Statistical Inference with Stochastic Gradient Algorithms

Jeffrey Negrea\(^1\), Jun Yang\(^2\), Haoyue Feng\(^3\), Daniel M. Roy\(^1\), and Jonathan H. Huggins\(^3\)

\(^1\)Department of Statistical Sciences, University of Toronto, Canada
\(^2\)Department of Statistics, Oxford University, UK
\(^3\)Department of Mathematics and Statistics, Boston University, USA

July 26, 2022

Abstract

Stochastic gradient algorithms are widely used for both optimization and sampling in large-scale learning and inference problems. However, in practice, tuning these algorithms is typically done using heuristics and trial-and-error rather than rigorous, generalizable theory. To address this gap between theory and practice, we novel insights into the effect of tuning parameters by characterizing the large-sample behavior of iterates of a very general class of preconditioned stochastic gradient algorithms with fixed step size. In the optimization setting, our results show that iterate averaging with a large fixed step size can result in statistically efficient approximation of the (local) M-estimator. In the sampling context, our results show that with appropriate choices of tuning parameters, the limiting stationary covariance can match either the Bernstein–von Mises limit of the posterior, adjustments to the posterior for model misspecification, or the asymptotic distribution of the MLE; and that with a naive tuning the limit corresponds to none of these. Moreover, we argue that an essentially independent sample from the stationary distribution can be obtained after a fixed number of passes over the dataset. We validate our asymptotic results in realistic finite-sample regimes via several experiments using simulated and real data. Overall, we demonstrate that properly tuned stochastic gradient algorithms with constant step size offer a computationally efficient and statistically robust approach to obtaining point estimates or posterior-like samples.

1 Introduction

Stochastic gradient algorithms were originally proposed as optimization and root finding methods by Robbins and Monro \cite{Robbins1951}, and have become the standard approach to large-scale optimization in statistics and machine learning. This success is because improvements in computational complexity from subsampling outweigh the accuracy loss from stochastic approximation for empirical objectives. Hence, stochastic optimization methods scale more favourably with the sample size and model complexity than their deterministic counterparts \cite{Moulines2011, Goodfellow2016}. Over the past decade, this scalability has also lead to tremendous growth in the use stochastic gradient Markov chain Monte Carlo sampling algorithms, particularly in machine learning \cite{Welling2011, Nemeth2021}.

Most analyses of stochastic gradient optimization procedures such as stochastic gradient descent (SGD) focus on the parameter error or the optimality gap \cite[e.g.,][]{Moulines2011, Kushner2003, Nemirovski2009, Reddi2018}, while analyses of stochastic gradient sampling procedures such as stochastic gradient Langevin dynamics (SGLD) focus on how well the empirical distribution of the iterates approximates the posterior \cite[e.g.,][]{Teh2016, Vollmer2016, Brosse2018, Baker2019, Nemeth2021}. Yet in practice stochastic gradient algorithms are 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.

We take a step toward closing this gap between theory and practice when the step size is fixed across iterations (but decreases with the sample size). The fixed-step size setting proves to be practically relevant for optimization because convergence to a near-optimum is rapid and robust to the precise step size choice...
while for sampling, using a fixed step size leads to better mixing time behaviour: the number of iterations until the next approximately independent sample is constant, unlike in the decreasing-step size regime where the number of iterations until the next approximately independent sample increases without bound.

Our main result characterizes the statistical scaling limits of stochastic gradient algorithms as the sample size tends to infinity. We show that the sample paths of a very general class of preconditioned stochastic gradient algorithms – including stochastic gradient descent with and without additional Gaussian noise, momentum, and/or acceleration – converge to the sample paths of an Ornstein–Uhlenbeck process under relatively mild conditions. For optimization, we prove that iterate averaging of SGD with a large step size provides statistically optimal estimation of the MLE, complementing the results of Polyak and Juditsky [1992] which hold in the fixed-data setting with decreasing step size. Notably, however, while the asymptotic guarantees in the decreasing step size case often require an impractically large number of iterates, numerical experiments show that our constant step size averaging result tends to hold even after a single pass over the dataset. For sampling, we show that it is more computationally efficient (in terms of mixing) to use SGLD to sample from a distribution that is far from the posterior but has better robustness to model misspecification. This result suggests that stochastic gradients, rather than being a problem in need of solution, can be seen as a beneficial tool.

The assumptions required by our analysis are quite weak. We allow the batch size used to compute the stochastic gradient to be constant or 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. [2017], 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 Yang [1993], Kushner and Yin [2003], who provides the basis for the assumptions in Mandt et al. [2017], 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 cases where either (i) the objective function is itself stochastic (in particular arising as the random loss function based on a finite IID sample) or (ii) the joint scaling of the heuristic arguments of Mandt et al. [2017], and open the way for further generalizations to situations where the asymptotic statistical results.

1.1 Notation

Let $\mathcal{M}_{1,+}(\mathcal{X})$ denote the set of probability measures on the observation space $\mathcal{X}$ and let $\mathbb{N} := \{1, 2, \ldots\}$ denote the natural numbers. For $n \in \mathbb{N}$, define $[n] := \{1, \ldots, n\}$. For $d \in \mathbb{N}$, denote the $d$-dimensional Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and (positive semi-definite) covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ by $N_d(\mu, \Sigma)$. 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 $\xi$, 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 $f$, $E(f(\xi_k)) \to E(f(\xi))$ as $k \to \infty$. We write $L(\xi)$ for the distribution (law) of a random element $\xi$, and $\mathcal{L}(\xi)$ for the conditional distribution of $\xi$ given another random element $\nu$. For a square matrix $M$, define the symmetrization operator as $\text{Sym}(M) := (M + M^\top)/2$. For a function $f : \mathcal{A} \to L$ with $\mathcal{A}$ a set and $(L, \|\cdot\|)$ a normed linear space, define $\|f\|_\infty := \sup_{a \in \mathcal{A}} \|f(a)\|$.

2 Stochastic Gradient Optimization and Sampling

Let $X^{(n)} = (X_i)_{i=1}^n \in \mathcal{X}^n$ denote a dataset with observations $X_i$ independently and identically distributed (i.i.d.) from an unknown distribution $P$. For parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, consider the potential $U^{(n)}(\theta) := r(\theta) + \sum_{i=1}^n \ell_i(\theta; X_i)$, where typically $\ell$ represents a log-likelihood or a negative loss function, and $r(\theta)$ represents a regularizer or a (possibly improper) log prior density $\log \pi(\theta)$ that is everywhere positive on $\Theta$. 

[1992]
If $-\mathcal{U}^{(n)}(\theta)$ is interpreted as a (possibly regularized) loss, perhaps the most popular estimator for the (locally) optimal population parameter $\theta$, satisfying $\mathbb{E}\{\nabla \ell(\theta_0; X_1)\} = 0$, is the M-estimator $\hat{\theta}^{(n)}$ satisfying the first-order optimality condition $\nabla \mathcal{U}^{(n)}(\hat{\theta}^{(n)}) = 0$. If $-\mathcal{U}^{(n)}(\theta)$ is interpreted as the the negative log of the joint model density or as a generalized Bayesian loss [Bissiri et al., 2016], the quantity of interest is (usually) an expectation with respect to the (generalized) 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 the computational cost of evaluating $\ell(\theta; X_i)$, classical optimization methods for approximating $\hat{\theta}^{(n)}$ (e.g., gradient descent or Newton–Raphson) and sampling methods for estimating $\pi^{(n)}(f)$ (e.g., Metropolis–Hastings algorithms) become computationally prohibitive.

Stochastic gradient algorithms provide a means of reducing the per-iteration computational cost of optimization and sampling methods. 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 used instead to compute an unbiased gradient estimate

$$\hat{G}_k^{(n)} := \frac{1}{n} \nabla \ell \left( \theta_k^{(n)} \right) + \frac{1}{b^{(n)}} \sum_{j=1}^{b^{(n)}} \nabla \ell \left( \theta_k^{(n)}; X_{I_k^{(n)}(j)} \right),$$

(1)

where $(I_k^{(n)})_{k \in \mathbb{N}} \in \{[m]^b\}^\mathbb{N}$ 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.\(^1\)

### 2.1 Optimization

The canonical approach is stochastic gradient descent (SGD), which has the one-step update

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h_k^{(n)}}{2} \hat{G}_k^{(n)},$$

(2)

where $(h_k^{(n)})_{k \in \mathbb{N}}$ is a sequence positive step sizes. While optimal tuning of the last-iterate challenge is still challenging, averaging the iterates can provide automatic optimal uncertainty quantification [Polyak and Juditsky, 1992, Kushner and Yang, 1993, Kushner and Yin, 2003]. More precisely, when $h_k \sim k^{-\varsigma}$ for $\varsigma \in (0, 1)$, the iterate average $\bar{\theta}_k^{(n)} := \frac{1}{k} \sum_{k'=1}^{k} \theta_k^{(n)}$, satisfies

$$\lim_{n \to \infty} \lim_{k \to \infty} k \text{Cov}(\bar{\theta}_k^{(n)}) = \mathcal{J}_{\varsigma}^{-1} \mathcal{I}_{\varsigma} \mathcal{J}_{\varsigma}^{-1} = \lim_{n \to \infty} n \text{Cov}(\hat{\theta}^{(n)}),$$

(3)

where $\mathcal{I}_{\varsigma} := \mathbb{E}\{\nabla_{\theta} \ell(\theta_0; X) \otimes \nabla_{\theta} \ell(\theta_0; X)\}$ and $\mathcal{J}_{\varsigma} := -\mathbb{E}\{\nabla_{\theta}^2 \ell(\theta_0; X)\}$ are, respectively, the first- and second-order Fisher information matrices. Such results are, however, very sensitive to the choice of step size schedule, leading to impractically slow convergence rates in practice [Moulines and Bach, 2011, Toulis et al., 2021].

### 2.2 Sampling

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 led the use of approximate, unadjusted stochastic gradient MCMC methods based on discretizations of continuous time stochastic processes [Welling and Teh, 2011, Ahn et al., 2012, Chen et al., 2014, Ma et al., 2015, Baker et al., 2019, Nemeth and Fearnhead, 2021]. For example, stochastic gradient Langevin dynamics [SGLD; Welling and Teh, 2011] has one-step update given by

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h_k^{(n)}}{2} \hat{G}_k^{(n)} + \sqrt{\frac{h_k^{(n)}}{\beta}} \xi_k,$$

(4)

\(^1\)“With replacement” means that $(I_k^{(n)})_{k \in \mathbb{N}} \overset{\text{iid}}{\sim} \text{Unif}([n]^b)$ and “without replacement” means that $(I_k^{(n)})_{k \in \mathbb{N}} \overset{\text{iid}}{\sim} \text{Unif}([n]^b)$ s.t. $j_1 \neq j_2 \implies I(j_1) \neq I(j_2))$. 

3
where $\xi_k \sim \mathcal{N}_d(0, I)$ is independent standard Gaussian noise and $\beta \in (0, \infty)$ is the inverse temperature, which is usually taken to be $n$. While these methods may be asymptotically exact when run with decreasing step sizes, with the optimal choice of step sizes results in a slow $k^{-1/3}$ convergence rate [Teh et al., 2016, Vollmer et al., 2016]. Further, these results do not directly guarantee finite-time accuracy [Brosse et al., 2018].

3 Stochastic gradient algorithms and their scaling limits

In this section we develop a comprehensive framework that accurately predicts the large-sample behaviour of stochastic gradient algorithms with fixed step sizes for inference and parameter estimation, including in cases where the model is misspecified. We develop our methods and theory in the framework of a stochastic gradient meta-algorithm with one-step update

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h(n)}{2} \Gamma_k^{(n)} + \sqrt{\frac{h(n)\Lambda}{\beta(n)}} \xi_k,$$

(5)

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, $\xi_k$ are i.i.d. $\mathcal{N}_d(0, I_d)$, and $G_k^{(n)}$ implicitly depends on the batch size $b^{(n)}$ (which in turn may vary with the sample size $n$). Unless otherwise noted take the parameter space $\Theta = \mathbb{R}^d$. The meta-algorithm subsumes the SGD and SGLD algorithms described in Section 2. It also includes momentum-based methods; the details in the case of the underdamped stochastic Langevin dynamics are given in Section 5.1.

3.1 Scaling limit of the stochastic gradient meta-algorithm

We now characterize the behaviour of the sample path of the iterates of Eq. (5) in the region about $\hat{\theta}^{(n)}$, which will enable us 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

$$\hat{\varphi}_t^{(n)} := w^{(n)} \left( \theta_{[\alpha^{(n)} t]}^{(n)} - \hat{\theta}^{(n)} \right),$$

(6)

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: $h(n) = c_h n^{-b}$ for $b > 0$, $b^{(n)} = \lceil c_b n^b \rceil$ for $b \geq 0$, and $\beta^{(n)} = c_\beta n^t$ for $t \in \mathbb{R}$. Given these tuning parameters, in order to have a stable and non-trivial limit, we must take the time scaling to be $\alpha^{(n)} = n^b$ and the spatial scaling to be $w^{(n)} = n^m$ for $m = \min \{ b + h, t \} / 2$. In this setting we have the following result, under Assumptions 1 to 5 discussed in Section 3.2.

**Theorem 1** (Scaling limit of the meta-algorithm). If Assumptions 1 to 5 all hold, and there exists $\theta_* \in \Theta$ such that both $\hat{\theta}^{(n)} \overset{P}{\to} \theta_*$ and $\hat{\varphi}_0^{(n)} \overset{\mathcal{D}}{\to} \varphi_0$, then

$$\left( \hat{\theta}_t^{(n)} \right)_{t \in \mathbb{R}_+} \overset{\mathcal{D}}{\to} \left( \varphi_t \right)_{t \in \mathbb{R}_+}$$

(7)

in the Skorohod topology$^4$ in probability, where $\left( \varphi_t \right)_{t \in \mathbb{R}_+}$ is an Ornstein–Uhlenbeck process given by

$$d\varphi_t = -\frac{1}{2} B \varphi_t \, dt + \sqrt{A} dW_t,$$

(8)

with $W_t$ a $d$-dimensional standard Brownian motion, drift matrix $B = c_h \Gamma J_*$, positive semi-definite diffusion matrix $A = \mathbb{I}_{[b + h, b]} \frac{c_b}{2 c_\beta} \Gamma^\top \Gamma + \mathbb{I}_{[b + h, \infty]} \frac{c_b}{c_\beta} \Lambda$, and batch constant

$$c_b := \begin{cases} 1 - c_b & b = 1 \text{ and } \text{“no replacement”} \\ 1 & \text{otherwise.} \end{cases}$$

$^2$We take $\beta^{-1}$ to mean 0 when $\beta = +\infty$, in which case we recover SGD from Eq. (2).

$^3$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 Appendix A.3 for further discussion.
The proof of Theorem 1 is given in Appendix B.

Remark 1 (Effects of stochastic gradient noise). As expected, the mini-batch noise contributes in the large-sample regime when \( h + b \leq t \). 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, \( \xi_k \). We can interpret the phase transition as occurring because the variance of the mini-batch gradient scales as \( n^{-2h-b} \) while the variance of update due to the Gaussian innovations scale as \( n^{-h-t} \). The spatial scaling is chosen as \( w = \min\{b + h, t\}/2 \) 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 \( \hat{\theta}^{(n)} \).

Based on Theorem 1, we can establish the following corollaries which 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. A square matrix \( M \) is said to be Hurwitz (or stable) if every eigenvalue of \( M \) has negative real part.

**Corollary 2** (Marginal and stationary covariances). In the setting of Theorem 1, the following hold:

1. For any initial parameter \( \vartheta_0 \), at time \( t \) the marginal covariance of the limiting process is
   \[
   Q_t := \text{Cov}(\vartheta_t|\vartheta_0) = \int_0^t e^{-sB/2} A e^{-sB^\top/2} ds
   \]
   and the marginal distribution is \( \mathcal{L}^{\vartheta_0}(\vartheta_t) = N_d(e^{-sB/2}\vartheta_0, Q_t) \).

2. If \( -\Gamma J_\star \) is Hurwitz, then \( Q_\infty := \lim_{t \to \infty} Q_t \) exists and the stationary distribution of \((\vartheta_t)_{t \in \mathbb{R}}\) is \( \nu := N_d(0, Q_\infty) \). In this case, \( Q_\infty \) solves the equation
   \[
   \frac{1}{2} B Q_\infty + \frac{1}{2} Q_\infty B^\top = A. \tag{9}
   \]

The previous corollary leads to conditions for a Bernstein–von Mises-type result for the stationary distributions of the meta-algorithm.

**Corollary 3** (Bernstein–von Mises-type theorem). In the setting of Theorem 1, if \( -\Gamma J_\star \) is Hurwitz and \((\vartheta_\star^{(n)})_{n \in \mathbb{N}}\) has a sub-sequence with uniformly tight stationary measures, then the sub-sequence of stationary measures converges weakly to \( N_d(0, Q_\infty) \) in probability.

This can be interpreted as saying that if there is a subsequence of the stationary measures where no probability mass “escapes to infinity” along that subsequence, then that subsequence converges weakly to the stationary distribution of the limiting process.

3.2 Discussion of assumptions

Assumptions 1 to 5 are fairly mild. Assumption 1 requires that the likelihood has a minimal number of continuous derivatives, and that the regularizer is smooth.\(^5\)

**Assumption 1.** \( \nabla r \) is \( L_0 \)-Lipschitz, and \( \ell(\cdot; x) \in C^2(\Theta) \) for each \( x \in \mathcal{X} \).

Assumption 2 ensures that the gradient value of the log-likelihood at the limiting parameter is not too volatile via a moment condition.

**Assumption 2.** \( h - w - a/3 > 0 \) and \( \mathbb{E} \|\nabla \ell(\theta_\star; X_1)\|^p < \infty \) for some \( p > \frac{1}{h-w-a/3} \).

Assumption 3 ensures that the random likelihood functions from each data sample are sufficiently smooth via a moment condition on the random smoothness parameter.

\(^5\)“Smoothness” here is being used in the optimization theory sense of the word, referring to Lipschitz gradients.
Assumption 3. There exists $q_3 \in [0, w)$ such that for $p_3 := \frac{1}{b + q_3 - \sqrt{3}}$, 
\[
\|\hat{\theta}^{(n)} - \theta_*\| \in o_p(1/n^{\alpha_3}) \quad \text{and} \quad \mathbb{E} \left[ \left\| \nabla^{\otimes 2} \ell(\cdot;X_1) \right\|_{\infty}^{p_3} \right] < \infty.
\] (10)

Assumptions 4 and 5 require convergence of the first- and second-order empirical Fisher information matrices
\[
\hat{T}^{(n)}(\theta) = \frac{1}{n} \sum_{i \in [n]} \left[ \nabla \ell(\theta; X_i) \right]^{\otimes 2} \quad \text{and} \quad \hat{J}^{(n)}(\theta) = \frac{1}{n} \sum_{i \in [n]} \left[ -\nabla^{\otimes 2} \ell(\theta; X_i) \right].
\]

Assumption 4. There is a nondecreasing sequence $(r_{\mathcal{J}, n})_{n \in \mathbb{N}}$ with $r_{\mathcal{J}, n} \to \infty$, such that
\[
\sup_{\theta \in B(\hat{\theta}^{(n)}; r_{\mathcal{J}, n}/n^w)} \left\| \hat{J}^{(n)}(\theta) - \mathcal{J}(\theta_*) \right\| \overset{p}{\to} 0.
\]

Assumption 5. There is a non-decreasing sequence $(r_{\mathcal{I}, n})_{n \in \mathbb{N}}$ with $r_{\mathcal{I}, n} \to \infty$, such that
\[
\sup_{\theta \in B(\hat{\theta}^{(n)}; r_{\mathcal{I}, n}/n^w)} \left\| \hat{T}^{(n)}(\theta) - \mathcal{I}(\theta_*) \right\| \overset{p}{\to} 0.
\]

The assumptions all hold, for example, for generalized linear models with bounded covariates and either Lipschitz inverse-link functions, or suitably constrained parameter domains. Several sufficient conditions for each of Assumptions 4 and 5 are given in Appendix D.

4 Practical implications of the scaling limit

4.1 Mixing time

Let $\hat{\nu}^{(n)}_k(f) := k^{-1} \sum_{k'=1}^k f(\hat{\theta}^{(n)}_{k'})$ denote the Monte Carlo estimate of $\nu^{(n)}(f)$, where $\nu^{(n)}$ is the 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) $\tau^{(n)} := \sup \{ t : \text{Var}_{\nu^{(n)}}(f)/\text{Var}_{\nu^{(n)}}(f) \leq 1 \}$ to characterize the efficiency of MCMC algorithms. For the limiting process, define the “Monte Carlo average” $\hat{\nu}_t(f) := t^{-1} \int_0^t f(\hat{\theta}_s) \, ds$ and the mixing time $\tau := \sup \{ t : \text{Var}_{\hat{\nu}_t}(f)/\text{Var}_{\hat{\nu}_t}(f) \leq 1 \}$. When the limiting process is reversible, standard results\(^6\) allow us to upper bound $\tau$ 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. In Appendix I, we provide further details and describe a possible approach to making the mixing result rigorous.

4.2 Implications for optimization

The key implication of our results for optimization concern the average $\bar{\hat{\theta}}^{(n)}_k = \frac{1}{k} \sum_{j=1}^k \hat{\theta}^{(n)}_j$ of the first $k$ iterations of the algorithm. The accuracy of the iterate average is characterized by its covariance $\hat{Q}^{(n)}_k := \text{Cov}(\hat{\theta}^{(n)}_k)$. We can approximate $\hat{Q}^{(n)}_k$ in terms of the covariance of the averaged limiting process, which is defined as $\bar{\theta}_t := t^{-1} \int_0^t \hat{\theta}_s \, ds$. The following result is similar in spirit to Theorem 2.1 of Kushner and Yang [1993].

\(^6\)Apply the spectral theorem for self-adjoint operators [Rudin, 1991] to the Poincaré inequality [Bakry et al., 2014]
Proposition 4 (Path averaging). For a stationary initial parameter \( \vartheta_0 \sim N(0, Q_\infty) \), the covariance of the averaged limiting process is
\[
\hat{Q}_t := \text{Cov}(\hat{\vartheta}_t) = \frac{4}{t} \text{Sym} \left( (c_h \Gamma J_*)^{-1} Q_\infty \right) - \frac{8}{t^2} \text{Sym} \left( (c_h \Gamma J_*)^{-2} \left\{ I - e^{-t(c_h \Gamma J_*)/2} \right\} Q_\infty \right). \tag{11}
\]

The proof of this result is in Appendix E. Using Eq. (9), the leading term has the explicit form
\[
\frac{4}{t} \text{Sym} \left( (c_h \Gamma J_*)^{-1} Q_\infty \right) = \frac{1}{t} \left( [\mathbb{I}_{b+k \leq t}] \text{Sym} \left( (\Gamma J_*)^{-1} Q_\infty \right) - [\mathbb{I}_{b+k \leq t}] \text{Sym} \left( (\Gamma J_*)^{-2} \left[ I - e^{-\frac{t}{\beta}} \right] Q_\infty \right) \right).
\]

When either (i) \( b + k < t \) or (ii) \( b + k = t \) and \( c_\beta = +\infty \), this simplifies to
\[
\frac{4}{t} \text{Sym} \left( (c_h \Gamma J_*)^{-1} Q_\infty \right) = \frac{c_b}{t c_h} J_*^{-1} I_* J_*^{-1}.
\]

To describe the large-sample behaviour of the iterate average, 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 5 (Iterate averaging). Fix a number of passes over the dataset, \( m \in \mathbb{R}_+ \). Suppose Assumptions 1 to 5 all hold. If \( b + k \leq t \) and \( (\hat{d}_t^{(n)})_{t \in \mathbb{R}_+}^{n \in \mathbb{N}} \) have initial distributions converging weakly to \( \nu \), then
\[
n \text{Cov} \left( \hat{\theta}^{(n)}_{mn/b(n)} \right) \to 4c_b c_h m \left( [\mathbb{I}_{b+k \leq t}] \text{Sym} \left( (\Gamma J_*)^{-1} Q_\infty \right) - [\mathbb{I}_{b+k \leq t}] \text{Sym} \left( (\Gamma J_*)^{-2} \left[ I - e^{-\frac{km}{\beta}} \right] Q_\infty \right) \right).
\]

If in addition \( b + k < \min(1, t) \) or \( c_\beta = +\infty \), then
\[
n \text{Cov} \left( \hat{\theta}^{(n)}_{mn/b(n)} \right) \to \frac{J_*^{-1} I_* J_*^{-1}}{m}.
\]

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. In particular, we can conclude from Eqs. (3) and (13) that iterate averaging with large step size (that is, slowly decreasing in the mixing speed at stationarity. In particular, we can conclude from Eqs. (3) and (13) that iterate averaging with large step size (that is, slowly decreasing in \( n \), results in optimal numerical estimation of \( \hat{\theta}^{(n)} \), in the sense that after \( m \gg 1 \) passes over the dataset, the numerical error is small compared to the statistical error: \( \text{Cov}(\hat{\theta}^{(n)}_{mn/b(n)}) \approx \frac{1}{m} \text{Cov}(\hat{\theta}^{(n)}) \ll \text{Cov}(\hat{\theta}^{(n)}) \).

More generally, observe that a phase change occurs at \( b + k = 1 \). When \( b + k = 1 \), the rate of concentration for the Bernstein–von Mises-like result is classical, \( \psi = 1/2 \), 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 \( b + k > 1 \), 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 \( b + k < 1 \), the mixing time decreases, but is exactly offset by a slower Bernstein–von Mises-like concentration rate relative to when \( b + k = 1 \), overall yielding the same rate of concentration for the iterate averages as when \( b + k = 1 \).

4.3 Implications for sampling

4.3.1 Sampling from the posterior

The Bernstein-von Mises theorem tells us that the posterior distribution of the parameter \( \theta^{(n)} \sim \pi^{(n)} \) satisfies \( n^{1/2}(\theta^{(n)} - \hat{\theta}^{(n)}) \sim N_\beta(0, J_*^{-1}) \) in probability. In order for the large-sample stationary distribution of Eq. (5) to match the Bernstein–von Mises limit of the posterior, we must first enforce that \( \psi = 1/2 \). Then, there are three ways to ensure that the limiting process has the same distribution as the limiting posterior:

1. Choose the tuning parameter scalings so that \( k + b > t \), which then requires \( t = 1 \) to ensure \( \psi = 1/2 \). 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 \( \Lambda, c_\beta = 1 \), and arbitrary values of \( c_h, c_b \) will suffice.
2. 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.

   a. If $h + b < t$, select $\Gamma$ such that the matrix $Q_\infty$ that solves

   $$
   \frac{1}{2} \Gamma \mathcal{J}_* Q_\infty + \frac{1}{2} Q_\infty \mathcal{J}_*^\top \Gamma^\top = \frac{c_h c_b}{4 c_b} \Gamma \mathcal{I}_* \Gamma^\top
   $$

   is $Q_\infty = \mathcal{J}_*$. As can be verified directly, and is essentially argued in Mandt et al. [2017, Corollary 4], taking $\Gamma = \mathcal{I}_*$ and $c_h = \frac{4 c_b}{c_b}$, the limiting stationary measure will match the limiting posterior.

   b. If $h + b = t$, take $\Lambda = \mathcal{J}_*^{-1} = \Gamma$, and choose $c_h$ and $c_\beta$ jointly so that $\frac{c_h c_b}{4 c_b} + \frac{1}{c_\beta} = 1$.

3. Use the SGLD fixed point algorithm with $h + b = t = 1$. As in the first case, it then suffices to select $\Gamma = \Lambda$, $c_\beta = 1$, and arbitrary values of $c_h, c_b$.

In terms of the number of gradient queries per unit time in the scaling limit scales, options 2b and 3 are the most efficient as the query-count scales linearly with the dataset size (since $h + b = 1$), while for options 1 and 2a it scales super-linearly ($h + b > 1$). In practice, options 2b and 3 both first require an estimate of $\hat{\theta}^{(n)}$ to, respectively, construct the preconditioner $\Gamma = \mathcal{J}^{(n)}(\hat{\theta}^{(n)})^{-1} \approx \mathcal{J}_*^{-1}$ preconditioning or construct the control variates $\nabla t(\hat{\theta}^{(n)}; X_i)$. The latter option is more appealing, particularly if $d$ is large, is no matrix inversion or per-iteration multiplication is required. In either case, however, preconditioning with $\mathcal{J}^{(n)}(\hat{\theta}^{(n)})^{-1}$ will minimize the mixing time.

4.3.2 Alternative uncertainty quantification

When the model is misspecified or when generalized Bayesian inference is used, the (generalized) posterior distribution may thus provide less-than-robust uncertainty quantification [Kleijn and Van der Vaart, 2012, Müller, 2013] because the (local) M-estimator $\hat{\theta}^{(n)}$ is itself asymptotically normal, centered at the true parameter $\theta_*$, with covariance equal to the “sandwich” covariance matrix, $\mathcal{J}_*^{-1} \mathcal{I}_* \mathcal{J}_*^{-1}$. If the model is well-specified (i.e., $P = Q_{\theta}$ for some $\theta \in \Theta$), then $\mathcal{J}_* = \mathcal{I}_*$ and so $\mathcal{J}_*^{-1} \mathcal{I}_* \mathcal{J}_*^{-1} = \mathcal{J}_*^{-1}$. However, if the model is misspecified (i.e., $P \neq Q_{\theta}$ for any $\theta \in \Theta$), then the sandwich may differ from $\mathcal{J}_*^{-1}$ [Huber, 1967, White, 1982]. In this case, posterior credible sets are not asymptotically well-calibrated frequentist confidence sets [Kleijn and Van der Vaart, 2012, Müller, 2013].

The question of how to account for misspecification in the Bayesian setting has been addressed in a number of ways [e.g., Royall and Tsou, 2003, Müller, 2013, Stafford, 1996, Grünwald and Van Ommen, 2017, Huggins and Miller, 2019]. For example, we may want to match the sandwich covariance, as prescribed by [Müller, 2013], which by definition is robust to model misspecification in a frequentist sense. Or we may want to combine the sandwich and Bernstein–von Mises covariances, as in the bagged posterior [Huggins and Miller, 2019]. Either of these desiderata can be obtained by setting $\Gamma = \Lambda = \mathcal{J}_*^{-1}$, and any valid $h + b = 1 = t$. With this tuning, for any $w_1, w_2 > 0$, taking $c_h = 4 w_1 c_b$ and $c_\beta = w_2^{-1}$, gives

$$Q_\infty = w_1 \mathcal{J}_*^{-1} \mathcal{I}_* \mathcal{J}_*^{-1} + w_2 \mathcal{J}_*^{-1}.$$

This matches the asymptotic distribution of the bagged posterior with re-sampling rate $w_1$ when $w_1 = w_2$ [Huggins and Miller, 2019]. Moreover, we can obtain any convex combinations of the uncertainty quantification from the posterior and the asymptotics of the M-estimator by taking $w_1 + w_2 = 1$. This enables interpolation between frequentist-like and Bayesian-like forms of inference and results in a mixing time of $n/w_1$ passes over the dataset. Hence in principle if we use SGD (i.e., set $w_1 = 1, w_2 = 0$, and $c_\beta = +\infty$), can also obtain the sandwich covariance and minimize the mixing time to be a single pass over the dataset.

A final alternative it to leverage Eq. (13) by choosing $b + h < t$ and then using the iterate average of $n/b^{(n)}$ iterations as a single sample, which will have asymptotic variance equal to the sandwich covariance. In this case the mixing time is $2 \alpha^{(n)}/\lambda_{\min}(B) \ll n/b^{(n)}$ iterations for $n$ sufficiently large, so consecutive iterate averages provide nearly independent samples. This approach is particularly attractive when $d$ is large since it does not require any preconditioning and can leverage constant mini-batch sizes and very large step sizes by choosing $b = 0$ and $h$ close to zero.
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 [e.g., Pollock et al., 2020, Nemeth and Fearnhead, 2021, Vollmer et al., 2016]. 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 [Gupal and Bazhenov, 1972] and underdamped (stochastic gradient) Langevin dynamics [An et al., 2018, Lessard et al., 2016, Cyrus et al., 2018, Van Scoy et al., 2017, Kidambi et al., 2018, Ma and Yarats, 2018] 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 [Baker et al., 2019, Nagapetyan et al., 2017], 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.

5.1 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 [2018], which includes many momentum-based algorithms as special cases, such as momentum algorithm, Nesterov’s accelerated gradient, PID control algorithms [An et al., 2018], synthesized Nesterov variants [Lessard et al., 2016], noise-robust momentum [Cyrus et al., 2018], triple momentum [Van Scoy et al., 2017], least-squares acceleration of SGD [Kidambi et al., 2018]. See [Ma and Yarats, 2018, 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 \( \tilde{\ell}(\theta, \tilde{\theta}; x) = \ell(\theta; x) - \tilde{\theta}^T \Gamma^{-1} \tilde{\theta}/2 \), and lift the prior to phase space using the (improper) prior \( \tilde{\pi}(0) \). For (stochastic) heavy ball and underdamped (stochastic gradient) Langevin dynamics (cf., e.g., Duncan et al. [2017, 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

\[
\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h^{(n)}}{2} \Gamma^{-1} \tilde{\theta}_k^{(n)}
\]

\[
\tilde{\theta}_{k+1}^{(n)} = \left( I - \frac{h^{(n)}}{2} \Gamma^{-1} \right) \tilde{\theta}_k^{(n)} + \frac{h^{(n)}}{2} \tilde{G}_k + \sqrt{\frac{h^{(n)}}{\beta^{(n)}}} \Gamma \xi_k.
\]

The corresponding limiting process is

\[
d \begin{bmatrix} \tilde{\theta} \\ \tilde{\theta} \end{bmatrix} = -\frac{1}{2} \tilde{B} \begin{bmatrix} \tilde{\theta} \\ \tilde{\theta} \end{bmatrix} dt + \sqrt{A} d\tilde{W}_t
\]

with \( \tilde{W}_t \) a 2d-dimensional standard Brownian motion, drift and diffusion matrices, respectively,

\[
\tilde{B} = c_h \begin{bmatrix} 0 & -M^{-1} \\ \mathcal{J} \Gamma & \Gamma M^{-1} \end{bmatrix}, \quad \text{and}
\]

\[
A = \mathbb{I}_{[\theta + b \leq 0]} \frac{c_h^2}{4c_b} \begin{bmatrix} 2 & 0 \\ 0 & \mathcal{I} \end{bmatrix} + \mathbb{I}_{[\theta \leq b + \hat{b}]} \frac{c_h}{c_b} \begin{bmatrix} 0 & \mathcal{I} \\ 0 & \Gamma \end{bmatrix}.
\]

(16)
5.2 Extension to control variates

SGLD Methods with control variates [Baker et al., 2019, Nagapetyan et al., 2017] aim to 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 of Section 3. However, our analysis can be easily modified to apply to these control variate methods, as we show in Appendix H.

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 by taking \( \text{w} = t/2 \). Under this scaling, the corresponding limiting process the Ornstein–Uhlenbeck process:

\[
d\vartheta_t = -\frac{1}{2}B\vartheta_t \, dt + \sqrt{A} \, dW_t,
\]

with \( B = c_\beta \Gamma J_* \) the drift matrix, \( A = \frac{\Delta}{c_\beta} \Lambda \) the positive semi-definite diffusion matrix, and \( W_t \) a d-dimensional standard Brownian motion.

5.3 Extension to constrained parameter spaces

If \( \Theta \subseteq \mathbb{R}^d \), then the iterations given by Eq. (5) may exit \( \Theta \), resulting in undefined behaviour. The typical way to handle this case is to impose boundary dynamics. The two most common examples of such boundary dynamics are reflecting and projecting. Projecting maps iterates that would exit \( \Theta \) to the nearest point within \( \Theta \). 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 reached the boundary if it had not collided. In this section we consider boundary dynamics satisfying a generalized version of this property.

Let \( \mathcal{P} : \Theta \times (\mathbb{R}^d)^3 \rightarrow \Theta \) be a measurable function such that:

(i) \( \mathcal{P} \) is faithful to \( \Theta \), meaning that if \( \text{Conv}(\theta, \theta + \Delta x(0) + \Delta \ell + \Delta z) \subset \Theta \) then

\[
\mathcal{P}(\theta, \Delta x(0), \Delta \ell, \Delta z) = \theta + \Delta x(0) + \Delta \ell + \Delta z,
\]

where \( \text{Conv}(\theta_1, \theta_2) \) is the line segment from \( \theta_1 \) to \( \theta_2 \).

(ii) \( \mathcal{P} \) is local, meaning that there exists \( c_\mathcal{P} > 0 \) such that for all \( (\theta, \Delta x(0), \Delta \ell, \Delta z) \in \Theta \times (\mathbb{R}^d)^3 \)

\[
\| \mathcal{P}(\theta, \Delta x(0), \Delta \ell, \Delta z) - \theta \| \leq c_\mathcal{P} (\| \Delta x(0) \| + \| \Delta \ell \| + \| \Delta z \|).
\]

We will consider the iterative algorithm on \( \Theta \) given by

\[
\theta_{k+1}^{(n)} = \mathcal{P} \left( \theta_k^{(n)}, \frac{h\Gamma}{2n} \nabla \log \pi(0) \left( \theta_k^{(n)} \right), \frac{h\Gamma}{2} \sum_{j \in [b]} \nabla \ell \left( \theta_k^{(n)} ; \ X_{k+1}^{(n)}(j) \right), \sqrt{h^2 \beta^{-1} \Lambda} \right).
\]

When \( \Theta \subseteq \mathbb{R}^d \) and \( \theta_* \in \text{interior}(\Theta) \) the proof is essentially the same because the boundary dynamics are faithful and local. Intuitively, the assumption that \( \vartheta^{(n)}(0) \sim \vartheta(0) \) ensures that the processes we consider all start near \( \theta_* \) and away from the boundary of \( \Theta \), and thus the spatial scaling drives the boundary of \( \Theta \) 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 Appendix H.
6 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.

6.1 Experiment 1: Gaussian simulation study

In this experiment, we demonstrate the effect of model misspecification. Exact specifications for the experiment are given in Table 1. The combination of true distribution and likelihood function was chosen specifically to ensure that $J_\star \neq I_\star$, so that the effect of misspecification would be apparent. We run SGD with no preconditioning, with preconditioning by $J_\star$, and with preconditioning by $I_\star$, and SGLD with preconditioning by $J_\star$. We interpret this using our scaling limit with parameters $w = 1/2, b = 1, \beta = 0$. This combination of scaling parameters corresponds to the standard statistical local scaling, and a fixed batch size. For SGLD we also use $t = 0$ 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. The results are plotted and described in Fig. 1 and Table 2. As predicted by our results, preconditioning by $J_\star$ leads to an empirical distribution for the iterates of the algorithm matching the covariance of the MLE, preconditioning by $I_\star$ 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 $J_\star$ 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.

| Method | Empirical IACT | Predicted IACT |
|--------|----------------|----------------|
| SGD, no preconditioning | 3.2 epochs | 3.2 epochs |
| $J^{-1}_\star$-preconditioned SGD | 1.1 epochs | 1.0 epochs |
| $I^{-1}_\star$-preconditioned SGD | 2.3 epochs | 2.8 epochs |
| $J^{-1}_\star$-preconditioned SGLD | 2.2 epochs | 2.0 epochs |

Table 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.

Table 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.
6.2 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. [2020], using the pre-processed data they provided. The responses are binary and there are 3 covariates. We use the full dataset (≈ 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 \(w = 1/2, \beta = 1, a = 0\). The results are plotted and described in Figs. 2 and 3 and Table 3. An experimental finding of Pollock et al. [2020] 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 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.
6.3 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. [2020] were based upon, in order to examine a more complex, more misspecified model on thematically similar data. In this case the responses 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. The model has 25 parameters. We use the full 2008 data (≈ 1.5 million observations) to estimate the “ground truth” quantities \((\theta^*, J^*, I^*)\), 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 \(J^*\) and \(I^*\) differ significantly in scale, and hence both 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. The results are plotted and described in Fig. 4 and Table 4. Both preconditioned methods behave exactly as predicted by the asymptotic theory.

7 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. [2020] 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 [2017] 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. [2018] 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. [2016] study the consistency, CLT, and asymptotic bias–variance decomposition of SGD for a sequence of decreasing step sizes that converge zero. Vollmer et al. [2016] 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. [2017] 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. [2018] 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. [2020] 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. [2021] 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. [2019] 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. [2020] consider recursive stochastic algorithms as approximations of certain contraction operators and view them within the framework of iterated random operators. Sirignano and Spiliopoulos [2017, 2020] 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. [1997]. 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 Schmon et al. [2021], Schmon and Gagnon [2021] 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.

The present work 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. Li et al. [2020], Zhou et al. [2020], Wu et al. [2020]) and for Bayesian deep learning (see e.g. Wenzel &

Table 4: Mixing times for experiment 4 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.
References

S. Ahn, A. Korattikara, and M. Welling. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In International Conference on Machine Learning, 2012.

L. Aitchison. A statistical theory of cold posteriors in deep neural networks. arXiv preprint arXiv:2008.05912, 2020.

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, pages 8522–8531, 2018.

Y. F. Atchadé. Approximate spectral gaps for markov chain mixing times in high dimensions. SIAM Journal on Mathematics of Data Science, 3(3):854–872, 2021.

J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient MCMC. Statistics and Computing, 29(3):599–615, 2019. doi: 10.1007/s11222-018-9826-2.

D. Bakry, I. Gentil, M. Ledoux, et al. Analysis and geometry of Markov diffusion operators, volume 103. Springer, 2014.

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):1103, 2016.

N. Brosse, A. Durmus, and E. Moulines. The promises and pitfalls of Stochastic Gradient Langevin Dynamics. In Advances in Neural Information Processing Systems, 2018.

T. Chen, E. B. Fox, and C. Guestrin. Stochastic Gradient Hamiltonian Monte Carlo. In International Conference on Machine Learning, 2014.

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), pages 1376–1381. IEEE, 2018.

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):1348–1382, 2020.

A. Duncan, N. Nüsken, and G. Pavliotis. Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. Journal of Statistical Physics, 169(6):1098–1131, 2017.

R. Durrett. Probability: theory and examples, volume 49. Cambridge university press, 4th edition, 2019.

S. N. Ethier and T. G. Kurtz. Markov processes: characterization and convergence, volume 282. John Wiley & Sons, 2009.

I. Goodfellow, Y. Bengio, and A. Courville. Deep Learning. MIT Press, 2016.

P. Grünwald and T. Van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. Bayesian Analysis, 12(4):1069–1103, 2017.

A. Gupal and L. Bazhenov. Stochastic analog of the conjugant-gradient method. Cybernetics, 8(1):138–140, 1972.

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):967–1003, 2020.

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, volume 5, page 221. Univ of California Press, 1967.
J. H. Huggins and J. W. Miller. Using bagged posteriors for robust inference and model criticism. arXiv preprint arXiv:1912.07104, 2019.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. What are bayesian neural network posteriors really like?, 2021. arXiv: 2104.14421.

O. Kallenberg. Foundations of modern probability. Springer Science & Business Media, 2006.

I. Karatzas and S. Shreve. Brownian motion and stochastic calculus, volume 113. springer, 2014.

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), pages 1–9. IEEE, 2018.

B. J. K. Kleijn and A. W. Van der Vaart. The Bernstein-von-Mises theorem under misspecification. Electronic Journal of Statistics, 6:354–381, 2012.

H. Kushner and G. G. Yin. Stochastic approximation and recursive algorithms and applications, volume 35. Springer Science & Business Media, 2003.

H. J. Kushner and J. Yang. Stochastic Approximation with Averaging of the Iterates: Optimal Asymptotic Rate of Convergence for General Processes. SIAM Journal on Control and Optimization, 31(4):1045–1062, 1993. ISSN 0363-0129. doi: 10.1137/0331047.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. SIAM Journal on Optimization, 26(1):57–95, 2016.

Q. Li, C. Tai, and E. Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations. J. Mach. Learn. Res., 20:40–1, 2019.

Z. Li, K. Lyu, and S. Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. arXiv preprint arXiv:2010.02916, 2020.

K. Liu, L. Ziyin, and M. Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In International Conference on Machine Learning, pages 7045–7056. PMLR, 2021.

J. Ma and D. Yarats. Quasi-hyperbolic momentum and adam for deep learning. In International Conference on Learning Representations, 2018.

Y.-A. Ma, T. Chen, and E. B. Fox. A Complete Recipe for Stochastic Gradient MCMC. In Advances in Neural Information Processing Systems, pages 2917–2925, 2015.

S. Mandt, M. D. Hoffman, and D. M. Blei. Stochastic gradient descent as approximate Bayesian inference. The Journal of Machine Learning Research, 18(1):4873–4907, 2017.

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):1–64, 2021.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/40008b9a5380fcaacce397c08af5b-Paper.pdf.

U. K. Müller. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. Econometrica, 81(5):1805–1849, 2013.

T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. The true cost of stochastic gradient langevin dynamics. arXiv preprint arXiv:1706.02692, 2017.

C. Nemeth and P. Fearnhead. Stochastic gradient Markov chain Monte Carlo. Journal of the American Statistical Association, 533(116):433–450, 2021.
A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. \textit{SIAM Journal on Optimization}, 19(4):1574–1609, 2009.

M. Pollock, P. Fearnhead, A. M. Johansen, and G. O. Roberts. Quasi-stationary monte carlo and the scale algorithm. \textit{Journal of the Royal Statistical Society: Series B (Statistical Methodology)}, 82(5):1167–1221, 2020.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. \textit{SIAM journal on control and optimization}, 30(4):838–855, 1992.

S. J. Reddi, S. Kale, and S. Kumar. On the Convergence of Adam and Beyond. International Conference on Learning Representations, 2018.

H. Robbins and S. Monro. A Stochastic Approximation Method. \textit{The Annals of Mathematical Statistics}, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL \url{https://doi.org/10.1214/aoms/1177729586}.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. \textit{The annals of applied probability}, 7(1):110–120, 1997.

R. Royall and T.-S. Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. \textit{Journal of the Royal Statistical Society: Series B (Statistical Methodology)}, 65(2):391–404, 2003.

W. Rudin. Functional analysis, mcgrawhill. Inc, New York, 45:46, 1991.

S. M. Schmon and P. Gagnon. Optimal scaling of random walk metropolis algorithms using bayesian large-sample asymptotics. arXiv preprint arXiv:2104.06384, 2021.

S. M. Schmon, G. Deligiannidis, A. Doucet, and M. K. Pitt. Large-sample asymptotics of the pseudo-marginal method. \textit{Biometrika}, 108(1):37–51, 2021.

J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time. \textit{SIAM Journal on Financial Mathematics}, 8(1):933–961, 2017.

J. Sirignano and K. Spiliopoulos. Stochastic gradient descent in continuous time: A central limit theorem. \textit{Stochastic Systems}, 10(2):124–151, 2020.

J. E. Stafford. A robust adjustment of the profile likelihood. \textit{The Annals of Statistics}, 24(1):336–352, 1996.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. \textit{The Journal of Machine Learning Research}, 17(1):193–225, 2016.

P. Toulis and E. M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. \textit{The Annals of Statistics}, 45(4):1694–1727, 2017.

P. Toulis, T. Horel, and E. M. Airoldi. The proximal robbins–monro method. \textit{Journal of the Royal Statistical Society: Series B (Statistical Methodology)}, 83(1):188–212, 2021. doi: \url{https://doi.org/10.1111/rssb.12405}. URL \url{https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12405}.

B. Tzen, T. Liang, and M. Raginsky. Local optimality and generalization guarantees for the Langevin algorithm via empirical metastability. In Conference On Learning Theory, pages 857–875, 2018.

B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. \textit{IEEE Control Systems Letters}, 2(1):49–54, 2017.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-) asymptotic bias and variance of stochastic gradient langevin dynamics. \textit{The Journal of Machine Learning Research}, 17(1):5504–5548, 2016.

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), pages 681–688. Citeseer, 2011.
A 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.

A.1 Bernstein-von Mises under misspecification

**Definition 6.** The first and second order Fisher information matrices, \( I \) and \( J \) respectively, are defined for a log-likelihood function \( \ell \) and probability distribution \( P \) by

\[
I(\theta) = \mathbb{E}_{X \sim P} [\nabla_\theta \ell(\theta; X) \otimes \nabla_\theta \ell(\theta; X)], \quad \text{and} \quad J(\theta) = -\mathbb{E}_{X \sim P} \nabla_\theta^2 \ell(\theta; X).
\]

Let \( \mathcal{X} \) be a Polish space with \( \sigma \)-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 \( X^{(n)} := (X_i)_{i \in \mathbb{N}} \sim P^\otimes \mathbb{N} \). Let \( n \in \mathbb{N} \) denote a sample size, let \([n] := \{1, \ldots, n\}\), and let \( X^{(n)} := (X_i)_{i \in \{1, \ldots, n\}} \sim P^\otimes n \) 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 \( \Theta \) 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.\(^7\)

Suppose there exists a \( \sigma \)-finite measure, \( \mu \), on \( \mathcal{X} \), such that for all \( \theta \in \Theta \), \( Q_\theta \ll \mu \). Let \( q_\theta \) denote a version of \( \frac{dQ_\theta}{d\mu} \) 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 \( \mathcal{M} := \{Q_\theta \mid \theta \in \Theta\} \) to be a model for \( P \). The model is well-specified when \( P \in \mathcal{M} \), and is misspecified otherwise. The pseudo-true parameter of the model is defined as \( \theta_* := \arg \max_{\theta \in \Theta} \mathbb{E}_{X \sim P} \ell(\theta; X) \). If \( \mu \ll P \) then

\[
\theta_* = \arg \max_{\theta \in \Theta} \mathbb{E}_{X \sim P} \ell(\theta; X) = \arg \min_{\theta \in \Theta} \text{KL}(P||Q_\theta).
\]

Let \( \Pi^{(0)} \in \mathcal{M}_{1,+}(\Theta) \) be any distribution on \( \Theta \). Let \( P_{\Pi^{(0)}}, M \in \mathcal{M}_{1,+}(\Theta \otimes \mathcal{X}^\mathbb{N}) \), given by

\[
P_{\Pi^{(0)}, M}(A \times B) := \int \mathbb{I}_{[\theta \in A]} \left[ \int \mathbb{I}_{[x^{(n)} \in B]} Q_\theta^N(dx^{(n)}) \right] \Pi^{(0)}(d\theta)
\]

denote the joint distribution of the data and the parameter according to the model and the prior, where \( Q_\theta^N(dx^{(n)}) \) denotes the law of an I.I.D. sequence from \( Q_\theta \) (an infinite product measure on the cylinder \( \sigma \)-field). Let \( E_{\Pi^{(0)}, M} \) denote the expectation under \( P_{\Pi^{(0)}, M} \). The posterior for \( \theta \) under the model \( M \) given data \( X^{(n)} \) is the random probability measure on \( \Theta \) given by

\[
\Pi^{(X^{(n)})}(A) := E_{\Pi^{(0)}, M} \left[ \mathbb{I}_{[\theta \in A]} \right],
\]

\(^7\)\( \Theta \) is equipped with the Borel \( \sigma \)-field inherited from \( \mathbb{R}^d \)
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 \( \Pi(X^{(n)}) \) can be viewed as a probability kernel from \( X^n \) to \( \Theta \).

Let \( \lambda \) denote the Lebesgue measure. If \( \Pi^{(0)} \ll \lambda \) with \( d\Pi^{(0)}/d\lambda =: \pi^{(0)} \), then \( \Pi(X^{(n)}) \ll \lambda \) with \( d\Pi(X^{(n)})/d\lambda = \pi(X^{(n)}) \) given by

\[
\pi(X^{(n)}) = \pi^{(0)}(\theta) \prod_{i \in [n]} q_\theta(X_i) = \pi^{(0)}(\theta) \exp \left( \sum_{i \in [n]} \ell(\theta; X_i) \right).
\] (21)

Let \( \hat{\theta}^{(n)} := \arg\max_{\theta \in \Theta} \sum_{i \in [n]} \ell(\theta; X_i) \) denote the maximum likelihood estimator (MLE) of \( \theta_\star \) given the data \( X^{(n)} \). 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 \( \vartheta \in \sqrt{n} \left( \Theta - \hat{\theta}^{(n)} \right) \) based on the identification

\[
\vartheta = \sqrt{n} \left( \theta - \hat{\theta}^{(n)} \right)
\]

and the local model is given by

\[
\mathcal{M}(X^{(n)}) := \left\{ Q_{\hat{\theta}^{(n)} + \sqrt{n} \vartheta} \mid \vartheta \in \sqrt{n} \left( \Theta - \hat{\theta}^{(n)} \right) \right\}.
\]

The random localization map is given by

\[
\text{loc}_{X^{(n)}} : \theta \mapsto \sqrt{n} \left( \theta - \hat{\theta}^{(n)} \right)
\]

For a measurable function \( f : A \to B \) and a measure \( \mu \) on \( A \), the pushforward of \( \mu \) through \( f \) is the measure \( f_\#\mu \) on \( B \) defined by \( [f_\#\mu](B) = \mu(f^{-1}(B)) \) for all measurable \( B \subset B \).

**Proposition 7** (BvM under model misspecification, Kleijn and Van der Vaart [2012]). Under regularity conditions,

\[
\left\| \text{loc}_{X^{(n)}} \right\|_{TV} \rightarrow 0,
\]

with \( \theta_\star = \arg\max_{\theta \in \Theta} \mathbb{E}_{X \sim P} \ell(\theta; X) \), \( J_\star = - \mathbb{E}_{X \sim P} [\nabla^2 \ell(\theta_\star; X)] \), and \( \Phi = N(0, J_\star^{-1}) \).

**A.2 Convergence modes of measures and operators**

Let \( A \) be a measurable space, and let \( B(A) \) denote the collection of bounded measurable functions on \( A \). For a function \( f : A \to L \) with \( (L, \| \cdot \|) \) a normed linear space, define

\[
\| f \|_\infty := \sup_{a \in A} \| f(a) \|.
\]

For a sequence of probability measures, \( \{\mu_n\}_{n \in \mathbb{N}} \) and a probability measure \( \mu \) on a measurable space \( A \), we have the following modes of convergence:

- \( \mu_n \) converges under total variation to \( \mu \), denoted by \( \mu_n \xrightarrow{TV} \mu \), if and only if

  \[
  \sup_{f \in B(A)} \left| \mu_n f - \mu f \right| \rightarrow 0.
  \]

- if \( A \) is also a topological space and the σ-field on \( A \) is the Borel σ-field, then \( \mu_n \) converges in distribution (also called weakly) to \( \mu \), denoted by \( \mu_n \xrightarrow{\text{d}} \mu \), if and only if for all \( f \in \mathcal{C}(A) \), \( \| \mu_n f - \mu f \| \rightarrow 0 \).
Clearly
\[
\left( \mu_n \xrightarrow{T} \mu \right) \implies \left( \mu_n \xrightarrow{s} \mu \right) \implies \left( \mu_n \xrightarrow{w} \mu \right)
\]
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 \( \| y \| := \sup_{x \in L \setminus \{0\}} |f(x)|/\|x\| \) 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
  \[
  \|T_n - T\| = \sup_{(x,y) \in L \times L'} \frac{|\langle y, (T_n - T)x \rangle|}{\|x\|\|y\|} \to 0
  \]  
  \[
  \tag{22}
  \]

- \( T_n \) converges strongly to \( T \), denoted \( T_n \xrightarrow{s} T \) if and only if for all \( x \in L \)
  \[
  \sup_{y \in L'} \frac{|\langle y, (T_n - T)x \rangle|}{\|y\|} \to 0
  \]  
  \[
  \tag{23}
  \]  
  
Clearly
\[
\left( \|T_n - T\| \to 0 \right) \implies \left( T_n \xrightarrow{s} T \right)
\]
while the converse does not hold in general.

### A.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
\[
A f = \lim_{t \to 0^+} T(t) f - f
\]
for \( f \in \text{dom}(A) = \{ f \in L \mid \lim_{t \to 0^+} (T(t) f - f) / t \text{ exists} \} \). Let
\[
\hat{C}(\mathbb{R}^d) = \left\{ f \in C(\mathbb{R}^d) \mid \forall \epsilon > 0 \exists K_{f,\epsilon} \subset \mathbb{R}^d \text{ compact with } \sup_{\theta \not\in K_{f,\epsilon}} |f(\theta)| \leq \epsilon \right\}
\]
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
\[
\| \mu \|_{TV} = \sup_{\substack{f \in \hat{C}(\mathbb{R}^d) \\|f\|_\infty \leq 1}} \left\| \int f(\theta) \mu(d\theta) \right\|.
\]
We will work with \( (L, \| \cdot \|) = (\hat{C}(\mathbb{R}^d), \| \cdot \|_\infty) \). A semigroup on \( (\hat{C}(\mathbb{R}^d), \| \cdot \|_\infty) \) is positive if
v) \( f \geq 0 \implies Tf \geq 0 \).

A semigroup on \( (\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty) \) 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 transition kernels) is defined by

\[
T_t f(x) = \mathbb{E}_x f(X_t) = \int f(y) k_t(x, dy)
\]

(24)

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; \( \text{dom}(A) \) is dense in \( \hat{C}(\mathbb{R}^d) \). A Markov process for which the corresponding forward operators form a Feller semigroup is called a Feller process. Feller processes have a richly developed theory; see, for example, Ethier and Kurtz [2009] or Kallenberg [2006]. 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 \( t > 0 \), \( \lim_{s \downarrow t^-} X(s) \) exists and \( \lim_{t \downarrow t^+} X_t \).

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 [Kallenberg, 2006, Theorem A2.2]. The formula for the Skorohod metric is not particularly illuminating, so is omitted here and may be found in the references. This space is denoted by \( D(I, \mathbb{R}^d) \). The Borel \( \sigma \)-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 \( \mathcal{C}(\mathbb{R}^d) \).

**Proposition 8** (Approximation of Markov Chains (compiled from Ethier and Kurtz [2009])). Let \( A : C_c^\infty(\mathbb{R}^d) \to \mathcal{C}(\mathbb{R}^d) \) 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 \( (\vartheta_k)_{k \in \mathbb{R}} \) be a Markov process with forward operator semigroup \( T \). Let \( \left( \theta_k^{(n)} \right)_{k \in \mathbb{N} \cup \{0\}} \) 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

\[
A^{(n)} = \alpha^{(n)} \left( U^{(n)} - I \right) \quad T^{(n)}_t = \left( U^{(n)} \right)^{[\alpha^{(n)} t]} \quad \vartheta^{(n)}_t = \theta^{(n)}_t^{[\alpha^{(n)} t]}.
\]

If \( \|A^{(n)} f - Af\|_\infty \to 0 \) for all \( f \in C_c^\infty(\mathbb{R}^d) \), then

(a) \( T^{(n)}_t \to T_t \) for each \( t > 0 \), and

(b) If \( \vartheta^{(n)}(0) \to \vartheta(0) \) then \( \vartheta^{(n)}(\cdot) \to \vartheta(\cdot) \) in the Skorohod metric.

**Proof of Proposition 8.** (a) Follows from Chapter 1, Theorem 6.5 of Ethier and Kurtz [2009]. (b) Follows by combining Chapter 4, Theorem 8.2, Corollary 8.5, and Corollary 8.9 of Ethier and Kurtz [2009]. □

**A.4 Miscellaneous notation and definitions**

**Definition 9** (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 \( \sigma \)-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 \( X_n \overset{p}{\to} x \), when for every neighbourhood \( U \in \tau \) we have

\[
\lim_{n \to \infty} \mathbb{P}(X_n \in U^c) = 0.
\]

**Lemma 10.** Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space, let \( (\mathcal{X}, \tau) \) be a topological space endowed with the \( \sigma \)-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 \( n_{mk} \) such that \( X_{n_{mk}} \to x \) almost surely as \( k \to \infty \) then \( X_n \overset{p}{\to} x \).

If \( (\mathcal{X}, \tau) \) is first-countable then the converse also holds; if \( X_n \overset{p}{\to} x \) then for every sub-sequence \( n_m \) there is a sub-sub-sequence \( n_{mk} \) such that \( X_{n_{mk}} \to x \) almost surely as \( k \to \infty \).
The proof of this result is the same as in Durrett [2019, 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.

**B Proof of Theorem 1**

In this section we prove Theorem 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 3.

**Theorem 11 (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. (20) with step-size \(h(n) = c_h n^{-h}\), \(\beta(n) = c_\beta n^\beta\), all other tuning parameters constant in \(n\). Let \(\alpha \in \mathbb{R}^d\). Let \(X^{(n)} = (X_t)_{t \in \mathbb{N}} \sim P_{\alpha}^{(n)}\), and \(\hat{\theta}^{(n)}\) be a critical point of the log-likelihood function \(\sum_{t=1}^{n} \ell(t, X_t)\) for each \(n \in \mathbb{N}\); that is \(\sum_{t=1}^{n} \nabla \ell(\hat{\theta}^{(n)}, X_t) = 0\) for all \(n \in \mathbb{N}\).

Let \(\varphi_t = w^{(n)}(\hat{\theta}_{[\alpha(n)t]}^{(n)} - \hat{\theta}^{(n)})\), where \(w^{(n)} = n^w\), \(\alpha(n) = n^a\), \(w \in (0,1)\),

\[
a = \min \{h, (t + h - 2w), (b + 2h - 2w)\}.
\]

If Assumptions 1 to 5 all hold, \(a > 0\), and \(\varphi_t(0) \rightarrow \varphi_t(0)\) then

1. \((\varphi_t^{(n)})_{t \in \mathbb{R}_+} \rightsquigarrow (\varphi_t)_{t \in \mathbb{R}_+}\) in the Skorohod topology in probability, where \((\varphi_t)_{t \in \mathbb{R}}\) follows the Ornstein–Uhlenbeck process:

\[
d\varphi_t = -\frac{c_d}{2} J(\theta_*) \varphi_t dt + \sqrt{c_g A + c_{eb} \Gamma J(\theta_*) \Gamma'} dW_t,
\]

with

\[
c_d = \begin{cases} c_h & a = h \\ 0 & a < h \end{cases}, \quad c_g = \begin{cases} \frac{2c_d}{c_g} & a = h + t - 2w \\ 0 & a < h + t - 2w \end{cases}
\]

and

\[
c_{eb} = \begin{cases} \frac{c_d^2 (1-c_h)}{4c_h} & a = 1 + 2h - 2w \text{ and } b = 1 \text{ and no replacement} \\ \frac{c_d^2}{4c_h} & a = b + 2h - 2w \text{ and } (b \neq 1 \text{ or replacement}) \\ 0 & a < b + 2h - 2w. \end{cases}
\]

2. If \(T^{(n)}\) and \(T\) are defined as in Proposition 8, then under the conditions above, every subsequence of \((T^{(n)})_{n \in \mathbb{N}}\), then \((T^{(n)}_{m_{n}})_{m \in \mathbb{N}}, T^{(m_{n}+)}_{k \in \mathbb{N}}, \text{ such that with probability} 1, T^{(m_{n}+)}_{k \in \mathbb{N}}, T_{k \in \mathbb{N}} \rightarrow T_t\) for all \(t > 0\).

Before beginning the proof of this result, Theorem 11, 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 12.** Let \(\alpha : \mathbb{R}_+ \rightarrow \mathbb{R}_+\) be non-decreasing, right continuous with left limits, with \(\alpha(0) = 0\), and \(\lim_{t \rightarrow \infty} \alpha_t = \infty\). Let \(Z_i \sim \mu\) for all \(i \in \mathbb{N}\) (possibly not independent) with \(Z_i \geq 0\) 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 \(\alpha\). Then

\[
\mathbb{P}\left(\max_{t \in [n]} Z_t \geq \alpha^+(n) \text{ i.o.}\right) = 0.
\]

**Proof of Lemma 12.** Let \(S_t = \mathbb{P}(Z_1 > t)\) be the survival function of \(\mu\), and let \(W_n = \alpha(Z_n)\) for each \(n \in \mathbb{N}\). Note that \(\mathbb{P}(W_1 > t) = S(\alpha^+ t)\). Then

\[
\infty > \mathbb{E}[(\alpha(Z_1))] = \int_0^\infty \mathbb{P}(W_1 > t) dt \geq \sum_{n=1}^\infty \mathbb{P}(W_1 > n) = \sum_{n=1}^\infty \mathbb{P}(W_n > n).
\]
Therefore, from the Borel–Cantelli lemma $\mathbb{P}(W_n > n \ i.o.) = 0$, and equivalently $\mathbb{P}(W_n \leq n \ a.b.f.o.) = 1$. Now, whenever $W_n \leq n$ for all but finitely many $n$, then there exists $K \in \mathbb{N}$ and $I_1, \ldots, I_K \in \mathbb{N}$ 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 $\mathbb{P}(\max_{i \leq n} W_i \leq n \ a.b.f.o.) = 1$, and equivalently $\mathbb{P}(\max_{i \leq n} W_i > n \ i.o.) = 0$. Finally, $W_i > n$ if and only if $Z_t > \alpha^+(n)$, hence

$$\mathbb{P}(\max_{t \leq n} Z_t > \alpha^+(n) \ i.o.) = 0.$$ 

\[\square\]

## B.1 Proof of Theorem 11

Let $J_\star = J(\theta_\star)$ and $I_\star = I(\theta_\star)$.

The proof proceeds in the following stages. In Appendix B.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-surely along subsequences and construct appropriate such subsequences. In Appendix B.1.2 we introduce notation that will be useful in the remainder of the proof. In Appendix B.1.3 we discuss what is needed to apply Proposition 8 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 Appendix B.1.4, we will consider arguments sufficiently far from the support of the test function. Then, in Appendix B.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 Appendix B.1.6. The second type is terms that do not vanish and do have corresponding terms in the limiting generator; these are handled in Appendices B.1.7 to B.1.9. The third type of term is the remainder term, which is handled in Appendix B.1.10. Putting all of this together allows us to apply Proposition 8 along our subsequences, establishing the main result.

### B.1.1 Reduction to almost-sure convergence on subsequences

Let

$$\Upsilon^{(n)} = \max \left( \Upsilon_1^{(n)} ; \Upsilon_2^{(n)} ; \Upsilon_3^{(n)} \right),$$

$$\Upsilon_1^{(n)} = n^{\gamma_3} \left\| \hat{\theta}^{(n)} - \theta_\star \right\|,$$

$$\Upsilon_2^{(n)} = \sup_{\theta \in B(\hat{\theta}^{(n)}), r_{J,n}/n^m} \left\| \hat{J}^{(n)}(\theta) - J(\theta_\star) \right\|,$$

$$\Upsilon_3^{(n)} = \sup_{\theta \in B(\hat{\theta}^{(n)}), r_{T,n}/n^w} \left\| \hat{I}^{(n)}(\theta) - I(\theta_\star) \right\|.$$

Each of the $\Upsilon$ 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 2 which lets us use a weaker moment assumption for the sup-norm of the Hessian of the log-likelihood.

By assumption, $\Upsilon^{(n)} \overset{P}{\rightarrow} 0$. Then, by Lemma 10, for every subsequence $(n_m)_{m \in \mathbb{N}}$ there is a further sub-subsequence $(n_{mk})_{k \in \mathbb{N}}$ so that this convergence is almost sure. Along an arbitrary such sub-subsequence, we will verify that $(\hat{\theta}^{(n_{mk})})_{t \in \mathbb{R}_+} \rightsquigarrow (\theta_t)_{t \in \mathbb{R}_+}$ 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 $(n_{mk})_{k \in \mathbb{N}}$ such that $(\hat{\theta}_t^{(n_{mk})})_{t \in \mathbb{R}_+} \rightsquigarrow (\theta_t)_{t \in \mathbb{R}_+}$ a.s., by Lemma 10 it must hold that $(\hat{\theta}_t^{(n)})_{t \in \mathbb{R}_+} \rightsquigarrow (\theta_t)_{t \in \mathbb{R}_+}$ in probability.

Now, let $(n_m)_{m \in \mathbb{N}}$ be an arbitrary subsequence of $\mathbb{N}$ such that $\Upsilon^{(n_m)} \overset{P}{\rightarrow} 0$. Let $\Omega$ denote the underlying

\[\text{\footnotesize \textsuperscript{8}Since every sub-subsequence is itself a subsequence, we can simplify our notation from here onward.}\]
probability space. Let
\[ \Omega^{(0)} = \bigcap_{i=1}^{3} \Omega^{(i)}, \]
\[ \Omega^{(1)} = \left\{ Y^{(n)} \to 0 \right\}, \]
\[ \Omega^{(2)} = \left\{ \max_{i \in [n]} \| \nabla \ell(\theta; X_i) \| \leq n^{1/p_2} \text{ a.b.f.o} \right\}, \]
\[ \Omega^{(3)} = \left\{ \max_{i \in [n]} \| \nabla \otimes^2 \ell(\cdot; X_i) \| \leq n^{1/p_3} \text{ a.b.f.o} \right\}. \]

By assumption, and by applying Lemma 12 to power functions of the form \( \alpha : t \mapsto t^p \) and random variables \( \| \nabla \ell(\theta; X_i) \| \) and \( \| \nabla \otimes^2 \ell(\cdot; X_i) \| \), \( \Omega^{(0)} \) is a sure set.

B.1.2 Additional notation used in the proof

We notate the increments of the localized iterative algorithms (given that \( \hat{\theta}_0^{(n)} = \theta \)) due to the Gaussian innovation \( (\xi) \), the gradient step contribution of the prior (\( \pi^{(0)} \)), the mini-batch gradient step based on the log-likelihood \( (\ell) \), and the total increment, respectively, as
\[
\Delta^{(n)}_\xi := u^{(n)} \sqrt{h_\beta^{-1} \Lambda} \xi_1,
\]
\[
\Delta^{(n)}_{\pi^{(0)}}(\theta) := \frac{h u^{(n)} \Gamma}{2n} \nabla \log \pi^{(0)} \left( \hat{\theta}^{(n)} + (w^{(n)})^{-1} \theta \right),
\]
\[
\Delta^{(n)}_\ell(\theta) := \frac{h u^{(n)} \Gamma}{2n} \sum_{j \in [b^{(n)}]} \nabla \ell \left( \hat{\theta}^{(n)} + (w^{(n)})^{-1} \theta; X_{i_1^{(n)}(j)} \right),
\]
\[
\Delta^{(n)}(\theta) := \Delta^{(n)}_\xi + \Delta^{(n)}_{\pi^{(0)}}(\theta) + \Delta^{(n)}_\ell(\theta).
\]

We define the sequence of operators \( A^{(n)} \) by
\[
[A^{(n)}f](\theta) = \alpha^{(n)} \left( \mathbb{E}^{X^{(n)}} \left[ f(\theta + \Delta^{(n)}(\theta)) \right] - f(\theta) \right). \tag{25}
\]
for all \( n \in \mathbb{N} \), and all \( f \in C^\infty_c(\mathbb{R}^d) \), where \( \alpha^{(n)} = n \). The generator of the (presumed, at this point) limiting OU process is given by
\[
[Af](\theta) = -\left( \frac{\zeta d}{2} \Delta + \Gamma \nabla f(\theta) \right) + \frac{1}{2} \left( \kappa \Lambda + \kappa_{ab} \Gamma \Lambda \Gamma' \right) : \nabla \otimes^2 f(\theta) \tag{26}
\]

B.1.3 How Proposition 8 is applied

Consider a single realization of \( X^{(0)} \in \Omega^{(0)} \). Our goal, now, is to apply Proposition 8, treating \( X^{(0)} \) as fixed. To do so, it suffices to show that for each \( f \in C^\infty_c(\mathbb{R}^d) \) we have
\[
\lim_{m \to \infty} \sup_{\theta \in \mathbb{R}^d} \left| [A^{(n_m)}f](\theta) - [Af](\theta) \right| = 0.
\]

For an arbitrary test function, \( f \in C^\infty_c(\mathbb{R}^d) \), 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
\[
\lim_{m \to \infty} \sup_{\theta \in K_1} \left| [A^{(n_m)}f](\theta) - [Af](\theta) \right| = 0.
\]

Then we will separately show that
\[
\lim_{m \to \infty} \sup_{\theta \in K_1} \left| [A^{(n_m)}f](\theta) - [Af](\theta) \right| = 0.
\]
B.1.4 Convergence away from the test function support

For all $\vartheta \in K_0$, $f(\vartheta) = 0$, $\nabla f(\vartheta) = 0$, and $\nabla^2 f(\vartheta) = 0$. Therefore, for any $K_1 \supseteq K_0$,

$$
\sup_{\vartheta \in K_1^c} \left| \left( A^{(nm)} f \right)(\vartheta) - [Af](\vartheta) \right| \leq \alpha^{(nm)} \|f\|_{\infty} \sup_{\vartheta \in K_1^c} \mathbb{P} \left( \vartheta + \Delta^{(nm)}(\vartheta) \in K_0 \right) \tag{27}
$$

Let $R_0 = \sup_{\vartheta \in K_0} \|\vartheta\|$. Let $K_1 = \{ \vartheta \in \mathbb{R}^d \text{ s.t. } \|\vartheta\| \leq 2R_0 + 2c_0 \}$, where

$$
c_0 = c_h \|\Gamma\| \left( 3 + \left( \|\nabla \log \pi_0(0)\| \right) \right) + \sqrt{c_h/c_2 \|A\|}.
$$

Then, using Eq. (27) and $\Delta^{(nm)}(\vartheta) = \Delta_{\ell}^{(nm)}(\vartheta) + \Delta_{\pi}^{(n_0)}(\vartheta) + \Delta_{\ell}^{(nm)}(\vartheta)$,

$$
\sup_{\vartheta \in K_1^c} \left| \left( A^{(nm)} f \right)(\vartheta) - [Af](\vartheta) \right| \leq \alpha^{(nm)} \|f\|_{\infty} \sup_{\|\vartheta\| > 2R_0 + 2c_0} \mathbb{P} \left( \vartheta + \Delta^{(nm)}(\vartheta) \right) \leq R_0 \tag{28}
$$

For $\vartheta \in K_1^c$, using the assumption that $\nabla \log \pi_0$ is $L_0$-Lipschitz and $h^{(n)} = c_h n^h$ and $w^{(n)} = n^w$,

$$
\left\| \Delta_{\pi}^{(n_0)}(\vartheta) \right\| \leq \frac{h^{(nm)} w^{(nm)} \|\Gamma\|}{2n_m} \left\| \nabla \log \pi_0 \left( \vartheta \right) \right\| \left( \vartheta + \left( w^{(nm)} \right)^{-1} \vartheta \right)
$$

$$
\leq \frac{h^{(nm)} w^{(nm)} \|\Gamma\|}{2n_m} \left[ \left\| \nabla \log \pi_0 \left( \vartheta \right) \right\| + L_0 \left\| \vartheta \right\| \right] \leq \frac{c_h n^m - b - 1}{2} \left( \left\| \nabla \log \pi_0 \left( \vartheta \right) \right\| + L_0 \left\| \vartheta \right\| \right),
$$

and similarly

$$
\left\| \Delta_{\ell}^{(nm)}(\vartheta) \right\|
$$

$$
\leq \frac{h^{(nm)} w^{(nm)} \|\Gamma\|}{2b^{(nm)}} \sum_{j \in [b^{(nm)}]} \left\| \nabla \ell \left( \vartheta, X_{1_{(n_m)}(j)} \right) \right\|
$$

$$
\leq \frac{c h^{(nm)} - b - 1}{2} \left( \left\| \nabla \ell \left( \vartheta, X_{1_{(n_m)}(j)} \right) \right\| + L \left( X_{1_{(n_m)}(j)} \right) \left\| \vartheta \right\| \right) \leq \frac{c h^{(nm)} - b - 1}{2} \left( \left\| \nabla \ell \left( \vartheta, X_{1_{(n_m)}(j)} \right) \right\| + L \left( X_{1_{(n_m)}(j)} \right) \left\| \vartheta \right\| \right)
$$

where we define the (random) Lipschitz constants $L(X_i)$, $L_s(X_{(n_m)})$, and $L(X_{(n_m)})$ by:

$$L(X_i) := \left\| \nabla^2 \ell : X_i \right\|_{\infty},
$$

$$L_s(X_{(n_m)}) := \max_{1 \leq n_m} \left\| \nabla \ell \left( \vartheta, X_{1_{(n_m)}(j)} \right) \right\|, \text{ and}
$$

$$L(X_{(n_m)}) := \max_{1 \leq n_m} L(X_i).$$
Using that \( X^{(n)} \in \Omega^{(0)} \), so that \( Y^{(nm)} \to 0 \) etc., if \( m \) is large enough that all of the following hold:

\[
\sup_{m' \geq m} Y^{(nm')} \leq \min(1, L_0^{-1}),
\]

\[
1 \geq \sup_{m' \geq m} \frac{L_s(X^{(n_{m'})})}{n_{m'}^{1/p_3}},
\]

\[
n_{m} \geq \max((2c_h \| \Gamma \|^{1/(p_3 - h)}, (2c_h L_0 \| \Gamma \|)^{\frac{1}{p_3 - h}}), \quad \text{and}
\]

\[
1 \geq \sup_{m' \geq m} \frac{L(X^{(n_{m'})})}{n_{m'}^{1/p_3}};
\]

then, using that \( 0 < \varpi < 1 \),

\[
\left\| \Delta^{(n_m)}_{\pi(0)}(\vartheta) \right\| \leq \frac{c_h \| \Gamma \|}{2} \left( \left\| \nabla \log \pi^{(0)}(\theta_*) \right\| + 1 \right) + \frac{1}{4} \| \vartheta \|,
\]

and

\[
\left\| \Delta^{(n_m)}_{\pi}(\vartheta) \right\| \leq \frac{c_h n_{m}^{h + \varpi}}{2} \left( n_{m}^{1/p_2} + n_{m}^{1/p_3} Y^{(n_m)} + n_{m}^{1/p_3 - \varpi} \| \vartheta \| \right)
\]

\[\leq c_h \| \Gamma \| \left( n_{m}^{1/p_2 - h + \varpi} + n_{m}^{1/p_3 - h + \varpi} Y^{(n_m)} + n_{m}^{1/p_3 - h} \| \vartheta \| \right),\]

\[\leq c_h \| \Gamma \| + \frac{1}{4} \| \vartheta \|.
\]

Therefore, for \( \vartheta \in K_1^{c} \) (and hence \( \| \vartheta \| > 2R_0 + 2c_0 \)),

\[
\| \vartheta \| - \left\| \Delta^{(n_m)}_{\pi(0)}(\vartheta) \right\| - \left\| \Delta^{(n_m)}_{\pi}(\vartheta) \right\| - R_0
\]

\[\geq \frac{1}{2} \| \vartheta \| - \frac{c_h \| \Gamma \|}{2} \left( 3 + \left\| \nabla \log \pi^{(0)}(\theta_*) \right\| \right) - R_0
\]

\[\geq \sqrt{c_h/c_\beta} \| \vartheta \|.
\]

Therefore, combining this with Eq. (28) and the definition of \( \Delta^{(n_m)}_{\pi}(\vartheta) \),

\[
\lim_{m \to \infty} \sup_{\vartheta \in K_1^{c}} \left| [A^{(n_m)} f](\vartheta) - [Af](\vartheta) \right| \leq \lim_{m \to \infty} \alpha^{(n_m)} \| f \|_{\infty} \mathbb{P}^{X^{(0)}}(\| \xi_1 \| \geq n_{m}^{h/2 + t/2 - \varpi})
\]

\[\leq \lim_{m \to \infty} \alpha^{(n_m)} \| f \|_{\infty} d \mathbb{P}^{X^{(0)}}(\| \xi_{1,1} \| \geq \frac{1}{\sqrt{d}} n_{m}^{h/2 + t/2 - \varpi})
\]

\[\leq \lim_{m \to \infty} 2n_{m}^{a} \| f \|_{\infty} d \exp(-n_{m}^{h/2 + t/2 - \varpi}/2d)
\]

\[= 0.
\]

**B.1.5 Taylor expansion near the test function support**

Recalling the definition of \( A^{(n_m)} \) in Eq. (25), using the definition of the time-scaling factor \( \alpha^{(n)} = n^a \), taking a second-order Taylor expansion of the test function \( f \in C_c^{\infty} \), and applying the decomposition
\[ \Delta^{(n_m)}(\vartheta) = \Delta^{(n_m)}_\xi(\vartheta) + \Delta^{(n_m)}_\pi(\vartheta) + \Delta^{(n_m)}_\ell(\vartheta), \]

\[ [A^{(n_m)}f](\vartheta) = \Delta^{(n_m)}_\xi(f(\vartheta + \Delta^{(n_m)}(\vartheta))) - f(\vartheta) \]

\[ = n_m E^{(n)} \left[ \nabla f(\vartheta), \Delta^{(n_m)}(\vartheta) \right] + n_m E^{(n)} \left[ \nabla f(\vartheta), \Delta^{(n_m)}_\pi(\vartheta) \right] + n_m E^{(n)} \left[ \nabla f(\vartheta), \Delta^{(n_m)}_\ell(\vartheta) \right] \]

\[ + \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta^{(n_m)}_\ell(\vartheta) + \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta^{(n_m)}_\pi(\vartheta) + \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta^{(n_m)}(\vartheta) \]

for some \( S \in [0,1] \) depending on \( f, \vartheta, \Delta^{(n_m)}(\vartheta) \), where \( \nabla^{\otimes 3} f(\vartheta) \) is the trilinear form of third order partials of \( f \) at \( \vartheta \) (and hence is linear in each of its three arguments). Terms that are linear in \( \Delta^{(n_m)}_0(\vartheta) \) 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

\[ [A f](\vartheta) = -\left( \frac{c_d}{2} \Gamma J_\vartheta \vartheta, \nabla f(\vartheta) \right) + \frac{c_g}{2} \Lambda : \nabla^{\otimes 2} f(\vartheta) + \frac{c_h}{2} \Gamma Z \vartheta : \nabla^{\otimes 2} f(\vartheta) . \]

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 eliminating terms which are linear in \( \Delta^{(n_m)}_0(\vartheta) \), and thus have mean 0, the difference of approximate and limiting generator applied to the test function can be expressed as

\[ \left| [A^{(n_m)}f](\vartheta) - [A f](\vartheta) \right| \]

\[ \leq \left| [1, \pi]^{(n_m)}(\vartheta) + [2, \pi]^{(n_m)}(\vartheta) + [2, \ell \pi]^{(n_m)}(\vartheta) \right| + \left| [1, \ell]^{(n_m)}(\vartheta) - [I, \Gamma J_\vartheta](\vartheta) \right| + \left| [2, \ell \xi]^{(n_m)}(\vartheta) - [II, \Lambda](\vartheta) \right| + \left| [2, \ell \ell]^{(n_m)}(\vartheta) - [II, \Gamma Z \vartheta](\vartheta) \right| + \left| [3, R]^{(n_m)}(\vartheta) \right| . \]

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.

B.1.6 Terms that do not contribute to the limit

\[
\left| 1.\pi^{(0)}(n_m^m, \vartheta) \right| = n_m^a \left| \mathbb{E}X^{(i)} \left\langle \nabla f(\vartheta), \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\rangle \right|
\]

\[
\leq \frac{c_h n_m^{a-b+m-1} ||\Gamma||}{2} \left| \mathbb{E}X^{(i)} \left\langle \nabla f(\vartheta), \nabla \log \pi^{(0)}(\tilde{\vartheta}^{(n_m)} + n_m^{-w} \vartheta) \right\rangle \right|
\]

\[
\leq \frac{c_h n_m^{a-b+m-1} ||\Gamma||}{2} ||\nabla f||_\infty \left( \left\| \nabla \log \pi^{(0)}(\vartheta^\star) \right\| + L_0 \left( \Upsilon^{(n_m)} + \frac{2R_0 + 2c_0}{n_m^{w}} \right) \right),
\]

which vanishes uniformly on \( K_1 \), since \( a + w - h - 1 \leq w - 1 < 0 \).

\[
\left| 2.\pi^{(0)}\pi^{(0)}(n_m^m, \vartheta) \right|
\]

\[
= n_m^a \left| \mathbb{E}X^{(i)} \left\langle \frac{1}{2} \nabla \otimes^2 f(\vartheta) \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta), \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\rangle \right|
\]

\[
\leq n_m^a \left\| \nabla \otimes^2 f \right\|_\infty \left( \frac{c_h n_m^{a-b-1} ||\Gamma||}{2} \right)^2 \nonumber
\]

\[
\times \left( \left\| \nabla \log \pi^{(0)}(\vartheta^\star) \right\| + L_0 \left\| \tilde{\vartheta}^{(n_m)} - \vartheta^\star \right\| + L_0 \frac{2R_0 + 2c_0}{n_m^{w}} \right)^2
\]

which vanishes uniformly since \( a + 2w - 2h \leq (2w - 2) - h < 0 \) (which follows from \( h \geq a \) and \( w < 1 \)).

\[
\left| 2.2f\pi^{(0)}(n_m^m, \vartheta) \right|
\]

\[
= n_m^a \left| \mathbb{E}X^{(i)} \left\langle \frac{1}{2} \nabla \otimes^2 f(\vartheta) \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta), \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) \right\rangle \right|
\]

\[
\leq 2n_m^a \left\| \nabla \otimes^2 f \right\|_\infty \left( \frac{c_h n_m^{a-b-1} ||\Gamma||}{2} \right) \left( \frac{c_h n_m^{a-b} ||\Gamma||}{2} \right)
\]

\[
\times \left( \left\| \nabla \log \pi^{(0)}(\vartheta^\star) \right\| + L_0 \Upsilon^{(n_m)} + L_0 \frac{2R_0 + 2c_0}{n_m^{w}} \right) \nonumber
\]

\[
\times \left( n_m^{1/p_2} + n_m^{1/p_3} \Upsilon^{(n_m)} + n_m^{1/p_2 - w} \right)
\]

which vanishes uniformly due to the assumptions of the relationship between \( h, a, w, p_3, p_2 \) under each
assumption.
B.1.7 Convergence of the drift term

Third, using that \( \sum_{i \in [n_m]} \nabla \ell \left( \tilde{\vartheta}^{(n_m)}; X_i \right) = 0 \),

\[
\begin{align*}
\left[1.\ell\right]^{(n_m)} (\vartheta) &= n_m a \mathbb{E}^{x(0)} \left\langle \nabla f(\vartheta), \Delta^{(n_m)} (\vartheta) \right\rangle \\
&= \mathbb{E}^{x(0)} \left\langle \nabla f(\vartheta), \frac{c_h n_m^{a+m-b} \Gamma}{2 h^{(n_m)}} \sum_{j \in [h^{(n_m)}]} \nabla \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{1}{n_m^b} \vartheta; X_i^{(n_m)} (j) \right) \right\rangle \\
&= \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \sum_{i \in [n_m]} \nabla \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{1}{n_m^b} \vartheta; X_i \right) \right\rangle \\
&= \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \left( \int_0^1 \sum_{i \in [n_m]} \nabla^2 \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{s}{n_m^b} \vartheta; X_i \right) ds \right) \right\rangle \\
\end{align*}
\]  

(29)

Now, for all \( n_m \) large enough that \( r_{\mathcal{J}, n_m} \geq R_0 + c_0 \)

\[
\left\| \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \left( \int_0^1 \sum_{i \in [n_m]} \nabla^2 \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{s}{n_m^b} \vartheta; X_i \right) ds \right) \right\rangle \vartheta \right\| \\
\leq c_h \| \Gamma \| \| \nabla f \|_\infty \left( R_0 + c_0 \right) \left\| \int_0^1 \left[ \sum_{i \in [n_m]} \nabla^2 \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{s}{n_m^b} \vartheta; X_i \right) + \mathcal{J}_* \right] ds \right\| \\
\leq c_h \| \Gamma \| \| \nabla f \|_\infty \left( R_0 + c_0 \right) \cdot \mathcal{T}^{(n_m)},
\]

and thus vanishes uniformly on \( K_1 \).

When \( a > b \), so \( c_0 = 0 \) and hence \( \left[1.\ell \mathcal{J}_* \right] (\vartheta) = 0 \) (where \( \left[1.\ell \mathcal{J}_* \right] \) is the drift term appearing in the definition of the limiting generator \( \mathcal{A} \) in Eq. (26)), then the drift term will be inactive in the limit. We show this by using the fact that \( \left[1.\ell\right]^{(n_m)} (\vartheta) \) is a vanishing distance from a sequence that vanishes:

\[
\left| \left[1.\ell\right]^{(n_m)} (\vartheta) - \left[1.\ell \mathcal{J}_* \right] (\vartheta) \right| \\
\leq n_m^{b-a} \left| \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \left( \int_0^1 \sum_{i \in [n_m]} \nabla^2 \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{s}{n_m^b} \vartheta; X_i \right) ds \right) \right\rangle \vartheta \right| \\
+ n_m^{b-a} \left| \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \mathcal{J}_* \vartheta \right\rangle \right|;
\]

and hence vanishes uniformly on \( K_1 \).

When \( a = b \), then the drift term is active in the limit, and we show that \( \left[1.\ell\right]^{(n_m)} (\vartheta) \) converges to the drift term from the limiting process \( \left[1.\ell \mathcal{J}_* \right] \) (\( \vartheta \)):

\[
\left| \left[1.\ell\right]^{(n_m)} (\vartheta) - \left[1.\ell \mathcal{J}_* \right] (\vartheta) \right| \\
= n_m^{b-a} \left| \left\langle \frac{c_h \Gamma^+}{2} \nabla f(\vartheta), \left( \int_0^1 \sum_{i \in [n_m]} \nabla^2 \ell \left( \tilde{\vartheta}^{(n_m)} + \frac{s}{n_m^b} \vartheta; X_i \right) ds \right) \right\rangle \vartheta \right| \\
\]

vanishes uniformly on \( K_1 \).
B.1.8 Convergence of the diffusion term corresponding to Gaussian noise

\[ \left| [2, \xi_n^{(m)}] (\vartheta) - [\Pi, A] (\vartheta) \right| = \left| \frac{\partial}{\partial n} \mathbb{E} \left( \frac{1}{2} \nabla^2 f(\vartheta) \Delta \xi_n^{(m)}(\vartheta), \Delta \xi_n^{(m)}(\vartheta) \right) - \frac{c_h}{2c_\beta} A : \nabla^2 f(\vartheta) \right| \]

If \( a + 2m - h - t = 0 \) then, the corresponding diffusion term is active in the limit. Using the definition of \( \Delta \xi_n^{(m)} \) and that \( \beta^{(n)} = c_\beta n^t, \beta_h = c_h n^h, \) and \( \beta_w = n^w \)

\[ \left| [2, \xi_n^{(m)}] (\vartheta) - [\Pi, A] (\vartheta) \right| \leq \frac{c_h}{2c_\beta} \left| n_m^{a + 2m - h - t} \mathbb{E} \left( \nabla^2 f(\vartheta) \sqrt{A} \xi_1 \right) \right| = 0 \]

If \( a + 2m - h - t < 0 \) then the corresponding diffusion term is inactive in the limit, and so \( [\Pi, A] (\vartheta) = 0 \). In that case we show that \( [2, \xi_n^{(m)}] (\vartheta) \) vanishes uniformly.

\[ \left| [2, \xi_n^{(m)}] (\vartheta) - [\Pi, A] (\vartheta) \right| \leq \frac{c_h}{2c_\beta} \left| n_m^{a + 2m - h - t} \left\| A \right\|_F \left\| \nabla^2 f \right\|_F \right|, \]

which vanishes uniformly.

B.1.9 Convergence of the diffusion term corresponding to minibatch noise

\[ \left| [2, \xi_n^{(m)}] (\vartheta) - [\Pi, \Gamma \xi_n, \Gamma'] (\vartheta) \right| = \left| \frac{\partial}{\partial n} \mathbb{E} \left( \frac{1}{2} \nabla^2 f(\vartheta) \Delta \xi_n^{(m)}(\vartheta), \Delta \xi_n^{(m)}(\vartheta) \right) - \frac{c_n}{2} \Gamma \xi_n, \Gamma' : \nabla^2 f(\vartheta) \right| \]

\[ \leq \frac{\left\| \nabla^2 f \right\|_F}{2} \left( \left| n_m^{a} \mathbb{E} \left( \left( \Delta \xi_n^{(m)}(\vartheta) \right)^{\otimes 2} \right) : \nabla^2 f(\vartheta) - \frac{c_n}{2} \Gamma \xi_n, \Gamma' : \nabla^2 f(\vartheta) \right| \right) \]

\[ \leq \frac{\sqrt{d}}{2} \left\| \frac{\nabla^2 f}{\nabla^2 f} \right\| \left( \left| n_m^{a} \mathbb{E} \left( \left( \Delta \xi_n^{(m)}(\vartheta) \right)^{\otimes 2} \right) - \frac{c_n}{2} \Gamma \xi_n, \Gamma' \right| \right) \]
Now,
\[
\mathbb{E}^\mathcal{X}_n b_m \left( \Delta_{(n_m)} (\theta) \right) \otimes^2 \\
= \frac{c_{b_m}^2 + 2 a + 2 b - 2 b}{4 b_m^2} \Gamma \left( \mathbb{E}^\mathcal{X}_n \sum_{j \in [b_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j}^{(n_m)}(j) \right) \right) \otimes^2 \Gamma \\
+ \frac{c_{b_m}^2 + 2 a + 2 b - 2 b}{4 b_m^2} \Gamma \left( \mathbb{E}^\mathcal{X}_n \sum_{j \in [b_m]} \sum_{j' \in [b_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j}^{(n_m)}(j) \right) \right) \otimes \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j}^{(n_m)}(j') \right) \otimes \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j}^{(n_m)}(j) \right) \\
= \frac{c_{b_m}^2 + 2 a + 2 b - 2 b}{4 b_m^2} \Gamma \left( \frac{1}{n_m^2} \sum_{i \in [n_m]} \sum_{j' \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \otimes \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \right) \\
= \frac{c_{b_m}^2 + 2 a + 2 b - 2 b}{4 b_m^2} \Gamma \left( \frac{1}{n_m^2} \sum_{i \in [n_m]} \sum_{j' \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \right) \otimes^2 \Gamma \\
= \frac{c_{b_m}^2 + 2 a + 2 b - 2 b}{4 b_m^2} \Gamma \left( \frac{1}{n_m^2} \sum_{i \in [n_m]} \int_0^1 \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) ds \vartheta \right) \otimes^2 \Gamma \\

Thus, if \( a + 2 b - 2 b = 0 \), so that \( c_{ab} \neq 0 \) and the corresponding term is active in the limit, and the mini-batches are drawn with replacement, then combining the past several equations gives:

\[
\left\| 2 \nabla \ell \left( \hat{\theta}^{(n_m)} \right) \cdot [22.1 \Gamma, \Gamma'] \left( \vartheta \right) \right\| \\
\leq \frac{\sqrt{d} \| \Gamma \| \| \nabla \ell \|_\infty}{2} \left\| \frac{c_{b_m}^2}{4 b_m^2} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \right) \right\| + \frac{\sqrt{d} \| \Gamma \| \| \nabla \ell \|_\infty}{8} \left\| \frac{1}{n_m} \sum_{i \in [n_m]} \int_0^1 \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \right\| \otimes^2 \vartheta \\
+ \frac{\sqrt{d} \| \Gamma \| \| \nabla \ell \|_\infty}{8} \left\| \frac{1}{n_m} \sum_{i \in [n_m]} \int_0^1 \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m^2} \vartheta; X_{i_j} \right) \right\| \otimes^2 \vartheta \\
\right\|
\]
For $\vartheta \in K_1$, and for all $n_m$ large enough that $r_{\mathcal{I}, n_m} \geq R_0 + c_0$

$$
\begin{align*}
&= n_m^{-2m} \left\| \frac{1}{n_m} \sum_{i \in [n_m]} \int_0^1 \nabla \otimes \ell \left( \hat{\vartheta}^{(n_m)} + \frac{s}{n_m} \vartheta; X_i \right) ds \vartheta \right\|^2 \\
&\leq \frac{(2R_0 + 2c_0)^2}{n_m^{2m}} \left( \| \mathcal{I} \| + \tau^{(n_m)} \right)^2,
\end{align*}
$$

which vanishes uniformly.

Since the mini-batches are drawn with replacement, using the definition of $c_{ab}$, for all $n_m$ large enough that $r_{\mathcal{I}, n_m} \geq R_0 + c_0$

$$
\begin{align*}
&\leq \frac{c_h^2}{4c_b} \left\| \frac{c_b n_m^b}{c_h n_m^b} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\vartheta}^{(n_m)} + \frac{1}{n_m} \vartheta; X_i \right) \right) \right\| - c_{ab} \mathcal{I}_s \\
&\leq \frac{c_h^2}{4c_b} \left\| \frac{c_b n_m^b}{c_h n_m^b} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\vartheta}^{(n_m)} + \frac{1}{n_m} \vartheta; X_i \right) \right) \right\| - \mathcal{I}_s \\
&\quad + \left| \frac{c_h^2}{4c_b} \frac{c_b n_m^b}{c_h n_m^b} - \frac{c_h^2}{4c_b} \right| \| \mathcal{I}_s \| \\
&\leq \frac{c_h^2}{4c_b} \left\| \frac{c_b n_m^b}{c_h n_m^b} \tau^{(n_m)} \right\| + \left| \frac{c_h^2}{4c_b} \frac{c_b n_m^b}{c_h n_m^b} - \frac{c_h^2}{4c_b} \right| \| \mathcal{I}_s \|.
\end{align*}
$$

And, if $a + 2m - 2h - b < 0$ and the mini-batches are drawn with replacement, so that $c_{ab} = 0$, and the corresponding diffusion term is inactive in the limit and $[\Pi.\Gamma\mathcal{I}_s \Gamma'](\vartheta) = 0$, then

$$
\begin{align*}
\left| \mathbb{E}^{(n)} n_m \left( \Delta^{(n)} (\vartheta) \right) \right| &\leq \left| \mathbb{E}^{(n)} n_m \left( \Delta^{(n)} (\vartheta) \right) \right| - n_m^{a+2m-2h-b} \frac{c_h^2}{4c_b} \mathcal{I}_s \\
&\quad + n_m^{a+2m-2h-b} \frac{c_h^2}{4c_b} \| \mathcal{I}_s \|
\end{align*}
$$

which vanishes uniformly by the previous arguments.

Therefore, when the mini-batches are drawn with replacement, we find that

$$
\left| \mathbb{E}^{(n)} n_m \left( \Delta^{(n)} (\vartheta) \right) - \left[ 2.\ell^{(n)} (\vartheta) - \left[ \Pi.\Gamma\mathcal{I}_s \Gamma' \right] (\vartheta) \right) \right|
$$

vanishes uniformly on $K_1$. 

33
If the mini-batches are drawn without replacement.

\[
E\!X^{(n)} \sum_{j \in [b^{(n_m)}]} \sum_{j' \in [b^{(n_m)}] \setminus \{j\}} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_{1^{(n_m)}(r_j)} \right) \otimes \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_{1^{(n_m)}(r_{j'})} \right)
\]
\[
= b^{(n_m)}(b^{(n_m)} - 1) \frac{n_m(n_m - 1)}{n_m} \sum_{i \in [n_m], i' \in [n_m] \setminus \{i\}} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \otimes \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_{i'} \right)
\]
\[
= b^{(n_m)}(b^{(n_m)} - 1) \frac{n_m(n_m - 1)}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \otimes 2
\]
\[
= b^{(n_m)}(b^{(n_m)} - 1) \frac{n_m}{n_m - 1} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \right) \otimes 2
\]
and so,

\[
E\!X^{(n_m)} \left( \Delta^{(n_m)} (\bar{\theta}) \right) \otimes 2
\]
\[
= \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \Gamma \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \otimes 2 \right) \Gamma'
\]
\[
+ \frac{c_h^2}{4b^{(n_m)}} \Gamma \left( b^{(n_m)}(b^{(n_m)} - 1) \frac{n_m}{n_m - 1} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \right) \otimes 2 \right) \Gamma'
\]

In this case, for all \( n_m \) large enough that \( r_{\mathcal{I}, n_m} \geq R_0 + c_0 \)

\[
\left\| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \otimes 2 \right) - c_{nb} I_{\mathcal{I}} \right\|
\]
\[
\leq \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \left\| \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \hat{\theta}^{(n_m)} + \frac{1}{n_m} \bar{\theta}; X_i \right) \otimes 2 \right) - I_{\mathcal{I}} \right\|
\]
\[
+ \left\| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} - c_{nb} \right\| I_{\mathcal{I}}
\]
\[
\leq \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} \Gamma^{(n_m)} + \left\| \frac{c_h^2}{4b^{(n_m)}} \frac{n_m - b^{(n_m)}}{n_m - 1} - c_{nb} \right\| I_{\mathcal{I}}
\]

Thus, when the mini-batches are drawn without replacement, we find that

\[
\left\| [2, \ell]^{(n_m)} (\bar{\theta}) - [\Pi, \Gamma, \bar{\theta}] (\bar{\theta}) \right\|
\]
vanishes uniformly on \( K_1 \).
B.1.10 Convergence of the Remainder Term

\[ |3.R^{(n_m)}(\theta)| \]
\[ = n_m^a \mathbb{E}^{(n_m)} \left[ \frac{1}{6} \left( \nabla^3 f(\theta) + S\Delta^{(n_m)}(\theta) \right) \left( \Delta^{(n_m)}(\theta), \Delta^{(n_m)}(\theta), \Delta^{(n_m)}(\theta) \right) \right] \]
\[ \leq \frac{n_m^a}{6} \left\| \nabla^3 f \right\|_\infty \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \right] \]
\[ \leq \frac{27n_m^a}{6} \left\| \nabla^3 f \right\|_\infty \left( \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 + \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \right] + \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \right] \right) , \]

Now
\[ \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \right] \leq \left( \frac{c_h n_m^{-h-1+2m}}{2c_\beta} \left\| \Lambda \right\| \right)^{3/2} \mathbb{E}^{(n_m)} \left[ \left\| \xi_1 \right\|^3 \right] \]
\[ = n_m^{-3/2(\alpha+2\alpha)} \left( \frac{c_h}{2c_\beta} \left\| \Lambda \right\| \right)^{3/2} 2^{3/2} \Gamma \left( \frac{d+3}{2} \right) \left( \frac{2}{\Gamma \left( \frac{d}{2} \right)} \right) , \]

where \( \Gamma \) is the gamma function. Note that \( \alpha - 3/2 (h + t - 2m) \leq -1/2 (h + t - 2m) \leq -a/2 < 0 \)

Second,
\[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \leq \left( \frac{c_h n_m^{-h-1+2m}}{2c_\beta} \left\| \Gamma \right\| \right)^{3/2} \left( \nabla \log \pi^{(n_m)}(\theta) \right) + L_0 \left\| \theta^{(n_m)} - \theta_* \right\| + L_0 \frac{2R_0 + 2\lambda_0}{n_m} . \]

Note that \( a - 3h + 3m - 3 \leq -2h - 3(1 - m) < 0 . \)

Third,
\[ \mathbb{E}^{(n_m)} \left[ \left\| \Delta^{(n_m)}(\theta) \right\|^3 \right] \leq \left( \frac{c_h n_m^{-h+1+2m}}{2} \left\| \Gamma \right\| \right)^{3/2} \left( n_m^{1/p_2} + n_m^{1/p_3} T^{(n_m)}(\theta_*), n_m^{1/p_2} - n_m^{1/p_3} \right)^{3/2} \]
\[ \leq \left( \frac{c_h}{2} \left\| \Gamma \right\| \right)^{3/2} \left( n_m^{1/p_2-b+m} + n_m^{1/p_2-b+m} T^{(n_m)} + n_m^{1/p_3-b} \right)^{3/2} . \]

Therefore, \( |3.R^{(n_m)}(\theta)| \) vanishes uniformly. \( \Box \)

C Proof of Corollary 3

Proof of Corollary 3. To verify that that the stationary measures, \( \nu^{(n_m)} \) of \( T^{(n_m)} \) converge weakly in probability to \( \nu \), we need to verify that every sub-subsequence \( \nu^{(n_{mk^j})} \) has a sub-sub-subsequence \( \nu^{(n_{mk^j}k)} \) converging weakly to \( \nu \) almost surely. Since weak convergence of probability measures is metrizable, then applying Lemma 10 yields the desired result.

By the second part of Theorem 11, every sub-subsequence of \( \{ T^{(n_m)} \}_{m \in \mathbb{N}} \); \( \{ T^{(n_{mk^j})} \}_{k \in \mathbb{N}} \); has a further sub-sub-subsequence, \( \{ T^{(n_{mk^j}k)} \}_{j \in \mathbb{N}} \) such that with probability 1, \( T^{(n_{mk^j}k)} \to T \) on \( C[\mathbb{R}^d] \) for all \( t > 0 \).

Applying Ethier and Kurtz [2009, Part 4, Theorem 9.10], we have that every weak limit of \( \{ \nu^{(n_{mk^j}k)} \}_{j \in \mathbb{N}} \) is stationary for \( T \). As a consequence of the assumption that the spectrum of \( \Gamma J(\theta_*) \) 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 [2014]), \( \nu = N(0, Q_\infty) \). Thus every weak limit of \( \{ \nu^{(n_{mk^j}k)} \}_{j \in \mathbb{N}} \) must be \( \nu \).

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. \( \Box \)
D  Sufficient conditions for Assumptions 4 and 5

In this section we provide some sufficient conditions that ensure Assumptions 4 and 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 13 (Sufficient conditions for Assumption 4). Each of the following imply Assumption 4.

a) there exists a $\delta_1 > 0$ with $\sup_{\theta \in B(\theta_*)} \| \frac{1}{n} \sum_{i \in [n]} \nabla^2 \ell(\theta; X_i) + J(\theta) \| \overset{p}{\to} 0$ and $J$ is continuous at $\theta_*$,

b) $\{ \nabla^2 \ell(\cdot; x) \mid x \in X \}$ is equicontinuous at $\theta_*$,

c) there exists a $\delta_1 > 0$ with

$$\mathbb{E} \left[ \sup_{\theta \in B(\theta_*)} \frac{\| \nabla^2 \ell(\theta; X_1) - \nabla^2 \ell(\theta_*; X_1) \|}{\| \theta - \theta_* \|} \right] < \infty,$$

Proof of Proposition 13.

a) Let $r, J, n = \delta_1 n^{1/2}/2$. Then $B \left( \hat{\theta}(n), r, J, n^{1/2} \right) \subseteq B \left( \hat{\theta}(n), \delta_1/2 \right)$.

Given that $\hat{\theta}(n) \overset{p}{\to} \theta_*$, any subsequence of indices $n_m$ has a further sub-subsequence of indices $n_{mk}$ where both $\hat{\theta}(n_{mk}) \to \theta_*$ and

$$\sup_{\theta \in B(\theta_*)} \left\| \frac{1}{n_{mk}} \sum_{i \in [n_{mk}]} \nabla^2 \ell(\theta; X_i) + J(\theta) \right\| \overset{a.s.}{\to} 0.$$

Then there is a $k_0$ such that if $k \geq k_0$ then $\| \hat{\theta}(n_{mk}) - \theta_* \| \leq \delta_1/2$. Therefore if $k \geq k_0$ then $B \left( \hat{\theta}(n_{mk}), r, J, n^{1/2} \right) \subseteq B (\theta_*, \delta_1)$.

Thus, for $k \geq k_0$,

$$\sup_{\theta \in B(\hat{\theta}(n_{mk}), r, J, n^{1/2})} \left\| J(n_{mk})(\theta) - J(\theta) \right\| \leq \sup_{\theta \in B(\hat{\theta}(n_{mk}), r, J, n^{1/2})} \left\| J(n_{mk})(\theta) - J(\theta) \right\| + \sup_{\theta \in B(\hat{\theta}(n_{mk}), \delta_1 n^{1/2})} \left\| J(\theta) - J(\theta_*) \right\| \overset{\text{a.s.}}{\to} 0.$$

Therefore, every subsequence of $S_n = \sup_{\theta \in B(\hat{\theta}(n), r, J, n^{1/2})} \left\| J(n)(\theta) - J(\theta) \right\|$ 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 $\rho_{J_+} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t \to 0} \rho_{J_+}(t) = 0$, and

$$\sup_{x \in X} \sup_{\theta \in B(\theta_*)} \left\| \nabla^2 \ell(\theta; x) - \nabla^2 \ell(\theta_*; x) \right\| \leq \rho_{J_+}(\delta).$$
Let $r_{\mathcal{J}, n} = n^{m/2}$. Then
\[
\sup_{\theta \in B(\hat{\theta}(n), r_{\mathcal{J}, n}/n^{m})} \left\| \tilde{\mathcal{F}}(n)(\theta) - \mathcal{J}(\theta) \right\|
\leq \sup_{\theta \in B(\hat{\theta}(n), n^{-m/2})} \left\| \tilde{\mathcal{F}}(n)(\theta) - \tilde{\mathcal{F}}(n)(\theta^*) \right\| + \left\| \tilde{\mathcal{F}}(n)(\theta^*) - \mathcal{J}(\theta^*) \right\|
\leq \sup_{\theta \in B(\theta^*, \|\hat{\theta}(n) - \theta^*\| + n^{-m/2})} \left\| \tilde{\mathcal{F}}(n)(\theta) - \tilde{\mathcal{F}}(n)(\theta^*) \right\| + \left\| \tilde{\mathcal{F}}(n)(\theta^*) - \mathcal{J}(\theta^*) \right\|
\leq \rho_{\mathcal{J}} \left( \|\hat{\theta}(n) - \theta^*\| + n^{-m/2} \right) + \|\mathcal{J}(\theta^*) - \mathcal{J}(\theta^*)\|
\overset{p}{\rightarrow} 0.
\]

In the last step we used that the first term vanishes in probability because $\hat{\theta}(n) \overset{p}{\rightarrow} \theta^*$, and the second term vanishes in probability by the weak law of large numbers.

c) Let
\[
Q_n = \frac{1}{n} \sum_{i \in [n]} \left[ \sup_{\theta \in B_{i}(\theta_{i})} \left\| \frac{\nabla^{\otimes 2}\ell(\theta; X_{i})}{\|\theta - \theta_{i}\|} \right\| \right], \quad \text{and}
\]
\[
q = \mathbb{E} \left[ \sup_{\theta \in B_{i}(\theta_{i})} \left\| \frac{\nabla^{\otimes 2}\ell(\theta; X_{i})}{\|\theta - \theta_{i}\|} \right\| \right].
\]

By the weak law of large numbers, $Q_n \overset{p}{\rightarrow} q$ and $\tilde{\mathcal{F}}(n_{m_k}) (\theta^*) \overset{p}{\rightarrow} \mathcal{J}(\theta^*)$. Let $r_{\mathcal{J}, n} = \delta_1 n^{m/2} / 2$. As in part a), given that $\hat{\theta}(n) \overset{p}{\rightarrow} \theta^*$, any subsequence of indices $n_{m_k}$ has a further sub-subsequence of indices $n_{m_{k^*}}$ where both $\hat{\theta}(n_{m_{k^*}}) \rightarrow \theta^*$, $Q_{n_{m_{k^*}}} \rightarrow q$, and $\tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) \rightarrow \mathcal{J}(\theta^*)$ almost surely. Then there is a $k_0$ such that if $k \geq k_0$ then $\|\hat{\theta}(n_{m_{k^*}}) - \theta^*\| \leq \delta_1 / 2$. Therefore if $k \geq k_0$ then $B(\hat{\theta}(n_{m_{k^*}}), r_{\mathcal{J}, n}/n_{m_{k^*}}^{m/2}) \subseteq B(\theta^*, \delta_1)$.

Thus, for $k \geq k_0$,
\[
\sup_{\theta \in B(\hat{\theta}(n_{m_{k^*}}), r_{\mathcal{J}, n}/n_{m_{k^*}}^{m/2})} \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta) - \mathcal{J}(\theta) \right\|
\leq \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) - \mathcal{J}(\theta^*) \right\|
+ \left( \|\hat{\theta}(n_{m_{k^*}}) - \theta^*\| + \delta_1 n_{m_{k^*}}^{-m/2} \right) \sup_{\theta \in B(\hat{\theta}(n_{m_{k^*}}), \delta_1 n_{m_{k^*}}^{-m/2})} \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta) - \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) \right\|
\leq \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) - \mathcal{J}(\theta^*) \right\|
+ \left( \|\hat{\theta}(n_{m_{k^*}}) - \theta^*\| + \delta_1 n_{m_{k^*}}^{-m/2} \right) \sup_{\theta \in B(\theta^*, \delta_1)} \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta) - \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) \right\|
\leq \left\| \tilde{\mathcal{F}}(n_{m_{k^*}})(\theta^*) - \mathcal{J}(\theta^*) \right\|
+ \left( \|\hat{\theta}(n_{m_{k^*}}) - \theta^*\| + \delta_1 n_{m_{k^*}}^{-m/2} \right) \frac{1}{n_{m_{k^*}}} \sum_{i \in n_{m_{k^*}}} \left\| \nabla^{\otimes 2}\ell(\theta; X_i) - \nabla^{\otimes 2}\ell(\theta^*; X_i) \right\|
\overset{a.s.}{\rightarrow} 0.
\]

Therefore, every subsequence of $S_n = \sup_{\theta \in B(\hat{\theta}(n), r_{\mathcal{J}, n}/n^{m})} \left\| \tilde{\mathcal{F}}(n)(\theta) - \mathcal{J}(\theta) \right\|$ has a further sub-subsequence converging almost surely to 0, and hence $S_n$ converges in probability to 0.
Proposition 14 (Sufficient conditions for Assumption 5). Each of the following imply Assumption 5.

a) there exists a $\delta > 0$ with $\sup_{q \in B_{S_2}(\theta_*)} \left\| \frac{1}{n} \sum_{i \in [n]} \nabla \ell(q; X) \right\|^{\otimes 2} \rightarrow 0$ and $I$ is continuous at $\theta_*$,
b) $\{\nabla \ell(x) | x \in \mathcal{X}\}$ is equicontinuous at $\theta_*$,
c) $\mathbb{E} \left[ \left\| \nabla^{\otimes 2} \ell(q; X) \right\|^2 \right] < \infty$.

Proof of Proposition 14.

a), b) The proofs are the same as for Proposition 13 a), b).

c) Let $Q_n = \frac{1}{n} \sum_{i \in [n]} \left| \nabla^{\otimes 2} \ell(q; X) \right|^2$, $q = \mathbb{E} \left[ \left| \nabla^{\otimes 2} \ell(q; X) \right|^2 \right]$, and let $r_{\theta_n} = n^m/2$. By the weak law of large numbers, $Q_n \xrightarrow{p} q$, and $\hat{T}^{(n)}(\theta_n) \xrightarrow{p} I(\theta_*)$. Starting with

$$\sup_{\theta \in B(\hat{\theta}_n, r_{\theta_n}/n^m)} \left\| \hat{T}^{(n)}(\theta) - I_\theta \right\| \leq \left\| \hat{T}^{(n)}(\theta_n) - I(\theta_*) \right\| + \sup_{\theta \in B(\hat{\theta}_n, r_{\theta_n}/n^m)} \left\| \hat{T}^{(n)}(\theta) - \hat{T}^{(n)}(\theta_*) \right\|,$$

we can bound the second term with a Taylor series and Cauchy-Shwarz as

$$\leq \frac{1}{n} \sum_{i \in [n]} \left\| \frac{1}{n} \sum_{i \in [n]} \nabla \ell(q; X) \right\|^2 \left\| \ell(q; X) \right\|^2 \left\| \nabla \ell(q; X) \right\|^2$$

$$\leq \frac{1}{n} \sum_{i \in [n]} \left\| \nabla \ell(q; X) \right\|^2 \left\| \nabla \ell(q; X) \right\|^2 \left\| \ell(q; X) \right\|^2 \left\| \nabla \ell(q; X) \right\|^2$$

$$\leq 2 \left\| q - q_n \right\| \sqrt{\frac{1}{n} \sum_{i \in [n]} \nabla \ell(q; X)}^2 \sqrt{\frac{1}{n} \sum_{i \in [n]} \ell(q; X)^2} \left\| \ell(q; X) \right\|^2 \left\| \nabla \ell(q; X) \right\|^2$$

Plugging this back in,

$$\sup_{\theta \in B(\hat{\theta}_n, r_{\theta_n}/n^m)} \left\| \hat{T}^{(n)}(\theta) - I_\theta \right\| \leq \left\| \hat{T}^{(n)}(\theta_n) - I(\theta_*) \right\| + \sup_{\theta \in B(\hat{\theta}_n, r_{\theta_n}/n^m)} \left\| \hat{T}^{(n)}(\theta) - \hat{T}^{(n)}(\theta_*) \right\|$$

$$\leq \left\| \hat{T}^{(n)}(\theta_n) - I(\theta_*) \right\| + 2 \left\| q - q_n \right\| \sqrt{\frac{1}{n} \sum_{i \in [n]} \nabla \ell(q; X)}^2 \sqrt{\frac{1}{n} \sum_{i \in [n]} \ell(q; X)^2} \left\| \ell(q; X) \right\|^2 \left\| \nabla \ell(q; X) \right\|^2$$

$$\xrightarrow{p} 0.$$

$\square$
E Proof of Proposition 4

Recall that
\[ d\vartheta_t = -\frac{1}{2} B\vartheta_t dt + \sqrt{A} dW_t, \]
which implies
\[ \vartheta_t = \exp(-B/2\vartheta_0) + \int_0^t \exp(-Bt/2)A^{1/2}dW_t. \]

Assuming stationarity, \( \vartheta_t \sim N(0, Q_\infty) \) where \( Q_\infty = \int_0^\infty \exp(-Bs/2)A \exp(-Bs/2)ds \), we have
\[
\text{Cov} \left( \int_0^t \vartheta ds \right) = \mathbb{E} \left( \int_0^t \int_0^s \vartheta \vartheta_T dsdr \right) = \int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr + \int_0^t \int_0^r \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr.
\]

We focus on the first term since the second term can be written similarly:
\[
\int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr = \int_0^t \int_0^s \mathbb{E} \left[ \left( \exp(-B(s-r)/2)\vartheta_r + \int_r^s \exp(-Bu/2)A^{1/2}dW_u \right) \vartheta_T^T \right] drds
\]
\[
= \int_0^t \int_0^s \exp(-B(s-r)/2) \mathbb{E}(\vartheta_r \vartheta_T^T) drds
\]
\[
= \int_0^t \int_0^s \exp(-B(s-r)/2)Q_\infty drds
\]
\[
= \int_0^t -2B^{-1}(\exp(-Bs/2) - 1)Q_\infty ds
\]
\[
= [4B^{-2}(\exp(-Bt/2) - 1) + 2tB^{-1}] Q_\infty.
\]

We can write \( \int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr \) similarly and combine the two results
\[
\text{Cov} (\vartheta_t) = \frac{1}{t^2} \text{Cov}(\int_0^t \vartheta ds) = \frac{1}{t^2} \left[ \int_0^t \int_0^s \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr + \int_0^t \int_0^r \mathbb{E}(\vartheta_s \vartheta_T^T) dsdr \right]
\]
\[
= \frac{4}{t} \text{Sym} (B^{-1}Q_\infty) - \frac{8}{t^2} \text{Sym} \left( B^{-2} \left\{ I - e^{-tB/2} \right\} Q_\infty \right),
\]
which completes the proof.

F Proof of Corollary 5

Proof. For Eq. (12), we have
\[
\mathcal{Q}_k^{(n)} = \text{Cov} \left( \tilde{\vartheta}_m/\langle b(n) \rangle \right) \approx \frac{1}{\langle w(n) \rangle^2} \text{Cov} \left( \tilde{\vartheta}_m/\langle b(n) \rangle \right)
\]
\[
= \frac{4}{m} \frac{\alpha(n)b(\alpha(n))}{n(\langle w(n) \rangle^2)} \text{Sym} \left( \left\{ c_h \Gamma J_s \right\}^{-1} Q_\infty \right)
\]
\[
- \frac{8}{m^2} \frac{(\alpha(n)b(\alpha(n)))^2}{\langle w(n) \rangle^2} \text{Sym} \left( \left\{ c_h \Gamma J_s \right\}^{-2} \left\{ I - \exp \left[ -\frac{c_{h,mn}}{2b(n)\alpha(n)} \Gamma J_s \right] Q_\infty \right\} \right).
\]

Now, given \( b \leq t \),
\[
\lim_{n \to \infty} n\mathcal{Q}_k^{(n)} = \frac{4ch}{m} \text{Sym} \left( \left\{ c_h \Gamma J_s \right\}^{-1} Q_\infty \right)
\]
\[
- \mathbb{I}_{[b+h=1]} \frac{8c_h^2}{m^2} \text{Sym} \left( \left\{ c_h \Gamma J_s \right\}^{-2} \left\{ I - e^{-\frac{c_{h,mn}}{2b(n)\alpha(n)} \Gamma J_s} \right\} Q_\infty \right).
\]

The rest follows by combining this with Proposition 4 and the simplifications following it, and by noting that since \( b + h \leq 1 \) and \( h > 0 \) we must have \( b < 1 \), and hence \( c_b = 1 \).
G 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. In SGLD-FP, the stochastic gradient \( \nabla \ell(\theta; X_t) \) is replaced by \( \nabla \ell(\theta; X_t) - \nabla \ell(\theta_*; X_t) \). By construction this stochastic gradient is still unbiased, but its significantly lower variance leads to materially different behaviour in the asymptotic analysis. Specifically, the corresponding \( [2, \ell \theta]^{(n_m)}(\theta) \) from the proof of Theorem 1 in Appendix B is vanishing under any scaling limit where the drift term \([1, \ell]\) does not vanish.

\[
\frac{n_m}{2} \mathbb{E}^{(n_m)}_{\theta} \left[ \frac{1}{2} \nabla^{\otimes 2} f(\theta) \Delta^{(n_m)}(\theta), \Delta^{(n_m)}(\theta) \right] \\
= n_m \mathbb{E}^{(n_m)}_{\theta} \left[ \frac{1}{2} \nabla^{\otimes 2} f(\theta) : \left( \Delta^{(n_m)}(\theta) \right)^{\otimes 2} \right] \\
= n_m \mathbb{E}^{(n_m)}_{\theta} \left[ \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\theta) : \left( \sum_{j \in [b^{(n)}]} \nabla \ell \left( \hat{\theta}^{(n)} + (u^{(n)})^{-1} \theta; X_{I_{1}^{(n)}(j)} \right) - \nabla \ell \left( \hat{\theta}^{(n)}; X_{I_{1}^{(n)}(j)} \right) \right)^{\otimes 2} \right] \\
\approx \frac{c_2}{c_5} h^{-2b + 2m - 2b} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\theta) \Gamma^T : \mathbb{E}^{(n_m)}_{\theta} \left( \sum_{j \in [b^{(n)}]} \nabla^{\otimes 2} \ell \left( \hat{\theta}^{(n)}; X_{I_{1}^{(n)}(j)} \right) (w^{(n)})^{-1} \right)^{\otimes 2} \\
\approx n_m \frac{a^{-2b - 2b} h^{-2b + 2m - 2b}}{2} \Gamma \nabla^{\otimes 2} f(\theta) \Gamma^T : \mathbb{E}^{(n_m)}_{\theta} \left( \sum_{j \in [b^{(n)}]} \nabla^{\otimes 2} \ell \left( \hat{\theta}^{(n)}; X_{I_{1}^{(n)}(j)} \right) \right)^{\otimes 2} \\
\approx n_m \frac{a^{-2b - 2b} h^{-2b + 2m - 2b}}{2} \Gamma \nabla^{\otimes 2} f(\theta) \Gamma^T : \left[ \hat{b}^{(n)} (b^{(n)} - 1) J_{\ell_2} \theta \theta^T J_{\ell_2} + \hat{b}^{(n)} K(\theta_*; \theta) \right]
\]

where \( K(\theta_*; \theta) = \int \nabla^{\otimes 2} \ell(\theta_*; x) \vartheta^{\otimes 2} \nabla^{\otimes 2} \ell(\theta_*; x) P(dx) \).

Now, we recall that for the drift term to be non-zero in the limit, we need \( a = h \). However, at any such scaling the \( [2, \ell \theta]^{(n_m)}(\theta) \) term is \( \mathcal{O}(n^{-b - 2b}) \), and so is always 0 in the limit.

H Sketch Proof for constrained parameter spaces

The key idea is that, if \( \theta_* \in \text{interior}(\Theta) \), there is a \( r > 0 \) with \( \theta_* \in B(\theta_*, r) \subset \text{interior}(\Theta) \), 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 \( (n_m) \), 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 \( \mathcal{P} \) (defined in Section 5.3) and an application of the Lebesgue dominated convergence theorem to handle truncation of the Gaussian increments shows that the \( A_{n_m} f \rightarrow Af \) uniformly within the extension of the support of \( f \) when \( \Theta = \mathbb{R}^d \). Moreover, the local property of the boundary condition (defined in Section 5.3) 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.
I Further discussion of asymptotics of mixing times

The discussion of the implications on the mixing time from Section 4.1 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 $\infty$ 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 [2021] introduces the $\zeta$-spectral gap, defined as

$$\text{SpecGap}_\zeta := \inf \left\{ \pi[f^2] - (f, Pf)_{L^2(\pi)} \mid \pi[f^2] - \zeta/2 \right\}$$

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 $\zeta$-spectral gap will converge to that of the OU-process for all $\zeta > 0$. This is supported by the intuitive interpretation of the $\zeta$-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.