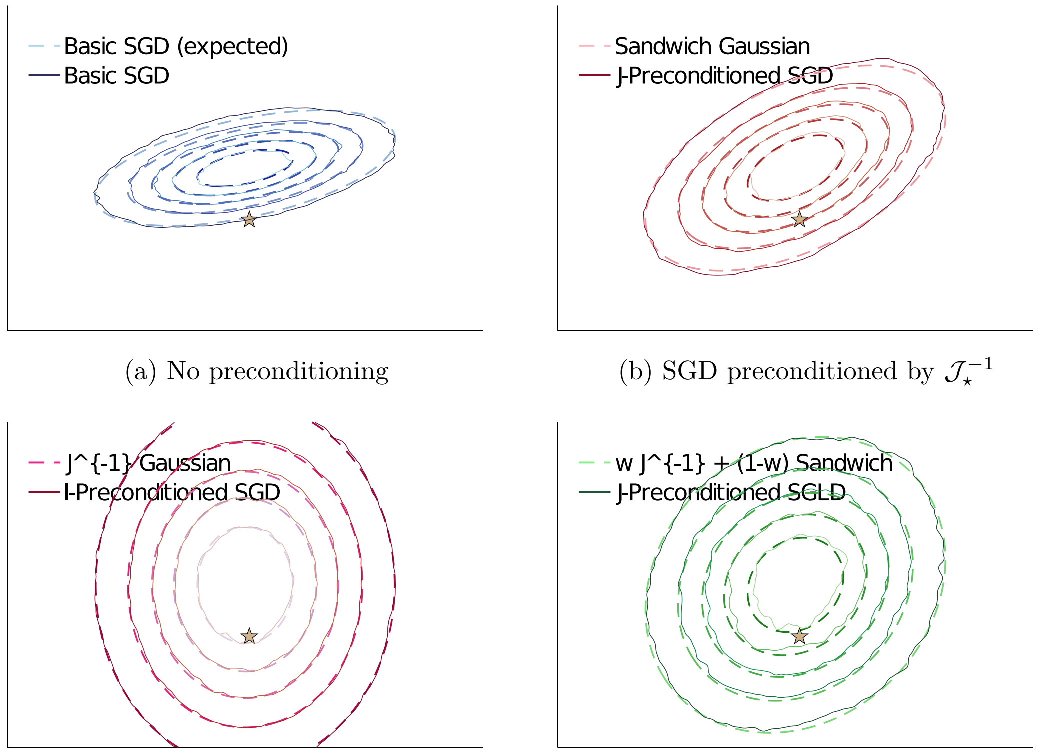
Table 4.1: Settings for experiments 1, 2, & 3. When the true distribution is unknown it is approximated by the empirical distribution on a larger version of the dataset for these experiments.

4.5.2 Experiment 2: Large-scale inference for airline delay data – logistic regression

In this experiment, we examine the same airline dataset and model as in Pollock et al. [91], using the pre-processed data they provided. The responses are binary and there are

Method	Empirical IACT	Predicted IACT
SGD, no preconditioning	3.2 epochs	3.2 epochs
$\mathcal{J}_\star^{-1}$-preconditioned SGD	1.1 epochs	1.0 epochs
$\mathcal{I}_\star^{-1}$-preconditioned SGD	2.3 epochs	2.8 epochs
$\mathcal{J}_\star^{-1}$-preconditioned SGLD	2.2 epochs	2.0 epochs

Table 4.2: Mixing times for experiment 1 as measured by integrated autocorrelation times (IACT). The empirical value is computed numerically from the run. The predicted value is computed based on the spectral gap of the limiting process.



(a) No preconditioning

(b) SGD preconditioned by $\mathcal{J}_\star^{-1}$

(c) SGD preconditioned by $\mathcal{I}_\star^{-1}$

(d) SGLD precond. by $\mathcal{J}_\star^{-1}$, $w = 1/2$

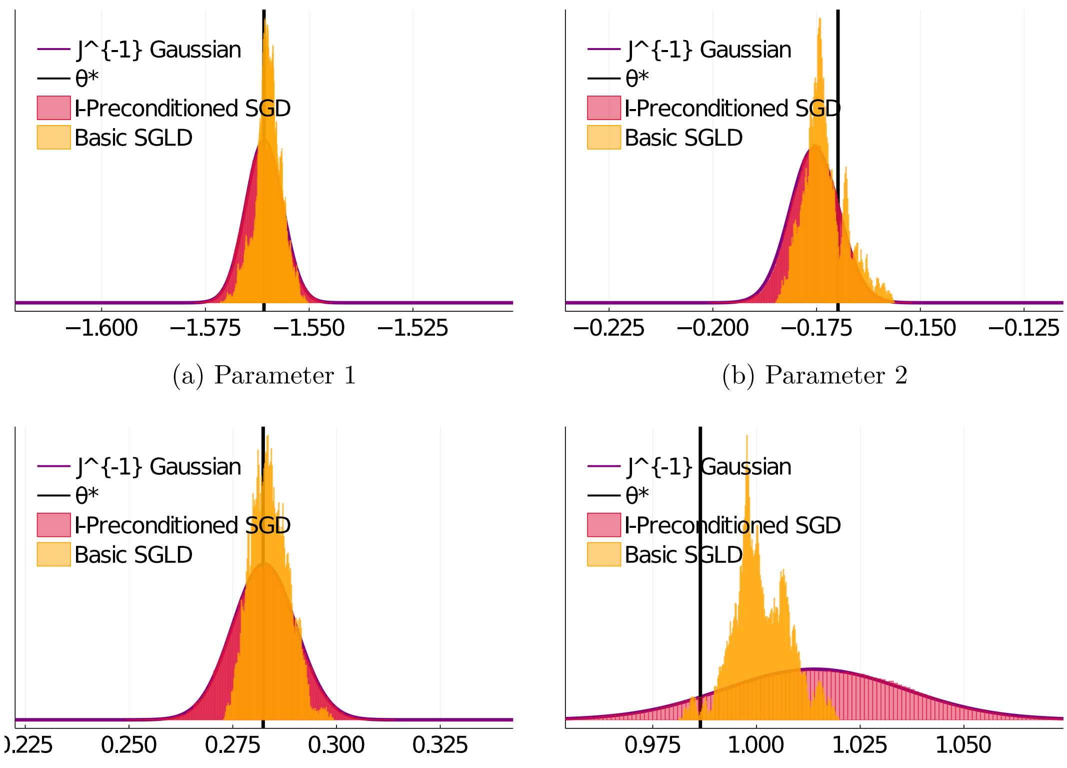
Figure 4.1: Results of experiment 1

3 covariates. We use the full dataset ($\approx$ 120 million observations) to estimate the "ground truth" quantities $(\theta_\star, \mathcal{J}_\star, \mathcal{I}_\star)$, and we apply the stochastic gradient algorithms using as a dataset a random subsample of size 1 million from the full dataset. In particular, we compare SGLD without preconditioning to SGD preconditioned by $\mathcal{I}_\star$. For this example, the matrices $\mathcal{J}_\star$ and $\mathcal{I}_\star$ are numerically indistinguishable, and hence all three preconditioned methods we examined in experiment 1 yield essentially identical results, and all are materially different from not preconditioning. Again, we interpret this using our scaling limit with parameters $\mathfrak{w} = 1/2$, $\mathfrak{h} = 1$, $\mathfrak{b} = 0$. An experimental finding of Pollock et al. [91] was that (non-preconditioned) SGLD had relatively poor mixing performance as compared with the ScaLE algorithm they introduce. Our experiment is consistent with their finding; we also find that without preconditioning, SGLD fails to properly quantify uncertainty in the true parameter (marginally for coordinate 4, and jointly) and mix slowly, which is not surprising since it was not properly tuned to. Furthermore, SGLD without preconditioning mixes materially

Method	Empirical IACT	Predicted IACT
SGD, no preconditioning	150 epochs	480 epochs
$\mathcal{J}_\star^{-1}$-preconditioned SGD	1.2 epochs	1.0 epochs
$\mathcal{I}_\star^{-1}$-preconditioned SGD	1.0 epochs	1.0 epochs
$\mathcal{J}_\star^{-1}$-preconditioned SGLD	2.3 epochs	2.0 epochs

Table 4.3: Mixing times for experiment 2 as measured by integrated autocorrelation times (IACT). The empirical value is computed numerically from the run. The predicted value is computed based on the spectral gap of the limiting process.

more slowly than preconditioned methods, as evidenced by the jagged histogram from its run, and the contour plot. However, we have shown that their findings would have been significantly different had they used the appropriate preconditioning as predicted by our theoretical results. Our experiments support the prediction made based on our theoretical results, that appropriate preconditioning accelerates the mixing of SGLD considerably and leads to more accurate uncertainty quantification.



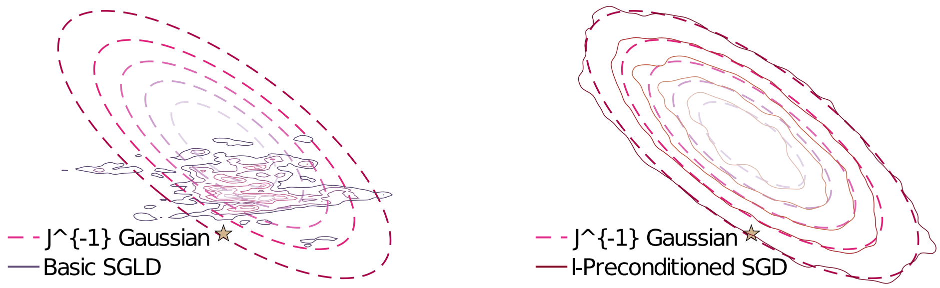
(a) Parameter 1

(b) Parameter 2

(c) Parameter 3

(d) Parameter 4

Figure 4.2: Univariate results of experiment 2



(a) SGLD without Preconditioning (b) SGD Preconditioned by $\mathcal{I}_\star$

Figure 4.3: Joint results of experiment 2: Parameters 1 and 4

4.5.3 Experiment 3: Large-scale inference for airline delay data – Poisson regression

In this experiment we examine the final year from the original airline dataset that the experiments in Pollock et al. [91] were based upon, in order to examine a more complex, more misspecified model on thematically similar data. In this case the responses are are non-negative integers and significantly 0-inflated (relative to a Poisson distribution), and we have opted not to model the zero-inflation to magnify the effect of misspecification. We use the full 2008 data ($\approx$ 1.5 million observations) to estimate the "ground truth" quantities $(\theta_\star, \mathcal{J}_\star, \mathcal{I}_\star)$, and we apply the stochastic gradient algorithms to a dataset generated as random subsample of size 150,000 from the full dataset. For this example, the matrices $\mathcal{J}_\star$ and $\mathcal{I}_\star$ differ significantly in scale, and hence all three preconditioned methods we examine yield materially different uncertainty quantification for the parameter. The non-preconditioned methods are numerically unstable at the comparable step-sizes, and quickly diverge. All three preconditioned methods behave exactly as predicted by the asymptotic theory, with the caveat that $\mathcal{I}_\star^{-1}$-preconditioned SGD mixes much slower than the $\mathcal{J}_\star^{-1}$-preconditioned methods in this example, and hence has not mixed as well as the $\mathcal{J}_\star^{-1}$-preconditioned methods for the number of epochs we have run. In this case, $\mathcal{I}_\star \approx r\mathcal{J}_\star$ for some $r \gg 1$, thus the faster mixing when preconditioning by $\mathcal{J}_\star^{-1}$ is to be expected since the spectral gap of the limiting process is roughly r times larger when preconditioning by $\mathcal{J}_\star^{-1}$ than when preconditioning by $\mathcal{I}_\star^{-1}$.

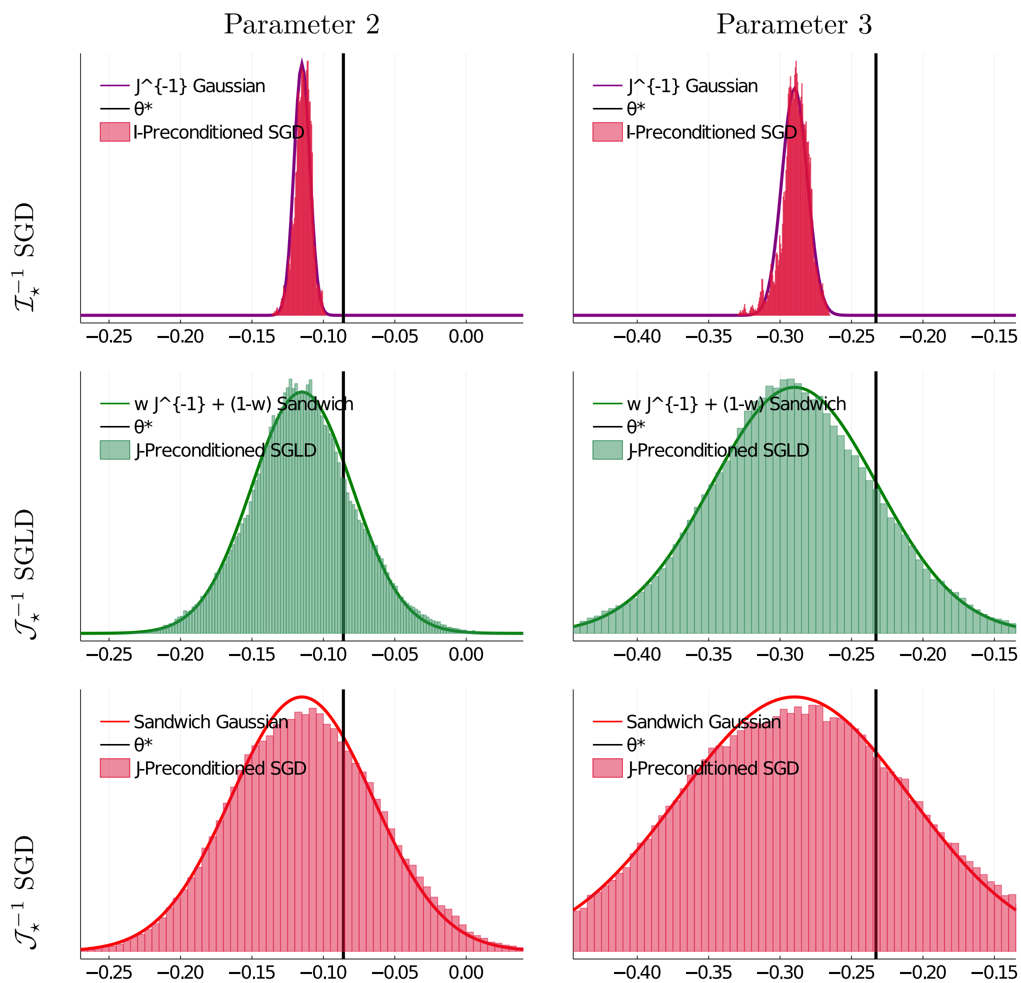


Figure 4.4: Univariate results of experiment 3

Method	Empirical IACT	Predicted IACT
$\mathcal{J}_\star^{-1}$-preconditioned SGD	1.1 epochs	1.0 epochs
$\mathcal{I}_\star^{-1}$-preconditioned SGD	130.0 epochs	98.0 epochs
$\mathcal{J}_\star^{-1}$-preconditioned SGLD	1.9 epochs	2.0 epochs

Table 4.4: Mixing times for experiment 3 as measured by integrated autocorrelation times (IACT). The empirical value is computed numerically from the run. The predicted value is computed based on the spectral gap of the limiting process.

4.6 Additional Definitions and Technical Results

Before presenting proofs of the various results of this work, we introduce some additional miscellaneous notations, definitions, and technical results that we will use.

4.6.1 Bernstein-von Mises under misspecification

Definition 4.1. *The first and second order* Fisher information matrices*, $\mathcal{I}$ and $\mathcal{J}$ respectively, are defined for a log-likelihood function ℓ and probability distribution P by*

$$\mathcal{I}(\theta) = \mathop{\mathbb{E}}_{X \sim P} \left[\nabla_\theta \ell(\theta; X) \otimes \nabla_\theta \ell(\theta; X)\right], \quad \textit{and} \quad \mathcal{J}(\theta) = -\mathop{\mathbb{E}}_{X \sim P} \nabla_\theta^{\otimes 2} \ell(\theta; X).$$

Let $\mathcal{X}$ be a Polish space with σ-field $\Sigma_\mathcal{X}$, $\mathcal{M}_{1,+}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$, and suppose that $P \in \mathcal{M}_{1,+}(\mathcal{X})$. Suppose that $\mathbf{X}^{(\mathbb{N})} := (X_i)_{i \in \mathbb{N}} \sim P^{\otimes \mathbb{N}}$. Let $n \in \mathbb{N}$ denote a sample size, let $[n] := \{1, \ldots, n\}$, and let $\mathbf{X}^{(n)} := (X_i)_{i \in [n]} \sim P^{\otimes n}$ be an I.I.D. sample of size n from P.

Let $\Theta \subseteq \mathbb{R}^d$ be open and nonempty, let Q be a regular conditional distribution from Θ to $(\mathcal{X}, \Sigma_\mathcal{X})$; i.e.:

(i) for all $\theta \in \Theta$, $Q_\theta \in \mathcal{M}_{1,+}(\mathcal{X})$, and

(ii) for all $A \in \Sigma_\mathcal{X}$, $Q_\cdot(A) : \theta \mapsto Q_\theta(A)$ is measurable[6].

Suppose there exists a σ-finite measure, μ, on $\mathcal{X}$, such that for all $\theta \in \Theta$, $Q_\theta \ll \mu$. Let q_θ denote a version of $\mathrm{d}Q_\theta / \mathrm{d}\mu$ for each $\theta \in \Theta$. Let $\ell(\theta; x) := \log q_\theta(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. We consider $\mathbb{M} := \{Q_\theta \mid \theta \in \Theta\}$ to be a *model* for P. The model is *well-specified* when $P \in \mathbb{M}$, and is *misspecified* otherwise. The *pseudo-true parameter* of the model is defined as $\theta_\star := \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{X \sim P} \ell(\theta; X)$. If $\mu \ll P$ then

$$\theta_\star = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{X \sim P} \ell(\theta; X) = \arg\min_{\theta \in \Theta} \mathrm{KL}\left(P \| Q_\theta\right).$$

[6] Θ is equipped with the Borel σ-field inherited from $\mathbb{R}^d$

Let $\Pi^{(0)} \in \mathcal{M}_{1,+}(\Theta)$ be any distribution on Θ. Let $\mathbb{P}_{\Pi^{(0)},\mathbb{M}} \in \mathcal{M}_{1,+}\left(\Theta \otimes \mathcal{X}^{\mathbb{N}}\right)$, given by

$$\mathbb{P}_{\Pi^{(0)},\mathbb{M}}(A \times B) := \int \mathbb{I}_{[\theta \in A]} \left[\int \mathbb{I}_{\left[x^{(\mathbb{N})} \in B\right]} {Q_\theta}^{\mathbb{N}}(\mathrm{d}x^{(\mathbb{N})}) \right] \Pi^{(0)}(\mathrm{d}\theta)$$

denote the joint distribution of the data and the parameter according to the model and the prior, where ${Q_\theta}^{\mathbb{N}}(\mathrm{d}x^{(\mathbb{N})})$ denotes the law of an I.I.D. sequence from Q_θ (an infinite product measure on the cylinder σ-field). Let $\mathbb{E}_{\Pi^{(0)},\mathbb{M}}$ denote the expectation under $\mathbb{P}_{\Pi^{(0)},\mathbb{M}}$. The posterior for θ under the model $\mathbb{M}$ given data $\mathbf{X}^{(n)}$ is the random probability measure on Θ given by

$$\Pi^{(\mathbf{X}^{(n)})}(A) := \mathbb{E}^{\mathbf{X}^{(n)}}_{\Pi^{(0)},\mathbb{M}} \left[\mathbb{I}_{[\theta \in A]} \right],$$

where for a random variable or σ-field G, an expectation operator $\mathbb{E}$ and a random variable Y, $\mathbb{E}^G(Y)$ is the conditional expectation of Y given G. The posterior $\Pi^{(\mathbf{X}^{(n)})}$ can be viewed as a probability kernel from $\mathcal{X}^n$ to Θ.

Let λ denote the Lebesgue measure. If $\Pi^{(0)} \ll \lambda$ with $\mathrm{d}\Pi^{(0)}/\mathrm{d}\lambda =: \pi^{(0)}$, then $\Pi^{(\mathbf{X}^{(n)})} \ll \lambda$ with $\mathrm{d}\Pi^{(\mathbf{X}^{(n)})}/\mathrm{d}\lambda = \pi^{(\mathbf{X}^{(n)})}$ given by

$$\pi^{(\mathbf{X}^{(n)})}(\theta) \propto \pi^{(0)}(\theta) \prod_{i \in [n]} q_\theta(X_i) = \pi^{(0)}(\theta) \exp\left(\sum_{i \in [n]} \ell(\theta; X_i) \right). \tag{4.22}$$

Let $\widehat{\theta}^{(n)} := \arg\max_{\theta \in \Theta} \sum_{i \in [n]} \ell(\theta; X_i)$ denote the maximum likelihood estimator (MLE) of $\theta_\star$ given the data $\mathbf{X}^{(n)}$. Posterior distributions have a general tendency to concentrate around the MLE as the sample size increases. Therefore, we will often reparameterize the model by considering a *local parametrization*, where to each parameter $\theta \in \Theta$ we associate a *local parameter*, $\vartheta \in \sqrt{n}\left(\Theta - \widehat{\theta}^{(n)}\right)$ based on the identification

$$\vartheta = \sqrt{n}\left(\theta - \widehat{\theta}^{(n)}\right)$$

and the *local model* is given by

$$\mathbb{M}^{(\mathbf{X}^{(n)})} := \left\{ Q_{\widehat{\theta}^{(n)} + \frac{1}{\sqrt{n}}\vartheta} \mid \vartheta \in \sqrt{n}\left(\Theta - \widehat{\theta}^{(n)}\right) \right\}.$$

The random localization map is given by

$$\mathrm{loc}_{\mathbf{X}^{(n)}} : \theta \mapsto \sqrt{n}\left(\theta - \widehat{\theta}^{(n)}\right)$$

For a measurable function $f : \mathcal{A} \to \mathcal{B}$ and a measure μ on $\mathcal{A}$, the *pushforward* of μ through f is the measure $f_\sharp\mu$ on $\mathcal{B}$ defined by $[f_\sharp\mu](B) = \mu(f^{-1}(B))$ for all measurable $B \subset \mathcal{B}$.

Proposition 4.2 (BvM under model misspecification, Kleijn and van der Vaart [58])**.** *Under regularity conditions,*

$$\left\| [\mathrm{loc}_{\mathbf{X}^{(n)}}]_\sharp \Pi^{(\mathbf{X}^{(n)})} - \Phi \right\|_{\mathrm{TV}} \xrightarrow{P} 0.$$

with $\theta_\star = \arg\max_{\theta\in\Theta} \mathop{\mathbb{E}}_{X\sim P} \ell(\theta; X)$*,* $\mathcal{J}_\star = -\mathop{\mathbb{E}}_{X\sim P}\left[\nabla^{\otimes 2}\ell(\theta_\star;\ X)\right]$*, and* $\Phi = \mathrm{N}\left(0, \mathcal{J}_\star^{-1}\right)$.

4.6.2 Convergence modes of measures and operators

Let $\mathcal{A}$ be a measurable space, and let $B(\mathcal{A})$ denote the collection of bounded measurable functions on $\mathcal{A}$. For a function $f : \mathcal{A} \to L$ with $(L, \|\cdot\|)$ a normed linear space, define

$$\|f\|_\infty := \sup_{a\in\mathcal{A}} \|f(a)\|.$$

For a sequence of probability measures, $\{\mu_n\}_{n\in\mathbb{N}}$ and a probability measure μ on a measurable space $\mathcal{A}$, we have the following modes of convergence:

- μ_n converges in *total variation* to μ, denoted by $\mu_n \xrightarrow{\mathrm{TV}} \mu$, *if and only if*

$$\sup_{f\in B(\mathcal{A})} \frac{|\mu_n f - \mu f|}{\|f\|_\infty} \to 0.$$

- if $\mathcal{A}$ is also a topological space and the σ-field on $\mathcal{A}$ is the Borel σ-field, then μ_n converges *in distribution* (also called *weakly*) to μ, denoted by $\mu_n \rightsquigarrow \mu$, *if and only if* for all $f \in \overline{C}(\mathcal{A})$, $|\mu_n f - \mu f| \to 0$.

Clearly

$$\left(\mu_n \overset{\mathrm{TV}}{\to} \mu\right) \implies \left(\mu_n \overset{\mathrm{s}}{\to} \mu\right) \implies (\mu_n \rightsquigarrow \mu)$$

while the converses do not hold in general.

For a Banach Space L with norm $\|\cdot\|$ denote its dual space (the space of all bounded linear operators on L) by L'. L' is a Banach space with norm $\|y\| := \sup_{x \in L \setminus \{0\}} |fx| \,/\, \|x\|$ for all $f \in L'$. Denote the set of bounded linear operators from L to itself by $\mathcal{B}(L)$. $\mathcal{B}(L)$ is also a Banach space with norm given by $\|T\| = \sup_{x \in L \setminus \{0\}} \|Tx\| \,/\, \|x\|$.

For a sequence of bounded linear operators, $\{T_n\}_{n \in \mathbb{N}}$, and a bounded linear operator, T, all mapping a Banach Space L to itself, we have the following modes of convergence:

- T_n converges *in norm* to T if and only if

$$\|T_n - T\| = \sup_{(x,y) \in L \times L'} \frac{|\langle y,\ (T_n - T)x \rangle|}{\|x\|\,\|y\|} \to 0 \tag{4.23}$$

- T_n converges *strongly* to T, denoted $T_n \overset{\mathrm{s}}{\to} T$ if and only if for all $x \in L$

$$\sup_{y \in L'} \frac{|\langle y,\ (T_n - T)x \rangle|}{\|y\|} \to 0 \tag{4.24}$$

Clearly

$$(\|T_n - T\| \to 0) \implies \left(T_n \overset{\mathrm{s}}{\to} T\right)$$

while the converse does not hold in general.

4.6.3 Operator Semigroups and Weak Convergence of Markov Processes

For a Banach space, $(L, \|\cdot\|)$, let $\mathcal{B}(L)$ denote the collection of all bounded linear operators from L to itself, and let I denote the identity operator. An *operator semigroup* on L is a function $T : \mathbb{R}_+ \to \mathcal{B}(L)$ such that

i) $T(0) = I$,

ii) $T(t + s) = T(t)T(s)$ for all $t, s \in \mathbb{R}$.

An operator semigroup is *strongly continuous* if

iii) $\lim_{t\to 0^+} \|T_t f - f\| = 0$ for all $f \in L$.

An operator semigroup is *contractive* if

iv) $\|T_t\| \leq 1$ for all $t \in \mathbb{R}_+$.

The *infinitessimal generator* (or just *generator*, for brevity) of the semigroup T is the (possibly unbounded) linear operator defined by

$$Af = \lim_{t\to 0^+} \frac{T_t f - f}{t}$$

for $f \in \mathrm{dom}(A) = \{f \in L \mid \lim_{t\to 0^+} (T_t f - f)/t \text{ exists}\}$. Let

$$\hat{C}(\mathbb{R}^d) = \left\{ f \in C(\mathbb{R}^d) \mid \forall \epsilon > 0 \; \exists K_{f,\epsilon} \subset \mathbb{R}^d \text{ compact with } \sup_{\theta \notin K_{f,\epsilon}} |f(\theta)| \leq \epsilon \right\}$$

Then $\hat{C}(\mathbb{R}^d)$ is a Banach space under the norm $\|f\|_\infty = \sup_{\theta\in\mathbb{R}^d} |f(\theta)|$. The dual space of $\hat{C}(\mathbb{R}^d)$ is the space of bounded signed measures under the *total variation norm*

$$\|\mu\|_{\mathrm{TV}} = \sup_{\substack{f\in\hat{C}(\mathbb{R}^d)\\ \|f\|_\infty \leq 1}} \left| \int f(\theta)\mu(d\theta) \right|.$$

We will work with $(L, \|\cdot\|) = \left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$. A semigroup on $\left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$ is *positive* if

v) $f \geq 0 \implies Tf \geq 0$.

A semigroup on $\left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$ is *Feller* if it is strongly continuous, contractive, and positive.

Semigroups naturally model the *forward operators* of Markov processes in continuous time. If X_t is a Markov process with transition kernels $k_t(\cdot,\cdot)$ then the forward operator corresponding to the Markov process (equivalently, corresponding to its transitio kernels) is defined by

$$T_t f(x) = \mathbb{E}_x f(X_t) = \int f(y) k_t(x, dy) \tag{4.25}$$

where $\mathbb{E}_x$ denotes expectation under the law of the Markov process given when $X(0) = x$ almost surely. The semigroup property is then equivalent to the Kolmogorov forward equation.

The generator, A, of a Feller semigroup T has a dense domain; $\mathrm{dom}(A)$ is dense in $\hat{C}(\mathbb{R}^d)$. A Markov process for which the corresponding forward operators form a a Feller semigroup is called a *Feller process*. Feller processes have a richly developed theory; see, for example, Ethier and Kurtz [33] or Kallenberg [53]. The following facts will be useful to us. First, every Feller process on $\mathbb{R}^d$ has a version with *càdlàg* (a.k.a *right continuous with left limits*, or *rcll*) paths, that is for all $t > 0$, $\lim_{s\to t^-} X(s)$ exists and $\lim_{s\to t^+} X_t$. Second for each $I \in \{[0,T] \mid T > 0\} \cup \{\mathbb{R}_+\}$, the collection of all càdlàg functions from I to $\mathbb{R}^d$ is a separable and complete metric space under the *Skorohod metric* [53, Theorem A2.2]. The formula for the Skorohod metric is not particularly illuminating, so is omitted here and may be found in the reference. This space is denoted by $D(I, \mathbb{R}^d)$. The Borel σ-field generated by the Skorohod metric is equal to $\sigma(\{\pi_t \mid t \in I'\})$ where $\pi_t(X) = X_t$ are the projection maps, and I' is any dense subset of I.

Let $C_c^\infty(\mathbb{R}^d)$ be the set of functions $\mathbb{R}^d \to \mathbb{R}$ with compact support and with continuous derivatives of all orders. $C_c^\infty(\mathbb{R}^d)$ is dense in $\overline{C}(\mathbb{R}^d)$.

Proposition 4.3 (Approximation of Markov Chains (compiled from Ethier and Kurtz [33])**.** *Let $A : C_c^\infty(\mathbb{R}^d) \to \overline{C}(\mathbb{R}^d)$ be linear and suppose that the closure of the graph of A (with respect to the graph norm defined by $\|f\|_A = \|f\|_\infty + \|Af\|_\infty$ for all $f \in L$) generates a Feller semigroup T on $\mathbb{R}^d$. Let $(\vartheta_t)_{t\in\mathbb{R}_+}$ be a Markov process with forward operator semigroup T. Let $\left((\theta_k^{(n)})_{k\in\mathbb{N}\cup\{0\}}\right)_{n\in\mathbb{N}}$ be a sequence of (discrete-time) Markov chains on $\mathbb{R}^d$ with respective transition kernels $(U^{(n)})_{n\in\mathbb{N}}$. Suppose that $0 < \alpha^{(n)} \to \infty$, and let*

$$A^{(n)} = \alpha^{(n)}\left(U^{(n)} - I\right) \qquad T_t^{(n)} = \left(U^{(n)}\right)^{\lfloor \alpha^{(n)}\ t\rfloor} \qquad \vartheta_t^{(n)} = \theta^{(n)}_{\lfloor \alpha^{(n)}\ t\rfloor}.$$

If $\left\|A^{(n)}f - Af\right\|_\infty \to 0$ for all $f \in C_c^\infty(\mathbb{R}^d)$, then

(a) $T_t^{(n)} \xrightarrow{s} T_t$ for each $t > 0$, and

(b) If $\vartheta^{(n)}(0) \rightsquigarrow \vartheta(0)$ then $\vartheta^{(n)}(\cdot) \rightsquigarrow \vartheta(\cdot)$ in the Skorohod metric.

Proof of Proposition 4.3. (a) Follows from Chapter 1, Theorem 6.5 of Ethier and Kurtz [33]. (b) Follows by combining Chapter 4, Theorem 8.2, Corollary 8.5, and Corollary 8.9 of Ethier and Kurtz [33]. □

4.6.4 Miscellaneous notation and definitions

Definition 4.2 (Convergence in Probability to a constant)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(\mathcal{X}, \tau)$ be a topological space endowed with the σ-field $\mathcal{F}_{\mathcal{X}} = \sigma(\tau)$, let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{X}$-valued random elements, and let $x \in \mathcal{X}$. Then X_n* converges to x in probability as $n \to \infty$*, denoted $X_n \xrightarrow{p} x$, when for every neighbourhood $x \in U \in \tau$ we have*

$$\lim_{n\to\infty} \mathbb{P}(X_n \in U^c) = 0.$$

Lemma 4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $(\mathcal{X}, \tau)$ be a topological space endowed with the σ-field $\mathcal{F}_{\mathcal{X}} = \sigma(\tau)$, let $(X_n)_{n\in\mathbb{N}}$ be a sequence of $\mathcal{X}$-valued random elements, and let $x \in \mathcal{X}$.*

If for every sub-sequence n_m there is a sub-sub-sequence n_{m_k} such that $X_{n_{m_k}} \to x$ almost surely as $k \to \infty$ then $X_n \xrightarrow{p} x$.

If $(\mathcal{X}, \tau)$ is first-countable then the converse also holds; if $X_n \xrightarrow{p} x$ then for every sub-sequence n_m there is a sub-sub-sequence n_{m_k} such that $X_{n_{m_k}} \to x$ almost surely as $k \to \infty$.

The proof of this result is the same as in Durrett [32, Theorem 2.3.2], generalizing the metric space definition of convergence in probability and replacing a sequence of balls of vanishing radius with a countable neighbourhood basis.

4.7 Proof of Theorem 4.1

In this section we prove Theorem 4.1, as well as an additional result along with what was stated, since both follow from the same premises. The full statement of what we prove is given below. Item 2 below is used in the proof of Corollary 4.2.

Theorem 4.2 (Scaling Limits of SGD/SGLD/LD (Full))**.** *Suppose that $(\theta_k^{(n)})_{k\in\mathbb{N}}$ evolves according to the gradient-based algorithm in Eq. (4.21) with step-size $h^{(n)} = c_h n^{-\mathfrak{h}}$, $b^{(n)} = \left\lfloor c_b n^{\mathfrak{b}} \right\rfloor$, $\beta^{(n)} = c_\beta n^{\mathfrak{t}}$, all other tuning parameters constant in n. Let $\theta_\star \in \mathbb{R}^d$. Let $\mathbf{X}^{(\mathbb{N})} = (X_i)_{i\in\mathbb{N}} \sim P^{\otimes\mathbb{N}}$, and $\widehat{\theta}^{(n)}$ be a critical point of the log-likelihood function $\sum_{i=1}^n \ell(\cdot, X_i)$ for each $n \in \mathbb{N}$; that is $\sum_{i=1}^n \nabla\ell(\widehat{\theta}^{(n)}, X_i) = 0$ for all $n \in \mathbb{N}$.*

Let $\vartheta_t^{(n)} = w^{(n)}\left(\theta^{(n)}_{\lfloor \alpha(n)t\rfloor} - \widehat{\theta}^{(n)}\right)$, *where* $w^{(n)} = n^{\mathfrak{w}}$, $\alpha^{(n)} = n^{\mathfrak{a}}$, $\mathfrak{w} \in (0,1)$,

$$\mathfrak{a} = \min\left\{\mathfrak{h},\ (\mathfrak{t}+\mathfrak{h}-2\mathfrak{w}),\ (\mathfrak{b}+2\mathfrak{h}-2\mathfrak{w})\right\}.$$

If Assumptions 4.1 to 4.5 all hold, $\mathfrak{a} > 0$, *and* $\vartheta^{(n)}(0) \rightsquigarrow \vartheta(0)$ *then*

1. $(\vartheta_t^{(n)})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ *in the Skorohod topology in probability, where* $(\vartheta_t)_{t\in\mathbb{R}}$ *follows the Ornstein–Uhlenbeck process:*

$$d\vartheta_t = -\frac{c_{\mathfrak{d}}}{2}\Gamma\mathcal{J}(\theta_\star)\vartheta_t dt + \sqrt{c_{\mathfrak{g}}\Lambda + c_{\mathfrak{mb}}\Gamma\mathcal{I}(\theta_\star)\Gamma'}\ dW_t,$$

with

$$c_{\mathfrak{d}} = \begin{cases} c_h & \mathfrak{a} = \mathfrak{h} \\ 0 & \mathfrak{a} < \mathfrak{h} \end{cases}, \qquad c_{\mathfrak{g}} = \begin{cases} \frac{c_h}{c_\beta} & \mathfrak{a} = \mathfrak{h}+\mathfrak{t}-2\mathfrak{w} \\ 0 & \mathfrak{a} < \mathfrak{h}+\mathfrak{t}-2\mathfrak{w} \end{cases}$$

and

$$c_{\mathfrak{mb}} = \begin{cases} \frac{c_h^2(1-c_b)}{4c_b} & \mathfrak{a} = 1+2\mathfrak{h}-2\mathfrak{w} \text{ and } \mathfrak{b}=1 \text{ and no replacement} \\ \frac{c_h^2}{4c_b} & \mathfrak{a} = \mathfrak{b}+2\mathfrak{h}-2\mathfrak{w} \text{ and } (\mathfrak{b}\neq 1 \text{ or replacement}) \\ 0 & \mathfrak{a} < \mathfrak{b}+2\mathfrak{h}-2\mathfrak{w}. \end{cases}$$

2. *If* $T^{(n)}$ *and* T *are defined as in Proposition 4.3, then under the conditions above, every subsequence of* $\left(T^{(n)}\right)_{n\in\mathbb{N}}$, $\left(T^{(n_m)}\right)_{m\in\mathbb{N}}$, *has a further sub-subsequence,* $\left(T^{(n_{m_k})}\right)_{k\in\mathbb{N}}$, *such that with probability* 1, $T_t^{(n_{m_k})} \xrightarrow{s} T_t$ *for all* $t > 0$.

Before beginning the proof of this result, Theorem 4.2, we require the following lemma, which is used to turn the moment conditions in our assumptions into bounds on the magnitudes of certain random variables that hold all but finitely often with probability 1.

Lemma 4.2. *Let* $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ *be non-decreasing, right continuous with left limits, with* $\alpha(0) = 0$, *and* $\lim_{t\to\infty}\alpha_t = \infty$. *Let* $Z_i \sim \mu$ *for all* $i \in \mathbb{N}$ *(possibly not independent) with* $Z_1 \geq 0$ *almost surely such that* $\mathbb{E}[\alpha(Z_1)] < \infty$. *Let* $\alpha^+ : u \mapsto \inf\{t \geq 0 \text{ s.t. } \alpha_t \geq u\}$ *be the generalized inverse of* α. *Then*

$$\mathbb{P}\left(\max_{i\in[n]} Z_i \geq \alpha^+(n) \quad \text{i.o.}\right) = 0.$$

Proof of Lemma 4.2. Let $S_t = \mathbb{P}(Z_1 > t)$ be the survival function of μ, and let $W_n = \alpha(Z_n)$ for each $n \in \mathbb{N}$. Note that $\mathbb{P}(W_1 > t) = S(\alpha_t^+)$. Then

$$\infty > \mathbb{E}\left[(\alpha(Z_1))\right] = \int_0^\infty \mathbb{P}(W_1 > t)dt \geq \sum_{n=1}^\infty \mathbb{P}(W_1 > n) = \sum_{n=1}^\infty \mathbb{P}(W_n > n)$$

Therefore, from the Borel–Cantelli lemma $\mathbb{P}(W_n > n \quad \text{i.o.}) = 0$, and equivalently $\mathbb{P}(W_n \leq n \quad \text{a.b.f.o.}) = 1$. Now, whenever $W_n \leq n$ for all but finitely many n, then there exists $K \in \mathbb{N}$ and $I_1, \dots I_K \in \mathbb{N}$ with $W_n \leq n$ for all $n \in \mathbb{N} \setminus \{I_j : j \in [K]\}$. Therefore, for all $n \geq \max_{j \leq K} W_{I_j}$, $\max_{i \leq n} W_i \leq n$. Therefore $\mathbb{P}(\max_{i \leq n} W_i \leq n \quad \text{a.b.f.o.}) = 1$, and equivalently $\mathbb{P}(\max_{i \leq n} W_i > n \quad \text{i.o.}) = 0$. Finally, $W_i > n$ if and only if $Z_i > \alpha^+(n)$, hence

$$\mathbb{P}(\max_{i \leq n} Z_i > \alpha^+(n) \quad \text{i.o.}) = 0.$$

$\square$

4.7.1 Proof of Theorem 4.2

Let $\mathcal{J}_\star = \mathcal{J}(\theta_\star)$ and $\mathcal{I}_\star = \mathcal{I}(\theta_\star)$.

The proof proceeds in the following stages. In Section 4.7.1.1, we will reduce the problem of weak convergence in the Skorohod topology in probability to one of weak convergence in the Skorohod topology almost-surly along subsequences and construct appropriate such subsequences. In Section 4.7.1.2 we introduce notation that will be useful in the remainder of the proof. In Section 4.7.1.3 we discuss what is needed to apply Proposition 4.3 to establish the processes converge weakly in the Skorohod topology almost-surely. This amounts to showing that the difference between the approximate generator and limiting generator evaluated a smooth test function with compact support vanishes uniformly. We will examine this difference in two regimes. First, in Section 4.7.1.4, we will consider arguments sufficiently far from the support of the test function. Then, in Section 4.7.1.5, we will consider arguments in or close to the support of the test function, and use a Taylor series expansion of the approximate generator to divide this into three types of non-zero terms. The first type is non-remainder terms that vanish and have no corresponding term in the limiting generator; these are handled in Section 4.7.1.6. The second type is terms that do not vanish and do have corresponding terms in the limiting generator; these are handled in Sections 4.7.1.7 to 4.7.1.9. The third type of term is the remainder term, which is handled in Section 4.7.1.10. Putting all of this together allows us to apply Proposition 4.3 along our subsequences, establishing the main result.

4.7.1.1 Reduction to almost-sure convergence on subsequences

Let

$$\begin{aligned}
\Upsilon^{(n)} &= \max\left(\Upsilon_1^{(n)}, \Upsilon_2^{(n)}, \Upsilon_3^{(n)}\right), \\
\Upsilon_1^{(n)} &= n^{q_3}\left\|\widehat{\theta}^{(n)} - \theta_\star\right\|, \\
\Upsilon_2^{(n)} &= \sup_{\theta\in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star)\right\|, \\
\Upsilon_3^{(n)} &= \sup_{\theta\in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}(\theta_\star)\right\|.
\end{aligned}$$

Each of the Υ terms corresponds to the important quantity that vanishes in probability for one of the assumptions. For example, $\Upsilon_1^{(n)}$ controls how quickly the local MLE converges under Assumption 4.2 which lets us use a weaker moment assumption for the sup-norm of the Hessian of the log-likelihood.

By assumption, $\Upsilon^{(n)} \xrightarrow{\text{p}} 0$. Then, by Lemma 4.1, for every subsequence $(n_m)_{m\in\mathbb{N}}$ there is a further sub-subsequence $(n_{m_k})_{k\in\mathbb{N}}$ so that this convergence is almost sure. Along an arbitrary such sub-subsequence, we will verify that $(\vartheta^{(n_{m_k})})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ in the Skorohod topology almost surely. Since weak convergence is metrizable (e.g., by the Levi–Prokhorov metric, and hence corresponds to a topology on probability distributions), and since for any subsequence $(n_m)_{m\in\mathbb{N}}$ we will have shown a further subsequence $(n_{m_k})_{k\in\mathbb{N}}$ such that $(\vartheta_t^{(n_{m_k})})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ a.s., by Lemma 4.1 it must hold that $(\vartheta_t^{(n)})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ in probability.

Now, let $(n_m)_{m\in\mathbb{N}}$ be an arbitrary subsequence[7] of $\mathbb{N}$ such that $\Upsilon^{(n_m)} \xrightarrow{\text{a.s.}} 0$. Let Ω

[7]Since every sub-subsequence is itself a subsequence, we can simplify our notation from here onward.

denote the underlying probability space. Let

$$\begin{aligned}
\Omega^{(0)} &= \bigcap_{i=1}^{3} \Omega^{(i)}, \\
\Omega^{(1)} &= \left\{ \Upsilon^{(n_m)} \to 0 \right\}, \\
\Omega^{(2)} &= \left\{ \max_{i \in [n]} \|\nabla \ell(\theta_\star; X_i)\| \leq n^{1/p_2} \quad \text{a.b.f.o} \right\} , \\
\Omega^{(3)} &= \left\{ \max_{i \in [n]} \left\| \nabla^{\otimes 2} \ell(\cdot; X_i) \right\|_\infty \leq n^{1/p_3} \quad \text{a.b.f.o.} \right\} .
\end{aligned}$$

By assumption, and by applying Lemma 4.2 to power functions of the form $\alpha : t \mapsto t^p$ and random variables $\|\nabla \ell(\theta_\star; X_i)\|$ and $\left\|\nabla^{\otimes 2} \ell(\cdot; X_i)\right\|_\infty$, $\Omega^{(0)}$ is a sure set.

4.7.1.2 Additional notation used in the proof

We notate the increments of the localized iterative algorithms (given that $\vartheta_0^{(n)} = \vartheta$) due to the Gaussian innovation (ξ), the gradient step contribution of the prior ($\pi^{(0)}$), the mini-batch gradient step based on the log-likelihood (ℓ), and the total increment, respectively, as

$$\begin{aligned}
\Delta_\xi^{(n)} &:= w^{(n)} \sqrt{h \beta^{-1} \Lambda}\; \xi_1, \\
\Delta_{\pi^{(0)}}^{(n)}(\vartheta) &:= \frac{h w^{(n)} \Gamma}{2n} \nabla \log \pi^{(0)} \left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1} \vartheta \right), \\
\Delta_\ell^{(n)}(\vartheta) &:= \frac{h w^{(n)} \Gamma}{2 b^{(n)}} \sum_{j \in \left[b^{(n)}\right]} \nabla \ell \left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1} \vartheta;\; X_{I_1^{(n)}(j)} \right), \text{ and} \\
\Delta^{(n)}(\vartheta) &:= \Delta_\xi^{(n)} + \Delta_{\pi^{(0)}}^{(n)}(\vartheta) + \Delta_\ell^{(n)}(\vartheta).
\end{aligned}$$

We define the sequence of operators $A^{(n)}$ by

$$[A^{(n)} f](\vartheta) = \alpha^{(n)} \left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[f(\vartheta + \Delta^{(n)}(\vartheta)) \right] - f(\vartheta) \right). \tag{4.26}$$

for all $n \in \mathbb{N}$, and all $f \in C_c^\infty(\mathbb{R}^d)$, where $\alpha^{(n)} = n$. The generator of the (presumed, at this point) limiting OU process is given by

$$[Af](\vartheta) = -\left\langle \frac{c_{\mathtt{d}}}{2} \Gamma \mathcal{J}_\star \vartheta,\ \nabla f(\vartheta) \right\rangle + \frac{1}{2} \left(c_{\mathtt{g}} \Lambda + c_{\mathtt{mb}} \Gamma \mathcal{I}_\star \Gamma' \right) : \nabla^{\otimes 2} f(\vartheta) \tag{4.27}$$

4.7.1.3 How Proposition 4.3 is applied

Consider a single realization of $\mathbf{X}^{(\mathbb{N})} \in \Omega^{(0)}$. Our goal, now, is to apply Proposition 4.3, treating $\mathbf{X}^{(\mathbb{N})}$ as fixed. To do so, it suffices to show that for each $f \in C_c^\infty(\mathbb{R}^d)$ we have

$$\lim_{m\to\infty} \sup_{\vartheta\in\mathbb{R}^d} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

For an arbitrary test function, $f \in C_c^\infty(\mathbb{R}^d)$, with compact support K_0, we will show this in two parts. First we will identify a compact extension, $K_1 \supset K_0$ to the compact support of f such that

$$\lim_{m\to\infty} \sup_{\vartheta\in K_1^c} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

Then we will separately show that

$$\lim_{m\to\infty} \sup_{\vartheta\in K_1} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

4.7.1.4 Convergence away from the test function support

For all $\vartheta \in K_0^c$, $f(\vartheta) = 0$, $\nabla f(\vartheta) = 0$, and $\nabla^{\otimes 2} f(\vartheta) = 0$. Therefore, for any $K_1 \supset K_0$,

$$\begin{aligned}
&\sup_{\vartheta\in K_1^c} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| \\
&\quad \leq \alpha^{(n_m)} \left\|f\right\|_\infty \sup_{\vartheta\in K_1^c} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left[\vartheta + \Delta^{(n_m)}(\vartheta) \in K_0\right].
\end{aligned} \tag{4.28}$$

Let $R_0 = \sup_{\vartheta\in K_0}\|\vartheta\|$. Let $K_1 = \left\{\vartheta \in \mathbb{R}^d \text{ s.t. } \|\vartheta\| \leq 2R_0 + 2c_0\right\}$, where

$$c_0 = \frac{c_h\left\|\Gamma\right\|}{2}\left(3 + \left\|\nabla \log \pi^{(0)}\left(\theta_\star\right)\right\|\right) + \sqrt{c_h/c_\beta\left\|\Lambda\right\|}.$$

Then, using Eq. (4.28) and $\Delta^{(n_m)}(\vartheta) = \Delta_\xi^{(n_m)}(\vartheta) + \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) + \Delta_\ell^{(n_m)}(\vartheta)$,

$$\begin{aligned}
&\sup_{\vartheta \in K_1^c} \left| [A^{(n_m)} f](\vartheta) - [Af](\vartheta) \right| \\
&\quad \leq \alpha^{(n_m)} \|f\|_\infty \sup_{\|\vartheta\| > 2R_0 + 2c_0} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}} \left[\left\| \vartheta + \Delta^{(n_m)}(\vartheta) \right\| \leq R_0 \right] \\
&\quad \leq \alpha^{(n_m)} \|f\|_\infty \sup_{\|\vartheta\| > 2R_0 + 2c_0} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}} \left[\left\| \Delta_\xi^{(n_m)} \right\| \geq \|\vartheta\| - \left\| \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\| - \left\| \Delta_\ell^{(n_m)}(\vartheta) \right\| - R_0 \right].
\end{aligned} \tag{4.29}$$

For $\vartheta \in K_1^c$, using the assumption that $\nabla \log \pi^{(0)}$ is L_0-Lipschitz and $h^{(n)} = c_h n^{\mathfrak{h}}$ and $w^{(n)} = n^{\mathfrak{w}}$,

$$\begin{aligned}
\left\| \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\| &\leq \frac{h^{(n_m)} w^{(n_m)} \|\Gamma\|}{2n_m} \left\| \nabla \log \pi^{(0)} \left(\widehat{\theta}^{(n_m)} + (w^{(n_m)})^{-1} \vartheta \right) \right\| \\
&\leq \frac{h^{(n_m)} w^{(n_m)} \|\Gamma\|}{2n_m} \left(\left\| \nabla \log \pi^{(0)} (\theta_\star) \right\| + L_0 \left\| \widehat{\theta}^{(n_m)} - \theta_\star \right\| + \frac{L_0 \|\vartheta\|}{w^{(n_m)}} \right) \\
&\leq \frac{c_h n_m^{\mathfrak{w} - \mathfrak{h} - 1} \|\Gamma\|}{2} \left(\left\| \nabla \log \pi^{(0)} (\theta_\star) \right\| + L_0 \left\| \widehat{\theta}^{(n_m)} - \theta_\star \right\| + \frac{L_0 \|\vartheta\|}{n_m^{\mathfrak{w}}} \right),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\left\| \Delta_\ell^{(n_m)}(\vartheta) \right\| \\
&\quad \leq \frac{h^{(n_m)} w^{(n_m)} \|\Gamma\|}{2b^{(n_m)}} \left\| \sum_{j \in [b^{(n_m)}]} \nabla \ell \left(\widehat{\theta}^{(n_m)} + (w^{(n_m)})^{-1} \vartheta;\ X_{I_1^{(n_m)}(j)} \right) \right\| \\
&\quad \leq \frac{c_h n_m^{\mathfrak{w} - \mathfrak{h}} \|\Gamma\|}{2b^{(n_m)}} \sum_{j \in [b^{(n_m)}]} \left(\left\| \nabla \ell \left(\theta_\star;\ X_{I_1^{(n_m)}(j)} \right) \right\| + L(X_{I_1^{(n_m)}(j)}) \left\| \widehat{\theta}^{(n_m)} - \theta_\star \right\| + \frac{L(X_{I_1^{(n_m)}(j)}) \|\vartheta\|}{{n_m}^{\mathfrak{w}}} \right) \\
&\quad \leq \frac{c_h n_m^{\mathfrak{w} - \mathfrak{h}} \|\Gamma\|}{2} \left(L_\star(\mathbf{X}^{(n_m)}) + L(\mathbf{X}^{(n_m)}) \left\| \widehat{\theta}^{(n_m)} - \theta_\star \right\| + L(\mathbf{X}^{(n_m)}) \frac{\|\vartheta\|}{{n_m}^{\mathfrak{w}}} \right)
\end{aligned}$$

where we define the (random) Lipschitz constants $L(X_i)$, $L_\star(\mathbf{X}^{(n_m)})$, and $L(\mathbf{X}^{(n_m)})$ by:

$$\begin{aligned}
L(X_i) &:= \left\| \nabla^{\otimes 2} \ell(\cdot; X_i) \right\|_\infty, \\
L_\star(\mathbf{X}^{(n_m)}) &:= \max_{i \leq n_m} \| \nabla \ell (\theta_\star;\ X_i) \|, \text{ and} \\
L(\mathbf{X}^{(n_m)}) &:= \max_{i \leq n_m} L(X_i).
\end{aligned}$$

Using that $\mathbf{X}^{(\mathbb{N})} \in \Omega^{(0)}$, so that $\Upsilon^{(n_m)} \to 0$ etc., if m is large enough that all of the following hold:

$$\sup_{m' \geq m} \Upsilon^{(n_m)} \leq \min(1, L_0^{-1}),$$

$$1 \geq \sup_{m' \geq m} \frac{L_\star(\mathbf{X}^{(n_{m'})})}{n_{m'}^{1/p_2}},$$

$$n_m \geq \max((2c_h \|\Gamma\|)^{1/(1/p_3 - \mathfrak{h})}, (2c_h L_0 \|\Gamma\|)^{\frac{1}{\mathfrak{h}+1-\mathfrak{a}-\mathfrak{w}}}), \qquad \text{and}$$

$$1 \geq \sup_{m' \geq m} \frac{L(\mathbf{X}^{(n_{m'})})}{n_{m'}^{1/p_3}};$$

then, using that $0 < \mathfrak{w} < 1$,

$$\left\|\Delta^{(n_m)}_{\pi^{(0)}}(\vartheta)\right\| \leq \frac{c_h \|\Gamma\|}{2} \left(\left\|\nabla \log \pi^{(0)} (\theta_\star)\right\| + 1\right) + \frac{1}{4} \|\vartheta\|,$$

and

$$\begin{aligned}\left\|\Delta^{(n_m)}_{\ell}(\vartheta)\right\| &\leq \frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}} \|\Gamma\|}{2} \left(n_m^{1/p_2} + n_m^{1/p_3} \Upsilon^{(n_m)} + n_m^{1/p_3 - \mathfrak{w}} \|\vartheta\|\right) \\ &\leq \frac{c_h \|\Gamma\|}{2} \left(n_m^{1/p_2 - \mathfrak{h} + \mathfrak{w}} + n_m^{1/p_3 - \mathfrak{h} + \mathfrak{w}} \Upsilon^{(n_m)} + n_m^{1/p_3 - \mathfrak{h}} \|\vartheta\|\right), \\ &\leq c_h \|\Gamma\| + \frac{1}{4} \|\vartheta\|.\end{aligned}$$

Therefore, for $\vartheta \in K_1^c$ (and hence $\|\vartheta\| > 2R_0 + 2c_0$),

$$\begin{aligned}&\|\vartheta\| - \left\|\Delta^{(n_m)}_{\pi^{(0)}}(\vartheta)\right\| - \left\|\Delta^{(n_m)}_{\ell}(\vartheta)\right\| - R_0 \\ &\geq \frac{1}{2} \|\vartheta\| - \frac{c_h \|\Gamma\|}{2} \left(3 + \left\|\nabla \log \pi^{(0)} (\theta_\star)\right\|\right) - R_0 \\ &\geq \sqrt{c_h / c_\beta \|\Lambda\|}.\end{aligned}$$

Therefore, combining this with Eq. (4.29) and the definition of $\Delta^{(n_m)}_{\xi}(\vartheta)$,

$$\begin{aligned}\lim_{m \to \infty} \sup_{\vartheta \in K_1^c} \left|[A^{(n_m)} f](\vartheta) - [Af](\vartheta)\right| &\leq \lim_{m \to \infty} \alpha^{(n_m)} \|f\|_\infty \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}} \left(\|\xi_1\| \geq n_m^{\mathfrak{h}/2 + \mathfrak{t}/2 - \mathfrak{w}}\right) \\ &\leq \lim_{m \to \infty} \alpha^{(n_m)} \|f\|_\infty d \, \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}} \left(|\xi_{1,1}| \geq \frac{1}{\sqrt{d}} n_m^{\mathfrak{h}/2 + \mathfrak{t}/2 - \mathfrak{w}}\right) \\ &\leq \lim_{m \to \infty} 2 n_m^{\mathfrak{a}} \|f\|_\infty d \exp(-n_m^{\mathfrak{h} + \mathfrak{t} - 2\mathfrak{w}} / 2d) \\ &= 0.\end{aligned}$$

since $\mathfrak{h} + \mathfrak{t} - 2\mathfrak{w} \geq \mathfrak{a} > 0$.

4.7.1.5 Taylor expansion near the test function support

Recalling the definition of $A^{(n_m)}$ in Eq. (4.26), using the definition of the time-scaling factor $\alpha^{(n)} = n^{\mathfrak{a}}$, taking a second-order Taylor expansion of the test function $f \in C_c^\infty$, and applying the decomposition $\Delta^{(n_m)}(\vartheta) = \Delta_\xi^{(n_m)}(\vartheta) + \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) + \Delta_\ell^{(n_m)}(\vartheta)$,

$$
\begin{aligned}
&[A^{(n_m)} f](\vartheta) \\
&= \alpha^{(n_m)} \left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[f\left(\vartheta + \Delta^{(n_m)}(\vartheta)\right)\right] - f(\vartheta)\right) \\
&= \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla f(\vartheta),\ \Delta_\xi^{(n_m)} \right\rangle}_{[1.\xi]^{(n_m)}(\vartheta)=0} + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla f(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\rangle}_{\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)} + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla f(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta) \right\rangle}_{[1.\ell]^{(n_m)}(\vartheta)} \\
&\quad + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta_\xi^{(n_m)},\ \Delta_\xi^{(n_m)} \right\rangle}_{[2.\xi\xi]^{(n_m)}(\vartheta)} + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla^{\otimes 2} f(\vartheta) \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_\xi^{(n_m)} \right\rangle}_{\left[2.\pi^{(0)}\xi\right]^{(n_m)}(\vartheta)=0} \\
&\quad + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla^{\otimes 2} f(\vartheta) \Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\xi^{(n_m)} \right\rangle}_{[2.\ell\xi]^{(n_m)}(\vartheta)=0} + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\rangle}_{\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)} \\
&\quad + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla^{\otimes 2} f(\vartheta) \Delta_\ell^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)} \right\rangle}_{\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)} + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta) \right\rangle}_{[2.\ell\ell]^{(n_m)}(\vartheta)} \\
&\quad + \underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\frac{1}{6} \left[\nabla^{\otimes 3} f(\vartheta + S\Delta^{(n_m)}(\vartheta))\right] \left(\Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta)\right)\right]}_{[3.R]^{(n_m)}(\vartheta)}
\end{aligned}
$$

for some $S \in [0,1]$ depending on $f, \vartheta, \Delta^{(n_m)}(\vartheta)$, where $\nabla^{\otimes 3} f(\vartheta)$ is the trilinear from of third order partials of f at ϑ (and hence is linear in each of its three arguments). Terms that are linear in $\Delta_\xi^{(n_m)}$ have mean 0 and can be eliminated outright, as indicated in their corresponding underbraces. Terms are labelled by the order of the term, followed by the increments that appear in the term; for example $[2.\ell\xi]^{(n_m)}(\vartheta)$ is the second order term involving a likelihood increment and a Gaussian noise (innovation) increment. The R in $[3.R]^{(n_m)}(\vartheta)$ denotes that it is the *remainder*.

Recall that

$$
[Af](\vartheta) = \underbrace{-\left\langle \frac{c_{\mathtt{d}}}{2} \Gamma \mathcal{J}_\star \vartheta,\ \nabla f(\vartheta) \right\rangle}_{[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)} + \underbrace{\frac{c_{\mathtt{g}}}{2} \Lambda : \nabla^{\otimes 2} f(\vartheta)}_{[\mathrm{II}.\Lambda](\vartheta)} + \underbrace{\frac{c_{\mathtt{mb}}}{2} \Gamma \mathcal{I}_\star \Gamma' : \nabla^{\otimes 2} f(\vartheta)}_{[\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)}.
$$

We have similarly labelled these terms, with the roman numeral denoting the order and the subsequent symbol denoting the coefficient matrix (up to scaling factors). Thus, after elminating terms which are linear in $\Delta_\xi^{(n_m)}$, and thus have mean 0, the difference of approximate and limiting generator applied to the test function can be expressed as

$$\begin{aligned}
&\left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| \\
&\quad\le \left|\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| + \left|\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| + \left|\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\qquad + \left|[1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)\right| \\
&\qquad + \left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\mathrm{II}.\Lambda](\vartheta)\right| \\
&\qquad + \left|[2.\ell\ell]^{(n_m)}(\vartheta) - \left[\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'\right](\vartheta)\right| \\
&\qquad + \left|[3.R]^{(n_m)}(\vartheta)\right|.
\end{aligned}$$

We will show that each of these seven terms vanish uniformly on K_1. The first three terms listed above, those non-remainder terms with no corresponding term in the limiting generator, will be handled first. Then we will handle each of the terms which corresponds to part of the limiting generator, and lastly we will handle the remainder term.

4.7.1.6 Terms that do not contribute to the limit

$$\begin{aligned}
\left|\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| &= n_m^{\mathfrak{a}}\left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla f(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\le \frac{c_h n_m^{\mathfrak{a}-\mathfrak{h}+\mathfrak{w}-1}\|\Gamma\|}{2}\left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla f(\vartheta),\ \nabla\log\pi^{(0)}\left(\widehat{\theta}^{(n_m)} + n_m^{-\mathfrak{w}}\vartheta\right)\right\rangle\right| \\
&\le \frac{c_h n_m^{\mathfrak{a}-\mathfrak{h}+\mathfrak{w}-1}\|\Gamma\|}{2}\|\nabla f\|_\infty\left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\left(\Upsilon^{(n_m)} + \frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right)\right),
\end{aligned}$$

which vanishes uniformly on K_1, since $\mathfrak{a}+\mathfrak{w}-\mathfrak{h}-1 \le \mathfrak{w}-1 < 0$.

$$\begin{aligned}
&\left|\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\quad = \left|n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\quad \leq n_m^{\mathfrak{a}}\left\|\nabla^{\otimes 2}f\right\|_\infty \left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}-1}\|\Gamma\|}{2}\right)^2 \\
&\qquad \times \left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\left\|\widehat{\theta}^{(n_m)}-\theta_\star\right\| + L_0\frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right)^2
\end{aligned}$$

which vanishes uniformly since $\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-2 \leq (2\mathfrak{w}-2)-h < 0$ (which follows from $\mathfrak{h}\geq\mathfrak{a}$ and $\mathfrak{w}<1$).

$$\begin{aligned}
&\left|\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\quad = \left|n_m^{\mathfrak{a}}2\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_{\ell}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\quad \leq 2n_m^{\mathfrak{a}}\left\|\nabla^{\otimes 2}f\right\|_\infty \left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}-1}\|\Gamma\|}{2}\right)\left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}}\|\Gamma\|}{2}\right) \\
&\qquad \times \left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\Upsilon^{(n_m)} + L_0\frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right) \\
&\qquad \times \left(n_m^{1/p_2} + n_m^{1/p_3}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{w}}\right)
\end{aligned}$$

which vanishes uniformly due to the assumptions of the relationship between $\mathfrak{h},\mathfrak{a},\mathfrak{w},p_3,p_2$ under each assumption.

4.7.1.7 Convergence of the drift term

Third, using that $\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)};\ X_i\right)=0$,

$$\begin{aligned}
&[1.\ell]^{(n_m)}(\vartheta) \\
&\quad = n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla f(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta)\right\rangle \\
&\quad = \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla f(\vartheta),\ \frac{c_h n_m^{\mathfrak{a}+\mathfrak{w}-\mathfrak{h}}\Gamma}{2b^{(n_m)}}\sum_{j\in\left[b^{(n_m)}\right]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)\right\rangle \\
&\quad = \left\langle \frac{c_h\Gamma^\dagger}{2}\nabla f(\vartheta),\ n_m^{\mathfrak{a}+\mathfrak{w}-\mathfrak{h}-1}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right\rangle \\
&\quad = \left\langle \frac{c_h\Gamma^\dagger}{2}\nabla f(\vartheta),\ \left(\int_0^1 n_m^{\mathfrak{a}-\mathfrak{h}-1}\sum_{i\in[n_m]}\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)}+\frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)ds\right)\right\rangle
\end{aligned} \tag{4.30}$$

Now, for all n_m large enough that $r_{\mathcal{J},n_m} \geq R_0 + c_0$

$$\begin{aligned}
&\left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2} \ell \left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) ds\ + \mathcal{J}_\star \right) \vartheta \right\rangle \right| \\
&\quad \leq c_h \left\| \Gamma \right\| \left\| \nabla f \right\|_\infty (R_0 + c_0) \left\| \int_0^1 \left[n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2} \ell \left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) + \mathcal{J}_\star \right] ds \right\| \\
&\quad \leq c_h \left\| \Gamma \right\| \left\| \nabla f \right\|_\infty (R_0 + c_0) \cdot \Upsilon^{(n_m)},
\end{aligned}$$

and thus vanishes uniformly on K_1.

When $\mathfrak{a} > \mathfrak{h}$, so $c_{\mathfrak{d}} = 0$ and hence $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) = 0$ (where $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)$ is the drift term appearing in the definition of the limiting generator A in Eq. (4.27)), then the drift term will be inactive in the limit. We show this by using the fact that $[1.\ell]^{(n_m)}(\vartheta)$ is a vanishing distance from a sequence that vanishes:

$$\begin{aligned}
&\left| [1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) \right| \\
&\quad \leq n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2} \ell \left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) ds\ + \mathcal{J}_\star \right) \vartheta \right\rangle \right| \\
&\qquad + n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \mathcal{J}_\star \vartheta \right\rangle \right|;
\end{aligned}$$

and hence vanishes uniformly on K_1.

When $\mathfrak{a} = \mathfrak{h}$, then the drift term is active in the limit, and we show that $[1.\ell]^{(n_m)}(\vartheta)$ converges to the drift term from the limiting process $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)$:

$$\begin{aligned}
&\left| [1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) \right| \\
&\quad = n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2} \ell \left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) ds\ + \mathcal{J}_\star \right) \vartheta \right\rangle \right|
\end{aligned}$$

vanishes uniformly on K_1.

4.7.1.8 Convergence of the diffusion term corresponding to Gaussian noise

$$\begin{aligned}
&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\mathrm{II}.\Lambda](\vartheta)\right| \\
&\quad = \left| n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2}\nabla^{\otimes 2} f(\vartheta)\Delta_\xi^{(n_m)},\ \Delta_\xi^{(n_m)} \right\rangle - \frac{c_h}{2c_\beta}\Lambda : \nabla^{\otimes 2} f(\vartheta) \right|
\end{aligned}$$

If $\mathfrak{a} + 2\mathfrak{w} - \mathfrak{h} - \mathfrak{t} = 0$ then, the corresponding diffusion term is active in the limit. Using the definition of $\Delta_\xi^{(n_m)}$ and that $\beta^{(n)} = c_\beta n^{\mathfrak{t}}$, $\beta_h = c_h n^{\mathfrak{h}}$, and $\beta_w = n^{\mathfrak{w}}$

$$\begin{aligned}
&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\mathrm{II}.\Lambda](\vartheta)\right| \\
&\quad \leq \frac{c_h}{2c_\beta}\left| n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla^{\otimes 2} f(\vartheta)\sqrt{\Lambda}\xi_1,\ \sqrt{\Lambda}\xi_1 \right\rangle - \Lambda : \nabla^{\otimes 2} f(\vartheta) \right| \\
&\quad = 0
\end{aligned}$$

If $\mathfrak{a} + 2\mathfrak{w} - \mathfrak{h} - \mathfrak{t} < 0$ then the corresponding diffusion term is inactive in the limit, and so $c_{\mathfrak{g}} = 0$ and so $[\mathrm{II}.\Lambda](\vartheta) = 0$. In that case we show that $[2.\xi\xi]^{(n_m)}(\vartheta)$ vanishes uniformly.

$$\begin{aligned}
&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\mathrm{II}.\Lambda](\vartheta)\right| \\
&\quad \leq \frac{c_h}{2c_\beta} n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}} \left| \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla^{\otimes 2} f(\vartheta)\sqrt{\Lambda}\xi_1,\ \sqrt{\Lambda}\xi_1 \right\rangle \right| \\
&\quad = \frac{c_h}{2c_\beta} n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}} \|\Lambda\|_F \left\| \left\| \nabla^{\otimes 2} f \right\|_F \right\|_\infty,
\end{aligned}$$

which vanishes uniformly.

4.7.1.9 Convergence of the diffusion term corresponding to minibatch noise

$$\begin{aligned}
&\left|[2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right| \\
&\quad = \left| n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2}\nabla^{\otimes 2} f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta) \right\rangle - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' : \nabla^{\otimes 2} f(\vartheta) \right| \\
&\quad = \frac{1}{2}\left| \left[n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\left(\Delta_\ell^{(n_m)}(\vartheta) \right)^{\otimes 2} \right] : \nabla^{\otimes 2} f(\vartheta) - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' : \nabla^{\otimes 2} f(\vartheta) \right] \right| \\
&\quad \leq \frac{\|\nabla^{\otimes 2} f_F\|_\infty}{2} \left\| \left[n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\left(\Delta_\ell^{(n_m)}(\vartheta) \right)^{\otimes 2} \right] - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' \right] \right\|_F \\
&\quad \leq \sqrt{d}\frac{\|\nabla^{\otimes 2} f_F\|_\infty}{2} \left\| \left[n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\left(\Delta_\ell^{(n_m)}(\vartheta) \right)^{\otimes 2} \right] - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' \right] \right\|
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m^{\mathfrak{a}} \left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\\
&= \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2}\Gamma\left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)^{\otimes 2}\right)\Gamma'\\
&\quad + \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2}\Gamma\Bigg(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]}\ \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j)}\right)\\
&\qquad\qquad\qquad \otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j')}\right)\Bigg)\Gamma'\\
&= \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4b^{(n_m)}}\Gamma\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)\Gamma'\\
&\quad + \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2}\Gamma\Bigg(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]}\ \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j)}\right)\\
&\qquad\qquad\qquad \otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j')}\right)\Bigg)\Gamma'
\end{aligned}$$

If the mini-batches are drawn with replacement, then

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]}\ \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)\otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j')}\right)\\
&= \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m^2}\sum_{i\in[n_m]}\sum_{i'\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right)\\
&= b^{(n_m)}(b^{(n_m)}-1)\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2}\\
&= b^{(n_m)}(b^{(n_m)}-1)\left(\frac{1}{n_m}\sum_{i\in[n_m]} \int_0^1 \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)ds\frac{1}{n_m^{\mathfrak{w}}}\vartheta\right)^{\otimes 2}
\end{aligned}$$

Thus, if $\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-\mathfrak{b}=0$, so that $c_{\mathtt{mb}}\neq 0$ and the corresponding term is active in the limit, and the minibtaches are drawn with replacement, then combining the past several equations gives:

$$
\begin{aligned}
&\left|[2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right| \\
&\quad \leq \frac{\sqrt{d}\left\|\Gamma\right\|^2\left\|\nabla^{\otimes 2} f_F\right\|_\infty}{2}\left\|\frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)-c_{\mathtt{mb}}\mathcal{I}_\star\right\| \\
&\qquad + \frac{\sqrt{d}c_h^2\left\|\Gamma\right\|^2\left\|\nabla^{\otimes 2} f_F\right\|_\infty n_m^{-2\mathfrak{w}}}{8}\left\|\left(\frac{1}{n_m}\sum_{i\in[n_m]}\int_0^1\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)}+\frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)ds\ \vartheta\right)^{\otimes 2}\right\|
\end{aligned}
$$

For $\vartheta \in K_1$, and for all n_m large enough that $r_{\mathcal{J},n_m} \geq R_0 + c_0$

$$
\begin{aligned}
n_m^{-2\mathfrak{w}}&\left\|\left(\frac{1}{n_m}\sum_{i\in[n_m]}\int_0^1\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)}+\frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)ds\ \vartheta\right)^{\otimes 2}\right\| \\
&= n_m^{-2\mathfrak{w}}\left\|\frac{1}{n_m}\sum_{i\in[n_m]}\int_0^1\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)}+\frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)ds\ \vartheta\right\|^2 \\
&\leq \frac{(2R_0+2c_0)^2}{n_m^{2\mathfrak{w}}}\left(\left\|\mathcal{J}_\star\right\|+\Upsilon^{(n_m)}\right)^2,
\end{aligned}
$$

which vanishes uniformly.

Since the mini-batches are drawn with replacement, using the definition of $c_{\mathtt{mb}}$, for all n_m large enough that $r_{\mathcal{I},n_m} \geq R_0 + c_0$

$$
\begin{aligned}
&\left\|\frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)-c_{\mathtt{mb}}\mathcal{I}_\star\right\| \\
&\quad \leq \frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}\left\|\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)-\mathcal{I}_\star\right\| \\
&\qquad + \left|\frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}-\frac{c_h^2}{4c_b}\right|\left\|\mathcal{I}_\star\right\| \\
&\quad \leq \frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}\Upsilon^{(n_m)}+\left|\frac{c_h^2}{4c_b}\frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}}\rfloor}-\frac{c_h^2}{4c_b}\right|\left\|\mathcal{I}_\star\right\|.
\end{aligned}
$$

And, if $\mathfrak{a} + 2\mathfrak{w} - 2\mathfrak{h} - \mathfrak{b} < 0$ and the mini-batches are drawn with replacement, so that $c_{\mathtt{mb}} = 0$, and the corresponding diffusion term is inactive in the limit and $[\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta) = 0$, then

$$\begin{aligned}&\left|[2.\ell\ell]^{(n_m)}(\vartheta) - [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right|\\&\quad\leq \left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m^{\mathfrak{a}}\left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2} - n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-\mathfrak{b}}\frac{c_h^2}{4c_b}\mathcal{I}_\star\right| + n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-\mathfrak{b}}\left|\frac{c_h^2}{4c_b}\mathcal{I}_\star\right|\end{aligned}$$

which vanishes uniformly by the previous arguments.

Therefore, when the mini-batches are drawn with replacement, we find that

$$\left|[2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right|$$

vanishes uniformly on K_1.

If the mini-batches are drawn without replacement.

$$\begin{aligned}&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\sum_{j\in[b^{(n_m)}]}\sum_{j'\in[b^{(n_m)}]\setminus\{j\}}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)\otimes\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j')}\right)\\&\quad=\frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)}\sum_{i\in[n_m]}\sum_{i'\in[n_m]\setminus\{i\}}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\otimes\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right)\\&\quad=\frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)}\sum_{i\in[n_m]}\sum_{i'\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\otimes\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right)\\&\qquad-\frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\\&\quad=b^{(n_m)}(b^{(n_m)}-1)\frac{n_m}{n_m-1}\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2}\\&\qquad-\frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2},\end{aligned}$$

and so,

$$\begin{aligned}&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m\left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\\&\quad=\frac{c_h^2}{4b^{(n_m)}}\frac{n_m-b^{(n_m)}}{n_m-1}\Gamma\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)\Gamma'\\&\qquad+\frac{c_h^2}{4(b^{(n_m)})^2}\Gamma\left(b^{(n_m)}(b^{(n_m)}-1)\frac{n_m}{n_m-1}\left(\frac{1}{n_m}\sum_{i\in[n_m]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+\frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2}\right)\Gamma'\end{aligned}$$

In this case, for all n_m large enough that $r_{\mathcal{I},n_m} \geq R_0 + c_0$

$$\begin{aligned}
&\left\| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \left(\frac{1}{n_m} \sum_{i\in[n_m]} \nabla \ell \left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right)^{\otimes 2} \right) - c_{\mathtt{mb}} \mathcal{I}_\star \right\| \\
&\quad \leq \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \left\| \left(\frac{1}{n_m} \sum_{i\in[n_m]} \nabla \ell \left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right)^{\otimes 2} \right) - \mathcal{I}_\star \right\| \\
&\qquad + \left| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} - c_{\mathtt{mb}} \right| \|\mathcal{I}_\star\| \\
&\quad \leq \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \Upsilon^{(n_m)} + \left| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} - c_{\mathtt{mb}} \right| \|\mathcal{I}_\star\|,
\end{aligned}$$

Thus, when the mini-batches are drawn without replacement, we find that

$$\left| [2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta) \right|$$

vanishes uniformly on K_1.

4.7.1.10 Convergence of the Remainder Term

$$\begin{aligned}
&\left| [3.R]^{(n_m)}(\vartheta) \right| \\
&\quad = n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\frac{1}{6} \left[\nabla^{\otimes 3} f(\vartheta + S\Delta^{(n_m)}(\vartheta)) \right] \left(\Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta) \right) \right] \\
&\quad \leq \frac{n_m^{\mathfrak{a}}}{6} \left\| \nabla^{\otimes 3} f \right\|_\infty \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\| \Delta^{(n_m)}(\vartheta) \right\|^3 \\
&\quad \leq \frac{27 n_m^{\mathfrak{a}}}{6} \left\| \nabla^{\otimes 3} f \right\|_\infty \left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\| \Delta_\xi^{(n_m)} \right\|^3 + \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\| \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\|^3 + \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\| \Delta_\ell^{(n_m)}(\vartheta) \right\|^3 \right),
\end{aligned}$$

Now

$$\begin{aligned}
\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\| \Delta_\xi^{(n_m)} \right\|^3 &\leq \left(\frac{c_h}{2c_\beta} n_m^{-\mathfrak{h}-\mathfrak{t}+2\mathfrak{w}} \|\Lambda\| \right)^{3/2} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \|\xi_1\|^3 \\
&= n_m^{-3/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w})} \left(\frac{c_h}{2c_\beta} \|\Lambda\| \right)^{3/2} 2^{3/2} \frac{\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d}{2}\right)},
\end{aligned}$$

where Γ is the gamma function. Note that $\alpha - 3/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}) \leq -1/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}) \leq -\mathfrak{a}/2 < 0$

Second,

$$\left\|\Delta^{(n_m)}_{\pi^{(0)}}(\vartheta)\right\|^3 \le \left(\frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}-1}\,\|\Gamma\|}{2}\right)^3 \left(\left\|\nabla \log \pi^{(0)}(\theta_\star)\right\| + L_0\left\|\widehat{\theta}^{(n_m)} - \theta_\star\right\| + L_0\frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right)^3.$$

Note that $\mathfrak{a} - 3\mathfrak{h} + 3\mathfrak{w} - 3 \le -2\mathfrak{h} - 3(1-\mathfrak{w}) < 0$.

Third,

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\|\Delta^{(n_m)}_{\ell}(\vartheta)\right\|^3 \\
&\quad\le \left(\frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}}\,\|\Gamma\|}{2}\right)^3 \left(n_m^{1/p_2} + n_m^{1/p_3}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{w}}\right)^3 \\
&\quad\le \left(\frac{c_h\,\|\Gamma\|}{2}\right)^3 \left(n_m^{1/p_2-\mathfrak{h}+\mathfrak{w}} + n_m^{1/p_3-\mathfrak{h}+\mathfrak{w}}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{h}}\right)^3
\end{aligned}$$

Therefore, $\left|[3.R]^{(n_m)}(\vartheta)\right|$ vanishes uniformly. □

4.8 Proof of Corollary 4.2

Proof of Corollary 4.2. To verify that that the stationary measures, $\nu^{(n_m)}$ of $T^{(n_m)}$ converge weakly in probability to ν, we need to verify that every sub-subsequence $\nu^{(n_{m_k})}$ has a sub-sub-subsequence $\nu^{(n_{m_{k_j}})}$ converging weakly to ν almost surely. Since weak convergence of probability measures is metrizable, then applying Lemma 4.1 yields the desired result.

By the second part of Theorem 4.2, every sub-subsequence of $(T^{(n_m)})_{m\in\mathbb{N}}$, $(T^{(n_{m_k})})_{k\in\mathbb{N}}$, has a further sub-sub-subsequence, $(T^{(n_{m_{k_j}})})_{j\in\mathbb{N}}$, such that with probability 1, $T_t^{(n_{m_{k_j}})} \xrightarrow{\text{s}} T_t$ on $\overline{C}(\mathbb{R}^d)$ for all $t > 0$.

Applying Ethier and Kurtz [33, Part 4, Theorem 9.10], we have that every weak limit of $\{\nu^{(n_{m_{k_j}})}\}_{j\in\mathbb{N}}$ is stationary for T. As a consequence of the assumption that the spectrum of $\Gamma\mathcal{J}(\theta_\star)$ is a subset of $\{x \in \mathbb{C} \text{ s.t. } \Re(x) > 0\}$, T has a unique stationary distribution (see, for example, Karatzas and Shreve [54]), $\nu = \mathrm{N}(0, Q_\infty)$. Thus every weak limit of $\{\nu^{(n_{m_{k_j}})}\}_{j\in\mathbb{N}}$ must be ν.

Since $\{\nu^{(n_m)}\}_{m\in\mathbb{N}}$ is assumed to be tight, then all of its sub-subsequences have a weakly converging sub-sub-subsequence, concluding the proof. □

4.9 Sufficient conditions for Assumptions 4.4 and 4.5

In this section we provide some sufficient conditions that ensure Assumptions 4.4 and 4.5. For each of the two assumptions, we one sufficient condition based on convergence of the corresponding information matrix empirical process, one sufficient condition based on equicontinuity of the derivatives of the likelihood function, and one sufficient condition based expected Lipschitz or local Lipschitz constants for the derivatives of the likelihood.

Proposition 4.4 (Sufficient conditions for Assumption 4.4)**.** *Each of the following imply Assumption 4.4.*

a) there exists a $\delta_1 > 0$ *with* $\sup_{\theta \in B_{\delta_1}(\theta_\star)} \left\| \frac{1}{n} \sum_{i\in[n]} \nabla^{\otimes 2}\ell(\theta; X_i) + \mathcal{J}(\theta) \right\| \xrightarrow{p} 0$ *and* $\mathcal{J}$ *is continuous at* $\theta_\star$*,*

b) $\left\{ \nabla^{\otimes 2}\ell(\cdot; x) \mid x \in \mathcal{X} \right\}$ *is equicontinuous at* $\theta_\star$*,*

c) there exists a $\delta_1 > 0$ *with*

$$\mathbb{E}\left[\sup_{\theta \in B_{\delta_1}(\theta_\star)} \frac{\left\| \nabla^{\otimes 2}\ell(\theta; X_1) - \nabla^{\otimes 2}\ell(\theta_\star; X_1) \right\|}{\|\theta - \theta_\star\|} \right] < \infty,$$

Proof of Proposition 4.4.

a) Let $r_{\mathcal{J},n} = \delta_1 n^{\mathfrak{w}/2}/2$. Then $B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right) \subseteq B\left(\widehat{\theta}^{(n)}, \delta_1/2\right)$.

Given that $\widehat{\theta}^{(n)} \xrightarrow{\mathrm{p}} \theta_\star$, any subsequence of indices n_m has a further sub-subsequence of indices n_{m_k} where both $\widehat{\theta}^{(n_{m_k})} \to \theta_\star$ and

$$\sup_{\theta \in B_{\delta_1}(\theta_\star)} \left\| \frac{1}{n_{m_k}} \sum_{i \in \left[n_{m_k}\right]} \nabla^{\otimes 2}\ell(\theta; X_i) + \mathcal{J}(\theta) \right\| \to 0 \text{ a.s.}$$

Then there is a k_0 such that if $k \geq k_0$ then $\left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| \leq \delta_1/2$. Therefore if $k \geq k_0$ then $B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right) \subseteq B\left(\theta_\star, \delta_1\right)$.

Thus, for $k \geq k_0$,

$$
\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r/n_{m_k}^{\mathfrak{w}}\right)} \left\| \mathcal{J}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B(\theta_\star, \delta_1)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1/n_{m_k}^{\mathfrak{w}/2}\right)} \left\| \mathcal{J}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B(\theta_\star, \delta_1)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\theta_\star, \left\|\widehat{\theta}^{(n_{m_k})} - \theta_\star\right\| + \delta_1/n_{m_k}^{\mathfrak{w}/2}\right)} \left\| \mathcal{J}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \xrightarrow{\text{a.s.}} 0.
\end{aligned}
$$

Therefore, every subsequence of $S_n = \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\|$ has a further sub-subsequence converging almost surely to 0, and hence S_n converges in probability to 0.

b) Equicontinuity implies there is a function $\rho_{\mathcal{J}_\star} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t \to 0} \rho_{\mathcal{J}_\star}(t) = 0$, and

$$
\sup_{x \in \mathcal{X}} \sup_{\vartheta \in B_\delta(\theta_\star)} \left\| \nabla^{\otimes 2} \ell(\vartheta; x) - \nabla^{\otimes 2} \ell(\theta_\star; x) \right\| \leq \rho_{\mathcal{J}_\star}(\delta).
$$

Let $r_{\mathcal{J},n} = n^{\mathfrak{w}/2}$. Then

$$
\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \widehat{\mathcal{J}}^{(n)}(\theta_\star) \right\| + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\theta_\star, \left\|\widehat{\theta}^{(n)} - \theta_\star\right\| + n^{-\mathfrak{w}/2}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \widehat{\mathcal{J}}^{(n)}(\theta_\star) \right\| + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \rho_{\mathcal{J}_\star} \left(\left\| \widehat{\theta}^{(n)} - \theta_\star \right\| + n^{-\mathfrak{w}/2} \right) + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \xrightarrow{\text{p}} 0.
\end{aligned}
$$

In the last step we used that the first term vanishes in probability because $\widehat{\theta}^{(n)} \xrightarrow{\text{p}} \theta_\star$, and the second term vanishes in probability by the weak law of large numbers.

c) Let

$$Q_n = \frac{1}{n}\sum_{i\in[n]}\left[\sup_{\theta\in B_{\delta_1}(\theta_\star)}\frac{\|\nabla^{\otimes 2}\ell(\theta;X_i)-\nabla^{\otimes 2}\ell(\theta_\star;X_i)\|}{\|\theta-\theta_\star\|}\right], \text{ and}$$
$$q = \mathbb{E}\left[\sup_{\theta\in B_{\delta_1}(\theta_\star)}\frac{\|\nabla^{\otimes 2}\ell(\theta;X_1)-\nabla^{\otimes 2}\ell(\theta_\star;X_1)\|}{\|\theta-\theta_\star\|}\right].$$

By the weak law of large numbers, $Q_n \xrightarrow{\text{p}} q$ and $\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \xrightarrow{\text{p}} \mathcal{J}(\theta_\star)$. Let $r_{\mathcal{J},n} = \delta_1 n^{\mathfrak{w}/2}/2$. As in part a), given that $\widehat{\theta}^{(n)} \xrightarrow{\text{p}} \theta_\star$, any subsequence of indices n_m has a further sub-subsequence of indices n_{m_k} where both $\widehat{\theta}^{(n_{m_k})} \to \theta_\star$, $Q_{n_{m_k}} \to q$, and $\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \to \mathcal{J}(\theta_\star)$ almost surely. Then there is a k_0 such that if $k \geq k_0$ then $\left\|\widehat{\theta}^{(n_{m_k})} - \theta_\star\right\| \leq \delta_1/2$. Therefore if $k \geq k_0$ then $B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right) \subseteq B(\theta_\star, \delta_1)$.

Thus, for $k \geq k_0$,

$$\begin{aligned}
&\sup_{\theta\in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta)-\mathcal{J}_\star\right\| \\
&\leq \left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)-\mathcal{J}(\theta_\star)\right\| + \sup_{\theta\in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)}\left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta)-\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)\right\| \\
&\leq \left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)-\mathcal{J}(\theta_\star)\right\| \\
&\qquad + \left(\left\|\widehat{\theta}^{(n_{m_k})}-\theta_\star\right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)\sup_{\theta\in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)}\frac{\left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta)-\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)\right\|}{\|\theta-\theta_\star\|} \\
&\leq \left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)-\mathcal{J}(\theta_\star)\right\| \\
&\qquad + \left(\left\|\widehat{\theta}^{(n_{m_k})}-\theta_\star\right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)\sup_{\theta\in B(\theta_\star,\delta_1)}\frac{\left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta)-\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)\right\|}{\|\theta-\theta_\star\|} \\
&\leq \left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)-\mathcal{J}(\theta_\star)\right\| \\
&\qquad + \left(\left\|\widehat{\theta}^{(n_{m_k})}-\theta_\star\right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)\sup_{\theta\in B(\theta_\star,\delta_1)}\frac{1}{n_{m_k}}\sum_{i\in\left[n_{m_k}\right]}\left[\frac{\|\nabla^{\otimes 2}\ell(\theta;X_i)-\nabla^{\otimes 2}\ell(\theta_\star;X_i)\|}{\|\theta-\theta_\star\|}\right] \\
&\leq \left\|\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star)-\mathcal{J}(\theta_\star)\right\| + \left(\left\|\widehat{\theta}^{(n_{m_k})}-\theta_\star\right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)Q_{n_{m_k}} \xrightarrow{\text{a.s.}} 0
\end{aligned}$$

Therefore, every subsequence of $S_n = \sup_{\theta\in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{J}}^{(n)}(\theta)-\mathcal{J}(\theta_\star)\right\|$ has a further sub-subsequence converging almost surely to 0, and hence S_n converges in probability to 0.

□

Proposition 4.5 (Sufficient conditions for Assumption 4.5)**.** *Each of the following imply Assumption 4.5.*

a) there exists a $\delta_2 > 0$ *with* $\sup_{\theta \in B_{\delta_2}(\theta_\star)} \left\| \frac{1}{n} \sum_{i \in [n]} \nabla \ell(\theta; X_i)^{\otimes 2} - \mathcal{I}(\theta) \right\| \xrightarrow{p} 0$ *and* $\mathcal{I}$ *is continuous at* $\theta_\star$*,*

b) $\{\nabla \ell(\cdot; x)^{\otimes 2} \mid x \in \mathcal{X}\}$ *is equicontinuous at* $\theta_\star$*,*

c) $\mathbb{E}\left[\left\|\nabla^{\otimes 2} \ell(\cdot; X_1)\right\|_\infty^2\right] < \infty$*,*

Proof of Proposition 4.5.

a), b) The proofs are the same as for Proposition 4.4 a), b).

c) Let $Q_n = \frac{1}{n} \sum_{i \in [n]} \left\|\nabla^{\otimes 2} \ell(\cdot; X_i)\right\|_\infty^2$, $q = \mathbb{E}\left\|\nabla^{\otimes 2} \ell(\cdot; X_1)\right\|_\infty^2$, and let $r_{\mathcal{I},n} = n^{\mathfrak{w}/2}$. By the weak law of large numbers, $Q_n \xrightarrow{\text{p}} q$, and $\widehat{\mathcal{I}}^{(n)}(\theta_\star) \xrightarrow{\text{p}} \mathcal{I}(\theta_\star)$. Starting with

$$\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)} \left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}_\star\right\| \leq \left\|\widehat{\mathcal{I}}^{(n)}(\theta_\star) - \mathcal{I}(\theta_\star)\right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \widehat{\mathcal{I}}^{(n)}(\theta_\star)\right\|,$$

we can bound the second term with a Taylor series and Cauchy-Shwarz as

$$\begin{aligned}
&\left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \widehat{\mathcal{I}}^{(n)}(\theta_\star)\right\| \\
&\quad \leq \frac{1}{n} \sum_{i \in [n]} \left\| \left(\nabla \ell\left(\theta_\star;\ X_i\right) + \int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right)^{\otimes 2} - \nabla \ell\left(\theta_\star;\ X_i\right)^{\otimes 2} \right\| \\
&\quad \leq \frac{2}{n} \sum_{i \in [n]} \left\|\nabla \ell\left(\theta_\star;\ X_i\right)\right\| \left\| \int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right\| \\
&\qquad + \frac{1}{n} \sum_{i \in [n]} \left\| \left(\int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right)^{\otimes 2} \right\| \\
&\quad \leq \frac{2}{n} \sum_{i \in [n]} \left\|\nabla \ell\left(\theta_\star;\ X_i\right)\right\| \left\|\nabla^{\otimes 2} \ell(\cdot; X_i)\right\|_\infty \left\|\theta - \theta_\star\right\| + \frac{1}{n} \sum_{i \in [n]} \left\|\nabla^{\otimes 2} \ell(\cdot; X_i)\right\|_\infty^2 \left\|\theta - \theta_\star\right\|^2 \\
&\quad \leq 2\left\|\theta - \theta_\star\right\| \sqrt{\frac{1}{n} \sum_{i \in [n]} \left\|\nabla \ell\left(\theta_\star;\ X_i\right)\right\|^2} \sqrt{\frac{1}{n} \sum_{i \in [n]} L(X_i)^2} + \left\|\theta - \theta_\star\right\|^2 Q_n \\
&\quad \leq 2\left\|\theta - \theta_\star\right\| \sqrt{\operatorname{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))} \sqrt{Q_n} + \left\|\theta - \theta_\star\right\|^2 Q_n,
\end{aligned}$$

Plugging this back in,

$$
\begin{aligned}
&\sup_{\theta\in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)} \left\|\widehat{\mathcal{I}}^{(n)}(\theta)-\mathcal{I}_\star\right\| \\
&\le \left\|\widehat{\mathcal{I}}^{(n)}(\theta_\star)-\mathcal{I}(\theta_\star)\right\| + \sup_{\theta\in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left(2\left\|\theta-\theta_\star\right\|\sqrt{\mathrm{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))}\sqrt{Q_n}+\left\|\theta-\theta_\star\right\|^2 Q_n\right) \\
&\le \left\|\widehat{\mathcal{I}}^{(n)}(\theta_\star)-\mathcal{I}(\theta_\star)\right\| + 2\left(\left\|\widehat{\theta}^{(n)}-\theta_\star\right\|+n^{-\mathfrak{w}/2}\right)\sqrt{\mathrm{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))}\sqrt{Q_n}+\left(\left\|\widehat{\theta}^{(n)}-\theta_\star\right\|+n^{-\mathfrak{w}/2}\right)^2 Q_n \\
&\xrightarrow{\mathrm{p}} 0.
\end{aligned}
$$

□

4.10 Proof of Proposition 4.1

Recall that

$$d\vartheta_t = -\frac{1}{2}B\vartheta_t\,dt + \sqrt{A}\,dW_t, \tag{4.31}$$

which implies

$$\vartheta_t = \exp(-B/2\vartheta_0) + \int_0^t \exp(-Bt/2)A^{1/2}dW_t. \tag{4.32}$$

Assuming stationarity, $\vartheta_t \sim \mathcal{N}(0, Q_\infty)$ where $Q_\infty = \int_0^\infty \exp(-Bs/2)A\exp(-Bs/2)ds$, we have

$$
\begin{aligned}
&\mathrm{Cov}(\int_0^t \vartheta_s\,ds) \\
&= \mathbb{E}(\int_0^t\int_0^t \vartheta_s\vartheta_r^T dsdr) \\
&= \int_0^t\int_0^s \mathbb{E}(\vartheta_s\vartheta_r^T)drds + \int_0^t\int_0^r \mathbb{E}(\vartheta_s\vartheta_r^T)dsdr.
\end{aligned} \tag{4.33}
$$

We focus on the first term since the second term can be written similarly:

$$
\begin{aligned}
&\int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_r^T) dr ds \\
&= \int_0^t \int_0^s \mathbb{E}\left[\left(\exp(-B(s-r)/2)\vartheta_r + \int_r^s \exp(-Bu/2) A^{1/2} dW_u\right) \vartheta_r^T\right] dr ds \\
&= \int_0^t \int_0^s \exp(-B(s-r)/2) \mathbb{E}(\vartheta_r \vartheta_r^T) dr ds \\
&= \int_0^t \int_0^s \exp(-B(s-r)/2) Q_\infty dr ds \\
&= \int_0^t -2B^{-1}(\exp(-Bs/2) - 1) Q_\infty ds \\
&= \left[4B^{-2}(\exp(-Bt/2) - 1) + 2tB^{-1}\right] Q_\infty.
\end{aligned}
\tag{4.34}
$$

We can write $\int_0^t \int_0^r \mathbb{E}(\vartheta_s \vartheta_r^T) ds dr$ similarly and combine the two results

$$
\begin{aligned}
&\operatorname{Cov}\left(\bar{\vartheta}_t\right) = \frac{1}{t^2} \operatorname{Cov}(\int_0^t \vartheta_s \, ds) \\
&= \frac{1}{t^2}\left[\int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_r^T) dr ds + \int_0^t \int_0^r \mathbb{E}(\vartheta_s \vartheta_r^T) ds dr\right] \\
&= \frac{4}{t} \operatorname{Sym}\left(B^{-1} Q_\infty\right) - \frac{8}{t^2} \operatorname{Sym}\left(B^{-2}\left\{I - e^{-tB/2}\right\} Q_\infty\right),
\end{aligned}
\tag{4.35}
$$

which completes the proof.

4.11 Sketch Proof of Scaling Limit for SGLD with Control Variates

We argue that the mini-batch noise is always lower order for SGLD with control variates by showing that the corresponding $[2.\ell\ell]^{(n_m)}(\vartheta)$ from the proof of Theorem 4.1 in Section 4.7

is vanishing under any scaling limit where the drift term [1.ℓ] does not vanish.

$$\begin{aligned}
&\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta)\right\rangle}_{[2.\ell\ell]^{(n_m)}(\vartheta)}\\
&= n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\frac{1}{2}\nabla^{\otimes 2}f(\vartheta) : \left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\\
&= n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\frac{1}{2}\Gamma\nabla^{\otimes 2}f(\vartheta) : \left(\frac{hw^{(n)}\Gamma}{2b^{(n)}}\sum_{j\in\left[b^{(n)}\right]}\left(\nabla\ell\left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1}\vartheta;\ X_{I_1^{(n)}(j)}\right) - \nabla\ell\left(\widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)}\right)\right)\right)^{\otimes 2}\\
&= \frac{c_h^2}{c_b^2}n_m^{\mathfrak{a}-2\mathfrak{h}+2\mathfrak{w}-2\mathfrak{b}}\frac{1}{2}\Gamma\nabla^{\otimes 2}f(\vartheta)\Gamma^\top\\
&\qquad : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left(\sum_{j\in\left[b^{(n)}\right]}\left(\nabla\ell\left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1}\vartheta;\ X_{I_1^{(n)}(j)}\right) - \nabla\ell\left(\widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)}\right)\right)\right)^{\otimes 2}\\
&\approx \frac{c_h^2}{c_b^2}n_m^{\mathfrak{a}-2\mathfrak{h}+2\mathfrak{w}-2\mathfrak{b}}\frac{1}{2}\Gamma\nabla^{\otimes 2}f(\vartheta)\Gamma^\top : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left(\sum_{j\in\left[b^{(n)}\right]}\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)}\right)(w^{(n)})^{-1}\vartheta\right)^{\otimes 2}\\
&= \frac{c_h^2}{c_b^2}n_m^{\mathfrak{a}-2\mathfrak{h}-2\mathfrak{b}}\frac{1}{2}\Gamma\nabla^{\otimes 2}f(\vartheta)\Gamma^\top : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left(\sum_{j\in\left[b^{(n)}\right]}\nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)}\right)\vartheta\right)^{\otimes 2}\\
&\approx n_m^{\mathfrak{a}-2\mathfrak{h}-2\mathfrak{b}}\frac{1}{2}\Gamma\nabla^{\otimes 2}f(\vartheta)\Gamma^\top : \left[b^{(n)}(b^{(n)}-1)\mathcal{J}_\star\vartheta\vartheta^\top\mathcal{J}_\star + b^{(n)}K(\theta_\star;\vartheta)\right]
\end{aligned}$$

where $K(\theta_\star;\vartheta) = \int \nabla^{\otimes 2}\ell(\theta_\star;x)\ \vartheta^{\otimes 2}\ \nabla^{\otimes 2}\ell(\theta_\star;x)P(dx)$.

Now, we recall that for the drift term to be non-zero in the limit, we need $\mathfrak{a} = \mathfrak{h}$. However, at any such scaling the $[2.\ell\ell]^{(n_m)}(\vartheta)$ term is $\mathcal{O}(n^{-\mathfrak{h}-2\mathfrak{b}})$, and so is never not 0 in the limit.

4.12 Sketch Proof for constrained parameter spaces

The key idea is that, if $\theta_\star \in \text{interior}(\Theta)$, there is a $r > 0$ with $\theta_\star \in B(\theta_\star, r) \subset \text{interior}(\Theta)$, and for any compactly supported test function f and compact extension of its support, K_1, for sufficiently large sample sizes n, $K_1 \subseteq B(0, w^{(n)}r)$. In the proof of the $\Theta = \mathbb{R}^d$ case we found that, along sub-sequences (n_{m_k}), the increments from the log-likelihood and from the prior vanish uniformly within a sufficiently large extension of the support of f. Combining this with faithfulness of P (defined in Section 4.4.3) and an application of the Lebesgue dominated convergence theorem to handle truncation of the Gaussian increments shows that the $A_{n_m} f \to Af$ uniformly within the extension of the support of f when $\Theta \neq \mathbb{R}^d$. Moreover, the local property of the boundary condition (defined in Section 4.4.3) ensures that for sufficiently large sample sizes, if the process were far enough outside of the support of f then it cannot re-enter the support via an arbitrarily large jump caused by the boundary condition. Thus, outside of the extension of the support of f, the deviation of $A_{n_m} f$ from 0 is essentially indistinguishable from the unconstrained case. Using those two facts we can rely on the faithfulness of the boundary dynamics to ensure that the process converges weakly to the same Ornstein-Uhlenbeck limit as in the unconstrained case.

Bibliography

[1] L. Aitchison. *A statistical theory of cold posteriors in deep neural networks.* 2020. arXiv: 2008.05912.

[2] P. Alquier, N. Friel, R. Everitt, and A. Boland. "Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels". *Statistics and Computing* 26.1–2 (2016), pp. 29–47.

[3] W. An, H. Wang, Q. Sun, J. Xu, Q. Dai, and L. Zhang. "A pid controller approach for stochastic optimization of deep networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018, pp. 8522–8531.

[4] C. Andrieu and G. O. Roberts. "The pseudo-marginal approach for efficient Monte Carlo computations". *The Annals of Statistics* 37.2 (2009), pp. 697–725.

[5] Y. F. Atchadé. "Approximate spectral gaps for Markov chain mixing times in high dimensions". *SIAM Journal on Mathematics of Data Science* 3.3 (2021), pp. 854–872.

[6] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. "Control variates for stochastic gradient MCMC". *Statistics and Computing* 29.3 (2019), pp. 599–615.

[7] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators.* Vol. 348. Springer Science & Business Media, 2013.

[8] D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators.* Vol. 103. Springer, 2014.

[9] P. H. Baxendale. “Renewal theory and computable convergence rates for geometrically ergodic Markov chains”. *The Annals of Applied Probability* 15.1B (2005), pp. 700–738.

[10] J. R. Baxter and J. S. Rosenthal. “Rates of convergence for everywhere-positive Markov chains”. *Statistics & probability letters* 22.4 (1995), pp. 333–338.

[11] M. Bédard. “Weak Convergence of Metropolis Algorithms for Non-IID Target Distributions”. *The Annals of Applied Probability* (2007), pp. 1222–1244.

[12] M. Bédard. *On the robustness of optimal scaling for random walk Metropolis algorithms*. Vol. 68. 01. 2006.

[13] M. Bédard and J. S. Rosenthal. “Optimal scaling of Metropolis algorithms: Heading toward general target distributions”. *Canadian Journal of Statistics* 36.4 (2008), pp. 483–503.

[14] P. G. Bissiri, C. C. Holmes, and S. G. Walker. “A general framework for updating belief distributions”. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 78.5 (2016), p. 1103.

[15] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[16] N. E. Breslow and D. G. Clayton. “Approximate inference in generalized linear mixed models”. *Journal of the American Statistical Association* 88.421 (1993), pp. 9–25.

[17] L. Breyer, G. Roberts, and J. Rosenthal. “A note on geometric ergodicity and floating–point roundoff error”. *Statistics and Probability Letters* 53 (2001), pp. 123–127.

[18] S. Brooks, A. Gelmand, G. L. Jones, and X.-.-L. Meng, eds. *Handbook of Markov chain Monte Carlo*. Chapman & Hall, 2011.

[19] N. Brosse, A. Durmus, and E. Moulines. “The promises and pitfalls of Stochastic Gradient Langevin Dynamics”. In: *Advances in Neural Information Processing Systems*. 2018.

[20] P. Bühlmann. "Discussion of big Bayes stories and BayesBag". *Statistical science* 29.1 (2014), pp. 91–94.

[21] P. Bühlmann and S. van de Geer. "High-dimensional inference in misspecified linear models". *Electronic Journal of Statistics* 9.1 (2015), pp. 1449–1473.

[22] O. Bunke and X. Milhaud. "Asymptotic behavior of Bayes estimates under possibly incorrect models". *The Annals of Statistics* 26.2 (1998), pp. 617–644.

[23] T. Campbell and T. Broderick. "Automated scalable Bayesian inference via Hilbert coresets". *The Journal of Machine Learning Research* 20.1 (2019), pp. 551–588.

[24] C.-F. Chen. "On asymptotic normality of limiting density functions with Bayesian implications". *Journal of the Royal Statistical Society: Series B (Methodological)* 47.3 (1985), pp. 540–546.

[25] L. H. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

[26] S. Cyrus, B. Hu, B. van Scoy, and L. Lessard. "A robust accelerated optimization algorithm for strongly convex functions". In: *2018 Annual American Control Conference (ACC)*. IEEE. 2018, pp. 1376–1381.

[27] A. Dieuleveut, A. Durmus, F. Bach, et al. "Bridging the gap between constant step size stochastic gradient descent and markov chains". *Annals of Statistics* 48.3 (2020), pp. 1348–1382.

[28] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer, 2018.

[29] A. Duncan, N. Nüsken, and G. Pavliotis. "Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions". *Journal of Statistical Physics* 169.6 (2017), pp. 1098–1131.

[30] A. Durmus and E. Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". *Bernoulli* 25.4A (2019), pp. 2854–2882.

[31] A. Durmus and E. Moulines. "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm". *The Annals of Applied Probability* 27.3 (2017), pp. 1551–1587.

[32] R. Durrett. *Probability: theory and examples*. 4th. Vol. 49. Cambridge university press, 2019.

[33] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*. Vol. 282. John Wiley & Sons, 2009.

[34] H. Federer. *Geometric measure theory*. Springer, 1969.

[35] D. Ferré, L. Hervé, and J. Ledoux. "Regular perturbation of V–geometrically ergodic Markov chains". *Journal of Applied Probability* 50.1 (2013), pp. 184–194.

[36] A. Gibbs. "Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration". *Stochastic Models* 20.4 (2004), pp. 473–492.

[37] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. "Measuring sample quality with diffusions". *The Annals of Applied Probability* 29.5 (2019), pp. 2884–2928.

[38] P. Grünwald and T. van Ommen. "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it". *Bayesian Analysis* 12.4 (2017), pp. 1069–1103.

[39] A. Gupal and L. Bazhenov. "Stochastic analog of the conjugant-gradient method". *Cybernetics* 8.1 (1972), pp. 138–140.

[40] A. Gupta, H. Chen, J. Pi, and G. Tendolkar. "Some Limit Properties of Markov Chains Induced by Recursive Stochastic Algorithms". *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 967–1003.

[41] L. Hervé and J. Ledoux. "Approximating Markov chains and V–geometric ergodicity via weak perturbation theory". *Stochastic Processes and their Applications* 124.1 (2014), pp. 613–638.

[42] J. P. Hobert and C. J. Geyer. "Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model". *Journal of Multivariate Analysis* 67.2 (1998), pp. 414–430.

[43] P. J. Huber. “The behavior of maximum likelihood estimates under nonstandard conditions”. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability: Weather modification*. Vol. 5. Univ of California Press. 1967, p. 221.

[44] J. Huggins, T. Campbell, and T. Broderick. “Coresets for scalable Bayesian logistic regression”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 4080–4088.

[45] J. H. Huggins and J. W. Miller. *Using bagged posteriors for robust inference and model criticism*. 2019. arXiv: `1912.07104`.

[46] L. Isserlis. “On Certain Probable Errors and Correlation Coefficients of Multiple Frequency Distributions with Skew Regression”. *Biometrika* 11.3 (1916), pp. 185–190. ISSN: 00063444.

[47] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. *What Are Bayesian Neural Network Posteriors Really Like?* 2021. arXiv: `2104.14421`.

[48] N. Jain and B. Jamison. “Contributions to Doeblin’s theory of Markov processes”. *Probability Theory and Related Fields* 8.1 (1967), pp. 19–40.

[49] A. Jakubowski. “On the Skorokhod topology”. *Ann. Inst. H. Poincaré Probab. Statist* 22.3 (1986), pp. 263–285.

[50] J. E. Johndrow and J. C. Mattingly. *Coupling and decoupling to bound an approximating Markov chain*. 2017. arXiv: `1706.02040`.

[51] J. E. Johndrow and J. C. Mattingly. *Error bounds for approximations of Markov chains used in Bayesian sampling*. 2017. arXiv: `1711.05382`.

[52] J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. *Approximations of Markov chains and high–dimensional Bayesian inference*. 2015. arXiv: `1508.03387v1`.

[53] O. Kallenberg. *Foundations of modern probability*. Springer Science & Business Media, 2006.

[54] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Vol. 113. springer, 2014.

[55] R. E. Kass, L. Tierney, and J. B. Kadane. "The validity of posterior expansions based on Laplace's method". *Bayesian and Likelihood Methods in Statistics and Econometrics* 7 (1990), p. 473.

[56] G. Keller and C. Liverani. "Stability of the spectrum for transfer operators". *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* 28.1 (1999), pp. 141–152.

[57] R. Kidambi, P. Netrapalli, P. Jain, and S. Kakade. "On the insufficiency of existing momentum schemes for stochastic optimization". In: *2018 Information Theory and Applications Workshop (ITA)*. IEEE. 2018, pp. 1–9.

[58] B. J. K. Kleijn and A. W. van der Vaart. "The Bernstein-von-Mises theorem under misspecification". *Electronic Journal of Statistics* 6 (2012), pp. 354–381.

[59] H. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*. Vol. 35. Springer Science & Business Media, 2003.

[60] M. Lalancette. "Convergence d'un algorithme de type Metropolis pour une distribution cible bimodale" (2017).

[61] Z. Landsman. "On the generalization of Stein's Lemma for elliptical class of distributions". *Statistics & probability letters* 76.10 (2006), pp. 1012–1016.

[62] L. Lessard, B. Recht, and A. Packard. "Analysis and design of optimization algorithms via integral quadratic constraints". *SIAM Journal on Optimization* 26.1 (2016), pp. 57–95.

[63] Q. Li, C. Tai, and E. Weinan. "Stochastic Modified Equations and Dynamics of Stochastic Gradient Algorithms I: Mathematical Foundations." *J. Mach. Learn. Res.* 20 (2019), pp. 40–1.

[64] Z. Li, K. Lyu, and S. Arora. *Reconciling Modern Deep Learning with Traditional Optimization Analyses: The Intrinsic Learning Rate*. 2020. arXiv: `2010.02916`.

[65] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[66] K. Liu, L. Ziyin, and M. Ueda. “Noise and Fluctuation of Finite Learning Rate Stochastic Gradient Descent”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 7045–7056.

[67] S. Livingstone, M. Betancourt, S. Byrne, M. Girolami, et al. “On the geometric ergodicity of Hamiltonian Monte Carlo”. *Bernoulli* 25.4A (2019), pp. 3109–3138.

[68] J. Ma and D. Yarats. “Quasi-hyperbolic momentum and Adam for deep learning”. In: *International Conference on Learning Representations*. 2018.

[69] F. Maggi. *Sets of finite perimeter and geometric variational problems: an introduction to Geometric Measure Theory*. 135. Cambridge University Press, 2012.

[70] S. Mandt, M. D. Hoffman, and D. M. Blei. “Stochastic gradient descent as approximate Bayesian inference”. *The Journal of Machine Learning Research* 18.1 (2017), pp. 4873–4907.

[71] J. C. Mattingly, N. S. Pillai, and A. M. Stuart. “Diffusion limits of the random walk Metropolis algorithm in high dimensions”. *The Annals of Applied Probability* 22.3 (2012), pp. 881–930.

[72] C. E. McCulloch and J. M. Neuhaus. “Generalized Linear Mixed Models”. *Encyclopedia of Biostatistics* 4 (2005).

[73] F. Medina–Aguayo, D. Rudolf, and N. Schweizer. “Perturbation bounds for Monte Carlo within Metropolis via restricted approximations”. *Stochastic Processes and their Applications* (2019).

[74] F. J. Medina–Aguayo, A. Lee, and G. O. Roberts. “Stability of noisy metropolis–hastings”. *Statistics and Computing* 26.6 (2016), pp. 1187–1211.

[75] G. Metafune, D. Pallara, and E. Priola. “Spectrum of Ornstein-Uhlenbeck operators in Lp spaces with respect to invariant measures”. *Journal of Functional Analysis* 196.1 (2002), pp. 40–60.

[76] C. Mingard, G. Valle-Perez, J. Skalse, and A. A. Louis. “Is SGD a Bayesian sampler? Well, almost”. *Journal of Machine Learning Research* 22.79 (2021), pp. 1–64.

[77] A. Y. Mitrophanov. “Sensitivity and convergence of uniformly ergodic Markov chains”. *Journal of Applied Probability* (2005), pp. 1003–1014.

[78] E. Moulines and F. Bach. “Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011.

[79] U. K. Müller. “Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix”. *Econometrica* 81.5 (2013), pp. 1805–1849.

[80] T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. 2017. arXiv: `1706.02692`.

[81] P. Neal and G. Roberts. “Optimal scaling for partially updating MCMC algorithms”. *The Annals of Applied Probability* 16.2 (2006), pp. 475–515.

[82] J. Negrea. *Optimal Scaling and Shaping of Random Walk Metropolis via Diffusion Limits of Block-IID Targets*. Available Online. 2019. arXiv: `1902.06603`.

[83] J. Negrea and J. S. Rosenthal. “Approximations of geometrically ergodic reversible Markov chains”. *Advances in Applied Probability* 53.4 (2021). Available Online, pp. 981–1022.

[84] J. Negrea, J. Yang, H. Feng, D. M. Roy, and J. H. Huggins. *Statistical Inference with Stochastic Gradient Algorithms*. Available Online. 2021.

[85] C. Nemeth and P. Fearnhead. “Stochastic gradient markov chain monte carlo”. *Journal of the American Statistical Association* 116.533 (2021), pp. 433–450.

[86] E. Nummelin and R. L. Tweedie. “Geometric ergodicity and R–positivity for general Markov chains”. *The Annals of Probability* (1978), pp. 404–420.

[87] D. Panchenko. *The Sherrington-Kirkpatrick model*. Springer Science & Business Media, 2013.

[88] G. A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*. Vol. 60. Springer, 2014.

[89] N. S. Pillai and A. Smith. *Ergodicity of approximate MCMC chains with applications to large data sets*. 2014. arXiv: 1405.0182.

[90] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions". *The Annals of Applied Probability* 22.6 (2012), pp. 2320–2356.

[91] M. Pollock, P. Fearnhead, A. M. Johansen, and G. O. Roberts. "Quasi-stationary Monte Carlo and the ScaLE algorithm". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5 (2020), pp. 1167–1221.

[92] B. T. Polyak and A. B. Juditsky. "Acceleration of stochastic approximation by averaging". *SIAM journal on control and optimization* 30.4 (1992), pp. 838–855.

[93] H. Robbins and S. Monro. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[94] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". *The annals of applied probability* 7.1 (1997), pp. 110–120.

[95] G. O. Roberts, J. S. Rosenthal, et al. "General state space Markov chains and MCMC algorithms". *Probability Surveys* 1 (2004), pp. 20–71.

[96] G. O. Roberts and J. S. Rosenthal. "Geometric ergodicity and hybrid Markov chains". *Electronic Communications in Probability* 2.2 (1997), pp. 13–25.

[97] G. O. Roberts and J. S. Rosenthal. "Optimal scaling for various Metropolis-Hastings algorithms". *Statistical science* 16.4 (2001), pp. 351–367.

[98] G. O. Roberts and J. S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[99] G. O. Roberts, J. S. Rosenthal, and P. O. Schwartz. "Convergence properties of perturbed Markov chains". *Journal of Applied Probability* (1998), pp. 1–11.

[100] G. O. Roberts and R. L. Tweedie. "Geometric L2 and L1 convergence are equivalent for reversible Markov chains". *Journal of Applied Probability* (2001), pp. 37–41.

[101] R. Royall and T.-S. Tsou. "Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 391–404.

[102] W. Rudin. *Functional Analysis*. 2nd ed. International series in pure and applied mathematics. McGraw-Hill, 1991. ISBN: 0070542368,9780070542365.

[103] D. Rudolf. *Explicit error bounds for Markov chain Monte Carlo*. Dissertationes Math. 485 (2012), 93 pp. 2011. arXiv: 1108.3201.

[104] D. Rudolf and N. Schweizer. *Perturbation theory for Markov chains via Wasserstein distance*. 2015. arXiv: 1503.04123.

[105] S. M. Schmon, G. Deligiannidis, A. Doucet, and M. K. Pitt. "Large-sample asymptotics of the pseudo-marginal method". *Biometrika* 108.1 (2021), pp. 37–51.

[106] S. M. Schmon and P. Gagnon. *Optimal scaling of random walk Metropolis algorithms using Bayesian large-sample asymptotics*. 2021. arXiv: 2104.06384.

[107] C. R. Shalizi. "Dynamics of Bayesian updating with dependent data and misspecified models". *Electronic Journal of Statistics* 3 (2009), pp. 1039–1074.

[108] J. Sirignano and K. Spiliopoulos. "Stochastic gradient descent in continuous time". *SIAM Journal on Financial Mathematics* 8.1 (2017), pp. 933–961.

[109] J. Sirignano and K. Spiliopoulos. "Stochastic gradient descent in continuous time: A central limit theorem". *Stochastic Systems* 10.2 (2020), pp. 124–151.

[110] R. L. Smith and L. Tierney. *Exact transition probabilities for the independence Metropolis sampler*. 1996. Author's Website: http://www.rls.sites.oasis.unc.edu/postscript/rs/exact.pdf.

[111] J. E. Stafford. "A robust adjustment of the profile likelihood". *The Annals of Statistics* 24.1 (1996), pp. 336–352.

[112] 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, Volume 2: Probability Theory*. The Regents of the University of California. 1972.

[113] C. Stein, P. Diaconis, S. Holmes, and G. Reinert. "Use of exchangeable pairs in the analysis of simulations". In: *Stein's Method*. Institute of Mathematical Statistics, 2004, pp. 1–25.

[114] T. Tao. *An introduction to measure theory*. American Mathematical Society Providence, RI, 2011.

[115] Y. W. Teh, A. H. Thiery, and S. J. Vollmer. "Consistency and fluctuations for stochastic gradient Langevin dynamics". *The Journal of Machine Learning Research* 17.1 (2016), pp. 193–225.

[116] L. Tierney and J. B. Kadane. "Accurate approximations for posterior moments and marginal densities". *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86.

[117] P. Toulis and E. M. Airoldi. "Asymptotic and finite-sample properties of estimators based on stochastic gradients". *The Annals of Statistics* 45.4 (2017), pp. 1694–1727.

[118] B. Tzen, T. Liang, and M. Raginsky. "Local Optimality and Generalization Guarantees for the Langevin Algorithm via Empirical Metastability". In: *Conference On Learning Theory*. 2018, pp. 857–875.

[119] B. van Scoy, R. A. Freeman, and K. M. Lynch. "The fastest known globally convergent first-order method for minimizing strongly convex functions". *IEEE Control Systems Letters* 2.1 (2017), pp. 49–54.

[120] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. "Exploration of the (non-) asymptotic bias and variance of stochastic gradient Langevin dynamics". *The Journal of Machine Learning Research* 17.1 (2016), pp. 5504–5548.

[121] M. Welling and Y. W. Teh. "Bayesian learning via stochastic gradient Langevin dynamics". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer. 2011, pp. 681–688.

[122] F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. *How good is the Bayes posterior in deep neural networks really?* 2020. arXiv: 2002.02405.

[123] H. White. "Maximum likelihood estimation of misspecified models". *Econometrica: Journal of the econometric society* (1982), pp. 1–25.

[124] J. Wu, D. Zou, V. Braverman, and Q. Gu. *Direction Matters: On the Implicit Regularization Effect of Stochastic Gradient Descent with Moderate Learning Rate*. 2020. arXiv: 2011.02538.

[125] L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. "An analysis of constant step size sgd in the non-convex regime: Asymptotic normality and bias". *Advances in Neural Information Processing Systems* 34 (2021).

[126] G. Zanella, M. Bédard, and W. S. Kendall. "A Dirichlet form approach to MCMC optimal scaling". *Stochastic Processes and their Applications* 127.12 (2017), pp. 4053–4082.

[127] P. Zhou, J. Feng, C. Ma, C. Xiong, and S. Hoi. *Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning*. 2020. arXiv: 2010.05627.