2022-11-14

# Statistical inference with stochastic gradient algorithms

| | |
|---|---|
| Version | First author draft |
| Citation (published version): | J. Negrea, J. Yang, H. Feng, D. Roy, J. Huggins. 2022. "Statistical Inference with Stochastic Gradient Algorithms" https://doi.org/10.48550/arXiv.2207.12395 |

# Tuning Stochastic Gradient Algorithms for Statistical Inference via Large-Sample Asymptotics

Jeffrey Negrea[1,2], Jun Yang[3], Haoyue Feng[4], Daniel M. Roy[5,2], and Jonathan H. Huggins[4,6]

[1] *Department of Statistics and Actuarial Science, University of Waterloo, Canada*
[2] *Vector Institute, Canada*
[3] *Department of Mathematical Sciences, University of Copenhagen, Denmark*
[4] *Department of Mathematics & Statistics, Boston University, USA*
[5] *Department of Statistical Sciences, University of Toronto, Canada*
[6] *Faculty of Computing & Data Sciences, Boston University, USA*

July 21, 2023

**Abstract**

The tuning of stochastic gradient algorithms (SGAs) for optimization and sampling is often based on heuristics and trial-and-error rather than generalizable theory. We address this theory–practice gap by characterizing the large-sample statistical asymptotics of SGAs via a joint step-size–sample-size scaling limit. We show that iterate averaging with a large fixed step size is robust to the choice of tuning parameters and asymptotically has covariance proportional to that of the MLE sampling distribution. We also prove a Bernstein–von Mises-like theorem to guide tuning, including for generalized posteriors that are robust to model misspecification. Numerical experiments validate our results and recommendations in realistic finite-sample regimes. Our work lays the foundation for a systematic analysis of other stochastic gradient Markov chain Monte Carlo algorithms for a wide range of models.

## 1 Introduction

Stochastic gradient algorithms, which were originally proposed as optimization and root finding methods by Robbins and Monro [1951], have become the standard approach to large-scale optimization in statistics and machine learning. Their success can be attributed to the reduction in per-iteration computational complexity from subsampling outweighing the accuracy loss from stochastic approximation for empirical objectives. Hence, stochastic gradient algorithms scale more favourably with the sample size and model complexity than their deterministic counterparts [Moulines and Bach, 2011, Goodfellow et al., 2016]. Over the past decade, this scalability has also lead to tremendous growth in the use of stochastic gradient Markov chain Monte Carlo sampling algorithms, particularly in machine learning [Welling and Teh, 2011, Nemeth and Fearnhead, 2021].

Most analyses of stochastic gradient optimization procedures such as stochastic gradient descent (SGD) focus on the parameter error or the optimality gap [e.g., Moulines and Bach, 2011, Kushner and Yin, 2003, Nemirovski et al., 2009, Reddi et al., 2018], 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 [e.g., Teh et al., 2016, Vollmer et al., 2016, Brosse et al., 2018, Baker et al., 2019, Nemeth and Fearnhead, 2021, Raginsky et al., 2017, Durmus and Moulines, 2017, 2019]. These results often rely on settings for tuning parameters that fall outside of standard practice. It is an important challenge to explain why, empirically, stochastic gradient algorithms appear successful with previously unvalidated tunings (e.g., large step size and small batch size). The lack of an explanatory theory forced users to rely on heuristic and problem-specific approaches to 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 [Moulines and Bach, 2011, Dieuleveut et al., 2020] 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 [Teh et al., 2016, Vollmer et al., 2016].

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 converge to the sample paths of an Ornstein–Uhlenbeck process under relatively mild conditions. The class of algorithms includes stochastic gradient descent with and without additional Gaussian noise, momentum, and/or acceleration. 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 can hold after a small number of passes over the dataset. For sampling, we show that it is even possible to leverage stochastic gradients to sample modifications to the posterior that have better robustness to model misspecification. This result suggests that stochastic gradients have a potentially beneficial (or at least benign) role to play, rather than one that creates accuracy problems in exchange for computational efficiency.

Because the guarantees we provide are asymptotic in the sample size, it is possible that they may not be representative for a particular dataset. Therefore, we complement our asymptotic results with three numerical experiments to demonstrate that the limiting behaviour often predicts actual performance. These include a simulation study with a Gaussian location model, and two real-data experiments (a logistic regression example with 1 million observations, and a misspecified Poisson regression example with 150,000 observations).

## 1.1 Implications for sampling

Nemeth and Fearnhead [2021] recently identified several key areas for stochastic gradient MCMC (SG-MCMC) research. Our work makes significant strides in two of these areas for fixed–step-size variants of SGLD through our analysis of their large-sample asymptotics. One key area they identify is the need for general theoretical results beyond the log-concave regime that are not asymptotic in the number of iterations. We move beyond the log-concave regime by using large-sample asymptotics, analogous to the applicability of the Bernstein-von Mises theorem regardless of the convexity of the likelihood. Under tuning regimes relevant to statistical inference, our results apply after a constant number of epochs (i.e., passes over the full dataset). Another key area identified by Nemeth and Fearnhead [2021] is the need for methods for robust and/or adaptive tunings. Ideally, tunings ought to be automatable for non-experts to use. We use our results to make recommendations on the tuning of these methods in the large-sample setting (see Table 1), which is especially relevant in practice since stochastic gradient MCMC algorithms are typically used when the sample size is large. In particular, a large class of bad tunings whose large-sample asymptotics do not match the large-sample asymptotics of the target, or whose large-sample asymptotic local mixing is very slow, can be immediately identified and ruled out. Good tunings with the correct large-sample asymptotics and rapid asymptotic local mixing can also be identified as candidates for use and possibly fine-tuned using other methods [e.g., Coullon et al., 2023]. The guidance we derive in this way does not require additional expertise to use and could be implemented in an automated way. Moreover, our statistical perspective on the large-sample asymptotics of these methods leads to the insight that other benefits can be obtained by targeting statistically robust modifications of the posterior distribution, a direction not foreseen by Nemeth and Fearnhead.

We illustrate the implications of our results with two recent applications of SG-MCMC in the statistics literature.

**Example 1.** *Pollock et al. [2020] benchmark their subsampling-based MCMC algorithm against SGLD. They tuned SGLD using the best-available-at-the-time theoretical guidance [Teh et al., 2016] and other best practices, including variance-reduced stochastic gradients [Baker et al., 2019]. However, their implementation of SGLD mixes slowly and does not appear to be sampling from the posterior. This can be attributed to two causes. First, because they use a decreasing step size, each nearly independent sample takes an increasing number of epochs to reach. Second, even if they were to use a fixed–step-size—or to run the Langevin diffusion for the posterior directly—it would have mixed slowly due to ill-conditioning of the posterior distribution.*

*Based on the scaling limit we derive, the poor approximation quality due to ill-conditioning would have been foreseen, and a fixed–step-size sampler could have been appropriately tuned. Our theory predicts that, because SGLD was not preconditioned adequately, mixing would be very slow—and how much slower it is relative to the optimal preconditioner. Furthermore, since the optimal preconditioner according to our theory was used in the implementation of their method, the numerical comparison overstates the relative benefits of their proposed method versus SGLD. We demonstrate this on the same data as used by Pollock et al. [2020] in our Section 5.2. In short, this example exhibits how our results are directly exploitable: our tuning recommendations would have resolved the mixing problems of SGLD seen by Pollock et al. [2020], and led to a more meaningful comparator to their method.*

**Example 2.** *Nemeth and Fearnhead [2021, Section 6.3] compare various SG-MCMC algorithms on a challenging matrix factorization problem. Due to the lack of actionable tuning advice in the literature, they use the kernel Stein discrepancy (KSD) to select the step size. They initialise the variance-reduced SG-MCMC algorithms at the maximum a posteriori solution. Due to a pathology of the KSD [Coullon et al., 2023], this results in selection of the smallest possible step size of* $10^{-10}$*, which leads to the variance-reduced chains essentially remaining at their initialization. As a result, Nemeth and Fearnhead [2021, Figure 6] incorrectly suggests the that the variance-reduced algorithms had much lower predictive error than other SG-MCMC algorithms, when in fact it illustrates that the maximum a posteriori solution provides small test error—but of course no uncertainty quantification. Our theory predicts the observed poor approximation to the posterior and lack of meaningful uncertainty quantification from the "stuck" chains. Using our recommendations would have avoided the undetected pathological slow-mixing behaviour resulting from the use of the KSD.*

## 1.2 Implications for optimization and frequentist inference

Our theory provide rigorous foundations and new insights into the use of iterate averaging with fixed–step-size SGAs. Our main result differs from the seminal works on scaling limits in stochastic approximations [Kushner and Huang, 1981, Pflug, 1986, Walk, 1977, Kushner and Yang, 1993, Kushner and Yin, 2003] in both the nature of our analysis and the required regularity conditions. We analyze the setting where the source of the stochastic gradients is itself random and undergoes stochastic convergence. This is an important distinction because this joint limit is the pivotal object that we study in the present work and is required to address our research questions. In further contrast to our work, Kushner and Huang [1981], Pflug [1986], Kushner and Yang [1993], Kushner and Yin [2003] require restrictive assumptions that are not readily lifted to this "doubly stochastic convergence" case. The assumptions required by our analysis, on the other hand, 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 those achieved by previous analyses since we characterize both the sample paths of the iterates and the complete stationary distribution. For example, Walk [1977] only considers decreasing step sizes and demonstrates asymptotic normality of the marginal distribution, unlike our a functional/path limit results. Characterising the sample-path distribution is critical to analysing not just iterate averages but also the mixing time.

Despite not directly applying to the statistical setting with fixed step size, the work of Kushner and Huang [1981], Pflug [1986], Walk [1977], Kushner and Yang [1993], Kushner and Yin [2003] has been an important source of motivation for more recent methodological developments. We highlight two examples.

**Example 3.** *Li et al. [2018] propose a method for constructing samples from a local asymptotic fiducial distribution (one whose credible regions are asymptotic confidence sets at the same significance) by magnifying the deviations of SGD from the mode. They point out that the intuition underlying their results aligns with the Ornstein–Uhlenbeck scaling limit of stochastic gradient algorithms (including references to [e.g., Kushner and Huang, 1981, Pflug, 1986]). However, the rigorous proof of their results does not leverage the continuous-time limit, and uses stronger assumptions – such as weak strong convexity (equivalently, strong convexity of the composite objective) – than we require. The gap between the seminal work on Ornstein–Uhlenbeck limits of SGAs and the intuition they formed is, again, the need for a joint stochastic limit of the decreasing step size and stochastically varying objective function. Using our results, their heuristic explanation based on the* *continuous-time limit could be made rigorous and their findings and methods could be extended beyond weak strong convexity.*

**Example 4.** *A special case of the scaling limit we derive, and some of resulting the tuning recommendations, was conjectured by Mandt et al. [2017] based on the heuristic combination of statistical asymptotics of the posterior distribution (Bernstein–von Mises) and results from the stochastic approximation literature. However, being led mainly by heuristics, Mandt et al. [2017] arrive at an erroneous conclusion on the effect of iterate averaging that. In particular, they claim that "there exist conditions [on the data-generating process and sample size] under which iterate averaging generates one true posterior sample per pass over the data." However, their heuristic approach fails to account for "low order" terms, which our analyses in Section 4.2 reveal are* not *actually low-order when only making a single pass over the dataset. This difference is observed empirically in the numerical experiments we present in Section 5.1. Our results also go far beyond those conjectured by Mandt et al. [2017], broadening the applicability of this category of result, including fixed and growing batch-size regimes, non-traditional spatial scalings and concentration rates, and incorporating comparison to other asymptotic target distributions of interest.*

Another practical benefit of our rigour for the stochastic limit of stochastic processes is that we can clearly distinguish previous heuristics which can be turned into precise claims (the functional Bernstein–von Mises results, for instance), and those which require additional qualification or restricted application (such as mixing time results).

## 1.3 Notation

Let $\mathcal{M}_{1,+}(\mathcal{A})$ denote the set of probability measures on the measure space $\mathcal{A}$ 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 $\mathrm{N}_d(\mu, \Sigma)$. For vectors $a, b \in \mathbb{R}^d$, define the outer product $a \otimes b \in \mathbb{R}^{d\times d}$ given by $(a \otimes b)_{ij} = a_i b_j$ and write $a^{\otimes 2} := a \otimes a$. Let $\nabla \otimes \nabla = \nabla^{\otimes 2}$ denote the Hessian operator. For random elements $(\xi_k)_{k\in\mathbb{N}}$ and $\xi$, 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 $f$, $\mathbb{E}\{f(\xi_k)\} \to \mathbb{E}\{f(\xi)\}$ as $k \to \infty$. We write $\mathcal{L}(\xi)$ for the distribution (law) of a random element $\xi$, and $\mathcal{L}^{\nu}(\xi)$ for the conditional distribution of $\xi$ given another random element $\nu$. For a square matrix $M$, define the symmetrization operator as $\mathrm{Sym}(M) := (M + M^{\top})/2$. 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 $\mathbf{X}^{(n)} = (X_i)_{i=1}^n \in \mathcal{X}^n$ denote a dataset with observations $X_i$ independently and identically distributed (i.i.d.) from an unknown distribution $P$. For parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, consider the potential $\mathcal{U}^{(n)}(\theta) := r(\theta) + \sum_{i=1}^n \ell(\theta; X_i)$, 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 $\log \pi^{(0)}(\theta)$ that is everywhere positive on $\Theta$.

If $-\mathcal{U}^{(n)}(\theta)$ is interpreted as a (possibly regularized) loss, perhaps the most popular estimator for the (locally) optimal population parameter $\theta_{\star}$ satisfying $\mathbb{E}\{\nabla\ell(\theta_{\star}; X_1)\} = 0$, is the M-estimator $\widehat{\theta}^{(n)}$ satisfying the first-order optimality condition $\nabla\mathcal{U}^{(n)}(\widehat{\theta}^{(n)}) = 0$. If $-\mathcal{U}^{(n)}(\theta)$ is interpreted as 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 $\widehat{\theta}^{(n)}$ (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

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

where $(I_k^{(n)})_{k\in\mathbb{N}} \in ([n]^b)^{\mathbb{N}}$ are an independent and identically distributed (i.i.d.) sequence of uniform random samples from $\{1,\dots,n\}$ of size $b^{(n)}$, which are formed either with or without replacement.[1]

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

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

where $(h_k^{(n)})_{k\in\mathbb{N}}$ is a sequence of positive step sizes. While optimal tuning of the last-iterate error is 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 $h_k \propto k^{-\varsigma}$ for $\varsigma \in (0,1)$, the iterate average $\bar{\theta}_k^{(n)} := \frac{1}{k}\sum_{k'=1}^{k}\theta_{k'}^{(n)}$ satisfies

$$\lim_{n\to\infty}\lim_{k\to\infty} k\,\mathrm{Cov}(\bar{\theta}_k^{(n)}) = \mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} = \lim_{n\to\infty} n\,\mathrm{Cov}(\widehat{\theta}^{(n)}), \tag{3}$$

where $\mathcal{I}_\star := \mathbb{E}\left\{\nabla_\theta\ell(\theta_\star;X)\otimes\nabla_\theta\ell(\theta_\star;X)\right\}$ and $\mathcal{J}_\star := -\mathbb{E}\{\nabla_\theta^{\otimes 2}\ell(\theta_\star;X)\}$ 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 [Moulines and Bach, 2011, Toulis et al., 2021].

For sampling, the canonical approach is stochastic gradient Langevin dynamics [SGLD; Welling and Teh, 2011], with one-step update

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \tfrac{h_k^{(n)}}{2}\hat{G}_k^{(n)} + \sqrt{\tfrac{h_k^{(n)}}{\beta}}\ \xi_k, \tag{4}$$

where $\xi_k \sim \mathrm{N}_d(0,I)$ is independent standard Gaussian noise and $\beta \in (0,\infty]$ is the inverse temperature, which is usually taken to be $n$.[2] The benefits of introducing stochastic gradients into an MCMC procedure are less clear than for optimization since retaining exactness would require an accept/reject step using the full-sample likelihood in the Metropolis–Hastings adjustment. While SGLD can be asymptotically exact when run with a decreasing step size, 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]. Despite these limitations, variants of SGLD has been an active area of methods development and seen adoption in practice [Ahn et al., 2012a, Chen et al., 2014, Ma et al., 2015, Baker et al., 2019, Nemeth and Fearnhead, 2021].

# 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

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h^{(n)}\Gamma}{2}\hat{G}_k^{(n)} + \sqrt{\frac{h^{(n)}\Lambda}{\beta^{(n)}}}\ \xi_k, \tag{5}$$

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. $\mathrm{N}_d(0,I_d)$, and $\hat{G}_k^{(n)}$ implicitly depends on the batch size $b^{(n)}$ (which in turn may vary with the sample size $n$). 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; see Appendix A.1 for details in the case of the underdamped stochastic Langevin dynamics.

[1] "With replacement" means $(I_k^{(n)})_{k\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathrm{Unif}([n]^b)$, and "without replacement" means $(I_k^{(n)})_{k\in\mathbb{N}} \overset{\text{iid}}{\sim} \mathrm{Unif}(\{I \in [n]^b : (j_1 \neq j_2 \Rightarrow I(j_1) \neq I(j_2))\})$.

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

### 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 $\widehat{\theta}^{(n)}$, 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

$$\vartheta_t^{(n)} := w^{(n)} \left( \theta^{(n)}_{\lfloor \alpha^{(n)} t \rfloor} - \widehat{\theta}^{(n)} \right), \tag{6}$$

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: $h^{(n)} = c_h n^{-\mathfrak{h}}$ for $\mathfrak{h} > 0$, $b^{(n)} = \lfloor c_b n^{\mathfrak{b}} \rfloor$ for $\mathfrak{b} \geq 0$, and $\beta^{(n)} = c_\beta n^{\mathfrak{t}}$ for $\mathfrak{t} \in \mathbb{R}$. Given these tuning parameters, in order to have a stable and non-trivial[3] limit, we must take the time scaling to be $\alpha^{(n)} = n^{\mathfrak{h}}$ and the spatial scaling to be $w^{(n)} = n^{\mathfrak{w}}$ for $\mathfrak{w} = \min\{\mathfrak{b} + \mathfrak{h}, \mathfrak{t}\}/2$. In this setting we have the following result, under Assumptions 1 to 5 discussed in Section 3.2. All proofs are deferred to the Supplementary Materials.

**Theorem 1** (Scaling limit of the meta-algorithm)**.** *If Assumptions 1 to 5 hold, there exists $\theta_\star \in \Theta$ such that $\widehat{\theta}^{(n)} \xrightarrow{p} \theta_\star$, and there exists $\vartheta_0 \in \mathcal{M}_{1,+}(\Theta)$ such that $\vartheta_0^{(n)} \rightsquigarrow \vartheta_0$, then $(\vartheta_t^{(n)})_{t \in \mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t \in \mathbb{R}_+}$ in the Skorohod topology[4] in probability, where $(\vartheta_t)_{t \in \mathbb{R}}$ is an Ornstein–Uhlenbeck process given by*

$$\mathrm{d}\vartheta_t = -\frac{1}{2} B \vartheta_t \, \mathrm{d}t + \sqrt{A}\, \mathrm{d}W_t, \tag{7}$$

*with $W_t$ a d-dimensional standard Brownian motion, $B := c_h \Gamma \mathcal{J}_\star$ the drift matrix, $A := \mathbb{I}_{[\mathfrak{b}+\mathfrak{h} \leq \mathfrak{t}]} \frac{c_h^2 \overline{c_b}}{4 c_b} \Gamma \mathcal{I}_\star \Gamma^\top + \mathbb{I}_{[\mathfrak{t} \leq \mathfrak{b}+\mathfrak{h}]} \frac{c_h}{c_\beta} \Lambda$ the positive semi-definite diffusion matrix, and $\overline{c_b} := 1 - c_b \mathbb{I}_{[\mathfrak{b}=1 \text{ and ``no replacement''}]}$ the batch constant.*

**Remark 1** (Assumptions)**.** *Assumptions 1 to 5 are quite weak and notably do not require convexity or bounded gradients. We require that the sequence of empirical critical points of the log-likelihood converges to a critical point of the expected log-likelihood. The critical point does not even need to be a minimizer, though in the case of a limiting critical point where the hessian is not positive definite, the paths of the process will move away from the critical point instead of towards it (hence the need for the Hurwitz condition for existence of the stationary distribution below). Further, we do not require a specific rate of convergence for the empirical critical point to the limiting one, but there is a trade-off determined by our proof strategy between this rate and the number of moments we must assume exist for various derivatives of the likelihood.*

**Remark 2** (Effects of stochastic gradient noise)**.** *As expected, the mini-batch noise contributes in the large-sample regime when $\mathfrak{h} + \mathfrak{b} \leq \mathfrak{t}$. This exactly corresponds to when the mini-batch noise in a single step is on the same order (=) or dominates (<) the noise from the Gaussian innovations, $\xi_k$. We can interpret the phase transition as occurring because the variance of the mini-batch gradient scales as $n^{-2\mathfrak{h}-\mathfrak{b}}$ while the variance of update due to the Gaussian innovations scale as $n^{-\mathfrak{h}-\mathfrak{t}}$. The spatial scaling is chosen as $\mathfrak{w} = \min\{\mathfrak{b} + \mathfrak{h}, \mathfrak{t}\}/2$ to ensure that at least one of (a) the mini-batch noise or (b) the Gaussian innovations contribute to the limit, as otherwise the limit would be a gradient flow instead of Ornstein–Uhlenbeck process, and hence fail to capture the asymptotically dominant local stochastic behaviour around $\widehat{\theta}^{(n)}$.*

**Remark 3** (SGLD with control variates)**.** *Modifications to SGLD that include control variates can be analyzed using similar techniques. These methods seek to reduce the variance of stochastic gradients using a control variate. In Appendices A.2 and H we examine the SGLD-FP algorithm [Baker et al., 2019, Nagapetyan et al., 2017], where the control variate is given by the random gradient function evaluated at (the current estimate of) the MLE. Formally, in an idealized setting where the MLE is known, it modifies the meta-algorithm by replacing Eq. (1) with*

$$\hat{G}_k^{(n)} := \frac{1}{n} \nabla r\left(\theta_k^{(n)}\right) + \frac{1}{b^{(n)}} \sum_{j=1}^{b^{(n)}} \left\{ \nabla \ell\left(\theta_k^{(n)};\ X_{I_k^{(n)}(j)}\right) - \nabla \ell\left(\widehat{\theta}^{(n)};\ X_{I_k^{(n)}(j)}\right) \right\}. \tag{8}$$

[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 B.3 for further discussion.

*We find that, when a non-trivial amount of additional Gaussian noise is included ($\beta > 0$), the use of control variates is sufficient to reduce the variance in minibatch gradients so much that at all non-trivial scalings the gradient noise is* $0$. *Hence, Theorem 1 holds for SGLD-FP except with* $\mathfrak{w} = \mathfrak{t}/2$, *and* $A = \frac{c_h}{c_\beta}\Lambda$.

Based on Theorem 1, we can establish the following corollaries which we will further leverage to explain the empirical behaviour of stochastic gradient methods and to make recommendations for how these methods could be best tuned. First, we have a characterization of the marginal and (when it exists) the stationary covariance of the limiting process, including conditions under which simplified forms are possible. A square matrix $M$ is said to be *Hurwitz* (or *stable*) if every eigenvalue of $M$ has negative real part.

**Corollary 1** (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 distribution is* $\mathcal{L}^{\vartheta_0}(\vartheta_t) = \mathrm{N}_d\big(e^{-sB/2}\vartheta_0, Q_t\big)$, *with*

$$Q_t := \mathrm{Cov}(\vartheta_t|\vartheta_0) = \int_0^t e^{-sB/2} A e^{-sB^\top/2} \mathrm{d}s.$$

2. *If* $-\Gamma\mathcal{J}_\star$ *is Hurwitz, then* $Q_\infty := \lim_{t\to\infty} Q_t$ *exists and the stationary distribution of* $(\vartheta_t)_{t\in\mathbb{R}}$ *is* $\nu := \mathrm{N}_d(0, Q_\infty)$. *In this case,* $Q_\infty$ *solves the equation*

$$\frac{1}{2}BQ_\infty + \frac{1}{2}Q_\infty B^\top = A. \tag{9}$$

Let $\nu^{(n)}$ denote the stationary measure of the stochastic gradient algorithm when the sample size is $n$, if it exists. The previous corollary leads to conditions for a Bernstein–von Mises-type result for these stationary measures.

**Corollary 2** (Bernstein–von Mises-type theorem). *In the setting of Theorem 1, if* $-\Gamma\mathcal{J}_\star$ *is Hurwitz and the collection* $\{\nu^{(n)}\}_{n\in\mathbb{N}}$ *is uniformly tight, the stationary-distributed parameters* $\theta^{(n)} \sim \nu^{(n)}$ *satisfy* $n^{\mathfrak{w}}(\theta^{(n)} - \widehat{\theta}^{(n)}) \rightsquigarrow \mathrm{N}_d(0, Q_\infty)$ *in probability.*

We can interpret Corollary 2 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 in the optimization theory sense of having Lipschitz gradients.

**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.** $\mathfrak{h} - \mathfrak{w} - \mathfrak{a}/3 > 0$ *and* $\mathbb{E}\left[\|\nabla\ell(\theta_\star; X_1)\|^{p_2}\right] < \infty$ *for some* $p_2 > \frac{1}{\mathfrak{h} - \mathfrak{w} - \mathfrak{a}/3}$.

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

**Assumption 3.** *For some* $q_3 \in [0, \mathfrak{w})$ *and* $p_3 := \frac{1}{\mathfrak{h} + q_3 - \mathfrak{w} - \mathfrak{a}/3}$, $\|\widehat{\theta}^{(n)} - \theta_\star\| \in o_p(1/n^{q_3})$, *and* $\mathbb{E}\big[\big\|\nabla^{\otimes 2}\ell(\cdot; X_1)\big\|_\infty^{p_3}\big] < \infty$.

Assumptions 4 and 5 require convergence of the first-and second-order empirical Fisher information matrices $\widehat{\mathcal{I}}^{(n)}(\theta) = \frac{1}{n}\sum_{i\in[n]}[\nabla\ell(\theta; X_i)]^{\otimes 2}$ and $\widehat{\mathcal{J}}^{(n)}(\theta) = \frac{1}{n}\sum_{i\in[n]}[-\nabla^{\otimes 2}\ell(\theta; X_i)]$. For any $r > 0$, define the ball $B^{(n)}(r) := \{\theta \in \Theta : \left\|\theta - \widehat{\theta}^{(n)}\right\| \le r/n^{\mathfrak{w}}\}$.

**Assumption 4.** *There is a non-decreasing sequence* $r_{\mathcal{J},n} \overset{n\to\infty}{\longrightarrow} \infty$ *such that* $\sup_{\theta \in B^{(n)}(r_{\mathcal{J},n})} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\| \overset{p}{\to} 0$.

**Assumption 5.** *There is a non-decreasing sequence* $r_{\mathcal{I},n} \overset{n\to\infty}{\longrightarrow} \infty$ *such that* $\sup_{\theta \in B^{(n)}\left(r_{\mathcal{I},n}\right)} \left\| \widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}(\theta_\star) \right\| \overset{p}{\to} 0$.

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 E.

# 4 Practical implications of the scaling limit

We now turn to assessing the implications of our statistical scaling limit on the large-sample behavior of stochastic gradient algorithms used for optimization and sampling.

## 4.1 Mixing time

Because we characterize the full-path behavior of the meta-algorithm, we can obtain insights into its mixing speed. Let $\hat{\nu}_k^{(n)}(f) := k^{-1} \sum_{k'=1}^{k} f(\theta_{k'}^{(n)})$ denote the Monte Carlo estimate of $\nu^{(n)}(f)$. We can use the *mixing time* (or worst-case integrated autocorrelation time) $\tau^{(n)} := \sup_f \inf\{k : \operatorname{Var}_{\hat{\nu}_k^{(n)}}(f) / \operatorname{Var}_{\nu^{(n)}}(f) \le 1\}$ to characterize the efficiency of MCMC algorithms. For the limiting process, define the "Monte Carlo average" $\hat{\nu}_t(f) := t^{-1} \int_0^t f(\vartheta_s)\,\mathrm{d}s$ and the mixing time $\tau := \sup_f \inf\{t : \operatorname{Var}_{\hat{\nu}_t}(f) / \operatorname{Var}_{\nu}(f) \le 1\}$. When the limiting process is reversible, standard results[5] 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 $2\alpha^{(n)} b^{(n)}/\lambda_{\min}(B) = 2b^{(n)}/\{h^{(n)} \lambda_{\min}(\Gamma \mathcal{J}_\star)\}$ likelihood evaluations, or equivalently $2b^{(n)}/\{h^{(n)} \lambda_{\min}(\Gamma \mathcal{J}_\star)\}$ epochs. 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.

The reason these arguments are heuristic is because weak converge of the processes and stationary distributions is insufficient to conclude that the mixing times converge. In Appendix J, we provide further details and describe a possible approach to making the mixing result rigorous. We note, however, that comparison between the mixing time of a scaling limit and the mixing time of the corresponding pre-limiting processes is standard is MCMC tuning, even though it is technically only a heuristic. This is, for example, the nature of widely celebrated results in the optimal scaling literature [e.g., Gelman et al., 1997, Roberts and Rosenthal, 2001]. Thus, as a practical matter, a user with a dataset of size $n$ can conclude that using a step size $h$ and batch size $b$, will result in a mixing time of roughly

$$\frac{2b}{h\,\lambda_{\min}\{\Gamma \widehat{\mathcal{J}}^{(n)}(\widehat{\theta}^{(n)})\}} \tag{10}$$

epochs, thereby providing a valuable constraint when tuning $b$, $h$, and $\Gamma$. Some example tuning parameter combinations that lead to limiting stationary distributions of interest, and the corresponding mixing times of the limit process, are given in Table 1.

## 4.2 Optimization

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

[5] Apply the spectral theorem for self-adjoint operators [Rudin, 1991] to the Poincaré inequality [Bakry et al., 2014]

**Proposition 1** (Path averaging)**.** *For $(\vartheta_t)_{t\in\mathbb{R}_+}$ defined by Eq. (7), assuming $-B$ is Hurwitz and $\vartheta_0 \sim \mathrm{N}(0, Q_\infty)$, the covariance of the averaged limiting process is*

$$\bar{Q}_t := \mathrm{Cov}\left(\bar{\vartheta}_t\right) = \frac{4}{t}B^{-1}AB^{-\top} - \frac{8}{t^2}\,\mathrm{Sym}\left(B^{-2}\left\{I - e^{-tB/2}\right\}Q_\infty\right) \tag{11}$$

$$= \begin{cases} Q_\infty - \frac{t}{6}A + O(t^2) & \textit{if } t \ll 7\left\|B\right\|^2\left\|B^{-2}Q_\infty\right\|^{1/2} \\ \frac{4}{t}B^{-1}AB^{-\top} + O(t^{-2}) & \textit{if } t \gg 3\left\|B^{-2}Q_\infty\right\|^{1/2}. \end{cases} \tag{12}$$

*If either (i) $\mathfrak{b} + \mathfrak{h} < \mathfrak{t}$ or (ii) $\mathfrak{b} + \mathfrak{h} = \mathfrak{t}$ and $c_\beta = +\infty$, then $\frac{4}{t}B^{-1}AB^{-\top} = \frac{\overline{c_b}}{tc_b}\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}$.*

The proof of this result is in Appendix F. Using Proposition 1, we can characterize large-sample behaviour of $\bar{\theta}_k^{(n)}$ for $k = k^{(n)} := \lfloor m\alpha^{(n)}/c_b\rfloor = \lfloor mn^{\mathfrak{h}}/c_b\rfloor$, which corresponds to making $m$ passes over the dataset.

**Corollary 3** (Bernstein–von Mises-type theorem for iterate averaging)**.** *Suppose Assumptions 1 to 5 all hold. If $\mathfrak{b} + \mathfrak{h} \le \mathfrak{t} \le 1$ and $\mathcal{L}(\vartheta_0^{(n)}) \rightsquigarrow \mathrm{N}(0, Q_\infty)$, then $n^{\mathfrak{b}+\mathfrak{h}}(\bar{\theta}_{k^{(n)}}^{(n)} - \widehat{\theta}^{(n)})$ converges in distribution to a zero-mean Gaussian and*

$$n^{\mathfrak{b}+\mathfrak{h}}\,\mathrm{Cov}\left(\bar{\theta}_{k^{(n)}}^{(n)}\right) \to \frac{4c_b}{c_h m}\,\mathrm{Sym}\left((\Gamma\mathcal{J}_\star)^{-1}Q_\infty\right) - \frac{8c_b^2}{c_h^2 m^2}\,\mathrm{Sym}\left((\Gamma\mathcal{J}_\star)^{-2}\left\{I - e^{-m\Gamma\mathcal{J}_\star/(2c_b)}\right\}Q_\infty\right) \tag{13}$$

*in probability. If in addition $\mathfrak{b} + \mathfrak{h} = 1$ and $c_\beta = +\infty$, then*

$$n\,\mathrm{Cov}\left(\bar{\theta}_{k^{(n)}}^{(n)}\right) \to \frac{1}{m}\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} + R(m) \textit{ in probability,} \qquad \textit{where } \|R(m)\| \le \frac{8c_b^2}{c_h^2 m^2}\left\|(\Gamma\mathcal{J}_\star)^{-2}Q_\infty\right\|. \tag{14}$$

It follows from Eqs. (3) and (14) that for $m$ sufficiently large, iterate averaging with potentially large step size of order $n^{-\mathfrak{h}}$ ("large" constant-in-time step sizes here means that $\mathfrak{h} \ll 1$, while the results apply for any $\mathfrak{h} \le 1$) and batch size of order $n^{1-\mathfrak{h}}$ has numerical error $\mathrm{Cov}(\bar{\theta}_{k^{(n)}}^{(n)}) \approx \frac{1}{m}\mathrm{Cov}(\widehat{\theta}^{(n)})$ for estimation of $\widehat{\theta}^{(n)}$. This error is optimal, in the sense that after $m \gg 1$ passes over the dataset, it is small compared to the statistical error $\mathrm{Cov}(\widehat{\theta}^{(n)})$. It is instructive to consider two idealized cases:

1. Take $\Gamma = \Lambda = I$ and assume that $\mathcal{J}_\star$ and $\mathcal{I}_\star$ commute. The bound on the remainder term simplifies to $\|R(m)\| \le \frac{2c_b}{c_h m^2}\left\|\mathcal{J}_\star^{-1}\right\|$. Hence, we should only expect the remainder term to be small when $m^2 \gg \frac{2c_b}{c_h}\left\|\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-2}\right\|$.
2. Take $\Gamma = \Lambda = \mathcal{J}_\star^{-1}$. The bound on the remainder term simplifies to $\|R(m)\| \le \frac{2c_b}{c_h m^2}\left\|\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}\right\|$. Hence, we should only expect the remainder term to be small when $m^2 \gg \frac{2c_b}{c_h}$.

In either case, a large step size constant $c_h$ relative to the batch size constant $c_b$ leads to the remainder term being small even for small $m$. However, particularly without preconditioning, this regime may lead to numerical instability.

## 4.3 Sampling

**Sampling from the posterior.** The Bernstein-von Mises theorem states that the posterior-distributed parameter $\theta^{(n)} \sim \pi^{(n)}$ satisfies $n^{1/2}(\theta^{(n)} - \widehat{\theta}^{(n)}) \rightsquigarrow \mathrm{N}_d(0, \mathcal{J}_\star^{-1})$ 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 $\mathfrak{w} = 1/2$. Then, there are several ways to ensure that the limiting process has the same distribution as the limiting, six of which using various forms of SGD, SGLD, and SGLD-FP are shown in Table 1.

In terms of the number of gradient queries per unit mixing time the cases where $h \notin o(b/n)$ are the most efficient as the query-count scales linearly with the dataset size (since $\mathfrak{h} + \mathfrak{b} = 1$), while for the cases where $h \in o(b/n)$ it scales super-linearly ($\mathfrak{h} + \mathfrak{b} > 1$). In practice, options involving preconditioning matrices ($\Gamma \neq I$) or control variates (SGLD-FP) first require an estimate of $\widehat{\theta}^{(n)}$ to, respectively, construct estimates of the preconditioner $\Gamma = \widehat{\mathcal{J}}^{(n)}(\widehat{\theta}^{(n)})^{-1} \approx \mathcal{J}_\star^{-1}$ and/or construct the control variates $\nabla\ell(\widehat{\theta}^{(n)}; X_i)$. The latter

| Target | Target Asymp. Cov. | Algo. | $\Gamma$ | $\Lambda$ | $\beta$ | $h$ | Mix. Time (Epochs) |
|---|---|---|---|---|---|---|---|
| Posterior | $\mathcal{J}_\star^{-1}$ | SGD | $\mathcal{I}_\star^{-1}$ | n.a. | n.a. | $4b/n$ | $\lambda_{\min}^{-1}(\mathcal{I}_\star^{-1}\mathcal{J}_\star)$ |
| Posterior | $\mathcal{J}_\star^{-1}$ | SGLD | $\mathcal{J}_\star^{-1}$ | $\mathcal{J}_\star^{-1}$ | $c_\beta n$ | $\frac{4b(1-c_\beta)}{nc_\beta}$ | $\frac{c_\beta}{(1-c_\beta)}$ |
| Posterior | $\mathcal{J}_\star^{-1}$ | SGLD | $\mathcal{J}_\star^{-1}$ | $\mathcal{J}_\star^{-1}$ | $n$ | $o(b/n)$ | $\frac{b}{nh} \in \omega(1)$ |
| Posterior | $\mathcal{J}_\star^{-1}$ | SGLD | $I$ | $I$ | $n$ | $o(b/n)$ | $\frac{b\lambda_{\min}^{-1}(\mathcal{J}_\star)}{nh} \in \omega(1)$ |
| Posterior | $\mathcal{J}_\star^{-1}$ | SGLD-FP | $I$ | $I$ | $n$ | $4b/n$ | $\lambda_{\min}^{-1}(\mathcal{J}_\star)$ |
| Posterior | $\mathcal{J}_\star^{-1}$ | SGLD-FP | $\mathcal{J}_\star^{-1}$ | $\mathcal{J}_\star^{-1}$ | $n$ | $4b/n$ | 1 |
| Bagged Posterior | $w_1\mathcal{J}_\star^{-1} + w_2\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}$ | SGLD | $\mathcal{J}_\star^{-1}$ | $\mathcal{J}_\star^{-1}$ | $\frac{n}{w_2}$ | $\frac{4w_1b}{n}$ | $1/w_1$ |
| Local Fiducial | $\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}$ | SGD | $\mathcal{J}_\star^{-1}$ | n.a. | n.a. | $\frac{4b}{n}$ | 1 |

Table 1: Tuning parameter combinations for various target distributions, and their corresponding mixing times in epochs. If the mixing time is $\omega(1)$ (in $n$), then in the limit the process does not mix in a constant number of epochs.

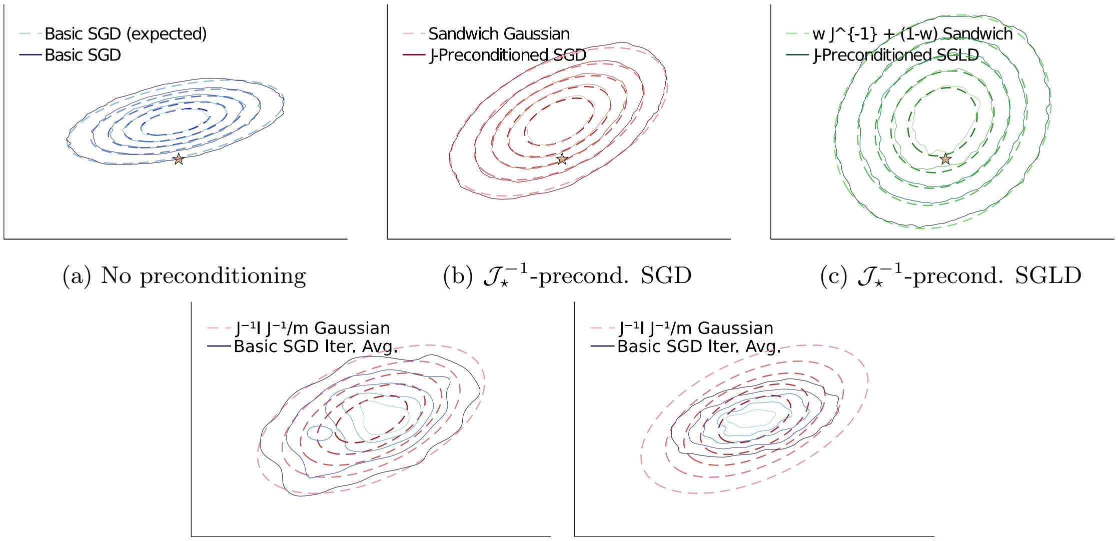
option is more appealing, particularly if $d$ is large, as no matrix inversion or per-iteration multiplication is required. In either case, however, preconditioning with $\widehat{\mathcal{J}}^{(n)}(\widehat{\theta}^{(n)})^{-1}$ will minimize the mixing time. Methods for estimating $\mathcal{J}_\star$, $\mathcal{I}_\star$, their inverses, and sparse approximations have been explored extensively in other work [e.g., Haario et al., 2001, Ahn et al., 2012b, Mandt et al., 2017, Pollock et al., 2020, Chen et al., 2020].

**Alternative uncertainty quantification.** When the model is misspecified or when generalized Bayesian inference based on a loss function is used [Bissiri et al., 2016], the (generalized) posterior distribution may provide less-than-robust uncertainty quantification because the (local) M-estimator $\widehat{\theta}^{(n)}$ is itself asymptotically normal, centered at the true parameter $\theta_\star$, with covariance equal to the "sandwich" covariance matrix, $\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1}$ [Kleijn and Van der Vaart, 2012, Müller, 2013]. If the model is well-specified (i.e., $P = Q_\theta$ for some $\theta \in \Theta$), then $\mathcal{J}_\star = \mathcal{I}_\star$, and so $\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} = \mathcal{J}_\star^{-1}$. However, if the model is misspecified (i.e., $P \neq Q_\theta$ for any $\theta \in \Theta$), then the sandwich may differ from $\mathcal{J}_\star^{-1}$ [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] and predictions (or other decision-theoretic quantities) can become unstable [Huggins and Miller, 2019, 2023].

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 $\Gamma = \Lambda = \mathcal{J}_\star^{-1}$, and any valid $\mathfrak{h}+\mathfrak{b} = 1 = \mathfrak{t}$. With this tuning, for any $w_1, w_2 > 0$, taking $c_h = 4w_1c_\mathfrak{b}$ and $c_\beta = w_2^{-1}$, gives $Q_\infty = w_1\mathcal{J}_\star^{-1}\mathcal{I}_\star\mathcal{J}_\star^{-1} + w_2\mathcal{J}_\star^{-1}$. This matches the asymptotic distribution of the bagged posterior with re-sampling rate $w_1$ when $w_1 = w_2$ [Huggins and Miller, 2019]. This is summarized in the "Bagged Posterior" row of Table 1. 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 $1/w_1$ epochs. Hence in principle we can use SGD (by setting $w_1 = 1$, $w_2 = 0$, and $c_\beta = +\infty$) to obtain the sandwich covariance and minimize the mixing time to be a single pass over the dataset. This can be interpreted as an asymptotic local fiducial distribution for the parameter, as it has credible sets which match frequentist confidence sets asymptotically. This tuning is summarized in the "Local Fiducial" row of Table 1.

## 5 Numerical Experiments

We present results for three experiments using both simulated and real data that show our theory closely reflects finite-sample behavior. Source code for experiments is at `https://github.com/jnegrea/stat-infr-sgas`.



(a) No preconditioning

(b) $\mathcal{J}_\star^{-1}$-precond. SGD

(c) $\mathcal{J}_\star^{-1}$-precond. SGLD

(d) SGD Iter. Avg., 8 epochs

(e) SGD Iter. Avg., 1 epoch

Figure 1: Results of experiment 1. The empirical results follow the theoretical predictions based on scaling limits. For preconditioned SGLD, $w = 1/2$.

## 5.1 Experiment 1: Gaussian simulation study

First we demonstrate the effect of model misspecification on tuning, highlighting both the sampling implications (Section 4.3 and Table 1) and the iterate averaging behaviour relevant to optimization (Section 4.2). We choose the combination of the data-generating distribution and likelihood function specifically to ensure that $\mathcal{J}_\star \neq \mathcal{I}_\star$, so that the effect of misspecification would be apparent. We run SGD with no preconditioning and with preconditioning by $\mathcal{J}_\star$, and SGLD with preconditioning by $\mathcal{J}_\star$. For SGLD we use the inverse temperature $\beta^{(n)} = n$, which corresponds to the canonical choice that would be made when not using stochastic gradients. We also compute iterate averages for SGD with no preconditioning over 1 epoch and 8 epochs. Exact specifications for the experiment are in Table 3 in the supplemental material.

We interpret our results using our scaling limit with parameters $\mathfrak{w} = 1/2$, $\mathfrak{h} = 1$, $\mathfrak{b} = 0$, which corresponds to the standard statistical local scaling and a fixed batch size. Figure 1 shows plots for the joint density of the first and last coordinates of the parameter vector for each of the five tunings. The density for the empirical run of the algorithms is given by a 2D kernel density estimate. The density for the predicted behaviour is given by the stationary distribution of the limiting process. As predicted by our results in Section 4.3, specifically in the "Local Fiducial" row of Table 1, preconditioning by $\mathcal{J}_\star$ leads to an empirical distribution for the iterates of the algorithm matching the covariance of the MLE (an asymptotic locally fiducial distribution), not preconditioning leads to behaviour that matches neither (but is still predictable using our results), and preconditioning by $\mathcal{J}_\star$ for SGLD leads to an empirical distribution for the iterates of the algorithm matching the asymptotics of a bagged posterior, which is given by a linear combination of the covariance of the MLE and the covariance of the posterior. Furthermore, as predicted by the results in Section 4.2 (in particular Eq. (14)), and in contrast to the predictions in Mandt et al. [2017], iterate averaging for a "large" number of epochs (8) is closely approximated by the scaled sandwich covariance, while iterate averaging over a "small" number of epochs (1) is not sufficient for the approximation by the scaled sandwich covariance to be accurate. Finally, Table 2 shows that the mixing times predicted by our theory closely match their empirical counterparts.

## 5.2 Experiment 2: Large-scale inference for airline delay data—logistic regression

Next we examine the same airline dataset and model as in Pollock et al. [2020] using their pre-processed data so we can directly compare our recommended settings to the results they obtained with suboptimal tuning parameters (see Example 1). The responses are binary and there are 3 covariates. We use the full dataset ($\approx$ 120 million observations) to estimate the "ground truth" quantities $(\theta_\star, \mathcal{J}_\star, \mathcal{I}_\star)$, and we apply the stochastic gradient algorithms using  a random subsample of size 1 million from the full dataset.

For the results regarding the marginal distribution of the iterates, 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 yield essentially identical results, and all are materially different from not preconditioning. Again, we interpret this using our scaling limit with parameters $\mathfrak{w} = 1/2$, $\mathfrak{h} = 1$, $\mathfrak{b} = 0$. An experimental finding of Pollock et al. [2020] was that (non-preconditioned) SGLD had relatively poor mixing performance as compared with the ScaLE algorithm they introduce. Figures 2 and 4 and Table 2 similarly show that, without preconditioning, SGLD fails to properly quantify uncertainty in the true parameter (marginally for coordinate 4, and jointly) and mixes slowly. Furthermore, SGLD without preconditioning mixes materially more slowly than preconditioned methods, as evidenced by the jagged histogram from its run (Fig. 2) and the contour plot (Fig. 4). Our numerical results also show that their findings would have been significantly different had they used the appropriate preconditioning as predicted by our results showing that preconditioning accelerates the mixing of SGLD considerably and leads to more accurate uncertainty quantification. These findings are consistent with our theoretical developments in Sections 4.1 and 4.3 (in particular, Eq. (10), and the preconditioned SGD and non-preconditioned SGLD and SGLD-FP rows of Table 1). Thus, we can conclude that the poor relative mixing of non-preconditioned SGLD-FP observed in Pollock et al. [2020], and the fast mixing with preconditioning could both have been predicted using our results. In particular, fixed–step-size preconditioned SGD would have been much more competitive with that work's proposed method than the non-preconditioned decreasing step size SGLD that was used.

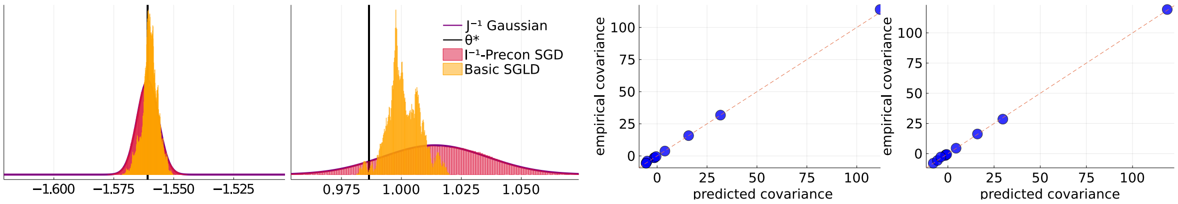
To further explore the value of our tuning guidance, we also consider the behavior of iterate averaging when using the preconditioner $\mathrm{diag}(\mathcal{J}_\star)^{-1}$, which is less computationally demanding in high dimensions, and examine different combinations of the step size and batch size scaling powers that both lead to statistically relevant scaling limits (in particular $(\mathfrak{h}, \mathfrak{b}) \in \{(1, 0), (1/2, 1/2)\}$). In both cases the iterate averages are computed for one epoch. As shown in Figs. 2 and 5, since this is a "small" number of epochs, the higher order approximation from Eq. (13) is required to have an accurate approximation. In particular, Fig. 5 confirms that one epoch is not sufficient for Eq. (14) to be accurate in this case. This is consistent with our theoretical developments in Section 4.2.

## 5.3 Experiment 3: Large-scale inference for airline delay data—Poisson regression

Finally, to validate the value of our tuning recommendations in a more complex, clearly misspecified model, we examine the  the original airline dataset [United States Department of Tansportation, 2008] that the experiments in Pollock et al. [2020] were based upon. In this case the responses are non-negative integers and significantly zero-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 ($\approx$ 1.5 million observations) to estimate the "ground truth" quantities $(\theta_\star, \mathcal{J}_\star, \mathcal{I}_\star)$, and we apply the stochastic gradient algorithms to a dataset consisting of a random subsample of size 150,000 from the full 2008 dataset. For this example, the matrices $\mathcal{J}_\star$ and $\mathcal{I}_\star$ differ significantly in scale, and hence both

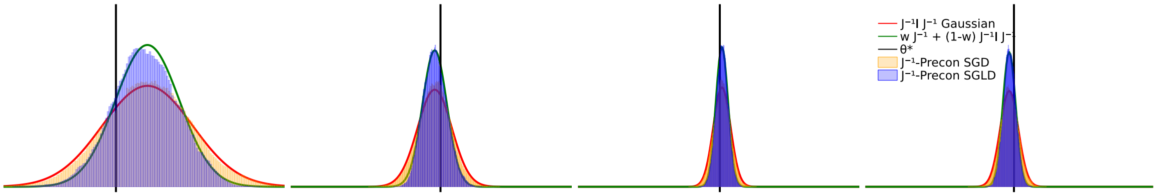
| | Experiment 1 | | Experiment 2 | | Experiment 3 | |
|---|---|---|---|---|---|---|
| Method | Emp. | Pred. | Emp. | Pred. | Emp. | Pred. |
| SGD, no preconditioning | 3.2 | 3.2 | 150 | 480 | - | - |
| $\mathcal{J}_\star^{-1}$-preconditioned SGD | 1.1 | 1.0 | 1.2 | 1.0 | 1.5 | 1.0 |
| $\mathcal{I}_\star^{-1}$-preconditioned SGD | 2.3 | 2.8 | 1.0 | 1.0 | - | - |
| $\mathcal{J}_\star^{-1}$-preconditioned SGLD | 2.2 | 2.0 | 2.3 | 2.0 | 3.0 | 2.0 |

Table 2: Comparison of empirical and predicted mixing times (in epochs) for all experiments measured by integrated autocorrelation times (IACT). The empirical value is computed numerically from the run. The predicted value is computed following Table 1.



(a) (b) (c) (d)

Figure 2: Results of experiment 2. Figs. 2a and 2b show the univariate results for the marginal distributions of parameters 1 and 4 respective when $(\mathfrak{h}, \mathfrak{b}) = (1, 0)$, and $\mathfrak{b} = 0$. Figs. 2c and 2d show the predicted and actual entries of the variance-covariance matrix for iterate averages when $(\mathfrak{h}, \mathfrak{b})$ is $(1, 0)$ and $(1/2, 1/2)$ respectively. The predictions for the iterate average are based upon Eq. (13).



(a) Parameter 2 (b) Parameter 3 (c) Parameter 24 (d) Parameter 25

Figure 3: Univariate results for experiment 3.

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. Figure 3 and Table 2 show that both preconditioned methods behave exactly as predicted by the asymptotic theory in Sections 4.1 and 4.3, and the results directly confirm the predictions made in, and support the recommendations implied by, Table 1.

# 6 Discussion

Given their ubiquity, stochastic gradient methods for optimization and sampling have been analyzed from a range of mathematical perspectives. Our work represents a convergence between non-statistical, continuous-time analyses [e.g, Kushner and Yin, 2003] (and also an often heuristic machine learning literature), statistical, discrete-time characterizations [e.g., Toulis and Airoldi, 2017], and Markov chain analyses of constant–step-size algorithms [e.g., Dieuleveut et al., 2020].

By focusing on the practically relevant fixed–step-size, large-sample setting, we are able to characterize the stationary distributions of the limiting stochastic processes. In combination with our statistical perspective, we are able to derive Bernstein–von Mises theorems: Corollary 2 for the marginal iterates, which is relevant to sampling applications, and Corollary 3 for iterative averages, which is relevant to optimization. The latter result complements analogous characterizations of iterate averages with decreasing step size schedules and fixed data [Polyak and Juditsky, 1992, Kushner and Yin, 2003]. Both results show that iterate averaging is robust to the choices of tuning parameters, including preconditioning, and can provide statistically optimal numerical estimates of the optimum. Our Bernstein–von Mises theorems offer insight into misspecified settings and clarify potential benefits of using stochastic gradients—something present in previous work from the statistical, discrete-time perspective, but limited to the marginal behaviour of individual iterates [Toulis and Airoldi, 2017].

Compared to previous heuristic arguments, our theory provides a more precise delineation of when continuous-time approximations are applicable [c.f. Mandt et al., 2017, Li et al., 2018]. For example we show that (a) there is no requirement for batch sizes to be large enough that single iteration increments are approximately Gaussian, and (b) these scaling limits exist for much broader combinations of joint scaling of step size and sample size leading to different rates of contraction. At the same time, the precise nature of our results allow us to more clearly understand the limitations of scaling limit analyses: heuristic calculations involving the behaviour across a large number of iterations can be replaced with corresponding approximations from the limiting process precisely when the time horizon involved is $O(1)$ on the limiting time scale. Our iterate averaging results provide a case in point: in order to have the same rate of contraction as the posterior distribution and/or the MLE, we must carefully choose the scaling of the step size and batch size together to ensure the spatial scaling is of order $1/\sqrt{n}$. Furthermore, Corollary 3 shows that, in general, iterate averaging must be done over $m \gg 1$ epochs for the covariance of the iterate average to accurately approximate the (rescaled) covariance of the MLE (which is in contrast to the claims of Mandt et al. [2017]).

Our results are advances in two of the key stochastic gradient MCMC research areas identified by Nemeth and Fearnhead [2021]. Namely, by invoking a large-sample limit we are able to provide results that circumvent strong convexity assumptions, and we are able to provide comprehensive analyses of various tuning combinations and to make tuning recommendations that can be implemented by or for practitioners. Using this new-found understanding, we were able to explain the empirical results of some critiques of SGLD-like methods and to show that with adequate tuning (that we identify) they would have performed significantly better.

Overall, our rigorous, continuous-time statistical approach to analyzing stochastic gradient algorithms complements existing work, yielding new insights into the practical effectiveness of stochastic gradient methods, and opens new avenues for future research. Because our results are expressed in terms of the joint scaling/choice of dataset size, step size, batch size, and other algorithm parameters, users can gain insight into a stochastic gradient algorithm's statistical behavior for specific choices of these values, which is not possible when taking the limit with the dataset size fixed.

Besides the concrete guidance for tuning SGAs and the explanations of prior work's empirical observations, our methods also lay the foundation for similar analyses of other SGAs and data generating models. Two such examples would be hierarchical models where the parameter dimension depends on the sample size and nonparametric models. Such analyses would allow for a systematic and fair comparison between inference methods and provide a better understanding of when stochastic gradient methods are effective. Another interesting new line of inquiry is to quantify in which finite-sample regimes our asymptotic results hold, which would enable more precise guidance for use in practice.

# References

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

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *International Conference on Machine Learning*, pages 1591–1598, 2012b.

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.

X. Chen, J. D. Lee, X. T. Tong, and Y. Zhang. Statistical inference for model parameters in stochastic gradient descent. *Annals of Statistics*, 48(1):251–273, 2020.

J. Coullon, L. South, and C. Nemeth. Efficient and generalizable tuning strategies for stochastic gradient MCMC. *Statistics and Computing*, 33(3):66, 2023. ISSN 0960-3174. doi: 10.1007/s11222-023-10233-3.

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.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

A. Durmus and E. Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

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.

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

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.

H. Haario, E. Saksman, and J. Tamminen. An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001. URL `http://www.jstor.org/stable/3318737`.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Univ of California Press, 1967.

J. H. Huggins and J. W. Miller. Using bagged posteriors for robust inference and model criticism. *arXiv*, arXiv:1912.07104 [stat.ME], 2019.

J. H. Huggins and J. W. Miller. Reproducible Model Selection Using Bagged Posteriors. *Bayesian Analysis*, 18(1):79–104, 2023.

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.

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*. Springer, 2003.

H. J. Kushner and H. Huang. Asymptotic properties of stochastic approximations with constant coefficients. *SIAM Journal on Control and Optimization*, 19(1):87–105, 1981. doi: 10.1137/0319007. URL `https://doi.org/10.1137/0319007`.

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.

T. Li, L. Liu, A. Kyrillidis, and C. Caramanis. Statistical inference using sgd. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, 2011.

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. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

G. C. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

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

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

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

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

H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22 (3):400 – 407, 1951.

G. O. Roberts and J. S. Rosenthal. Optimal scaling for various metropolis-hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

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

W. Rudin. Functional analysis. *McGraw Hill*, 1991.

J. E. Stafford. A robust adjustment of the profile likelihood. *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. *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. *The Annals of Statistics*, 45(4):1694–1727, 2017.

P. Toulis, T. Horel, and E. M. Airoldi. The proximal robbins–monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):188–212, 2021.

B. o. T. S. United States Department of Tansportation. Data expo 2009: Airline on time data, 2008. URL `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/HG7NV7`.

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

H. Walk. An invariance principle for the robbins-monro process in a hilbert space. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 39(2):135–150, 1977.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, pages 681–688, 2011.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

# A Further applications and extensions

We now discuss applications and extensions of our scaling limit to more complex, practically relevant stochastic gradient algorithms. 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, Ma and Yarats, 2018] aim to improve on SGLD by decreasing 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 Nagapetyan et al. [2017] and Baker et al. [2019] do this with a clever choice of control variates. Lastly, in practice, often parameter spaces are constrained, and we show that this does not affect the scaling limit.

## A.1 Applications to momentum-based algorithms

Special cases of our results include momentum-based acceleration of SGD such as the quasi-hyperbolic momentum algorithm of Ma and Yarats [2018], which includes many momentum-based algorithms as special cases (e.g., Nesterov's accelerated gradient, PID control algorithms [An et al., 2018], and more; see [Ma and Yarats, 2018, Table 1]). 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 $\tilde{\Theta} = \Theta \times \mathbb{R}^d$, for $\tilde{\theta} = (\theta, \psi) \in \tilde{\Theta}$, we extend the log-likelihood to the phase space according to $\tilde{\ell}(\tilde{\theta}; x) = \ell(\theta; x) - \psi^\top M^{-1}\psi/2$, and lift the prior to phase space using the (improper) prior $\tilde{\pi}^{(0)}(\tilde{\theta}) = \pi^{(0)}(\theta)$. 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 $\tilde{\Gamma} = \left[\begin{smallmatrix} 0 & -I \\ I & \Gamma \end{smallmatrix}\right]$ and $\tilde{\Lambda} = \left[\begin{smallmatrix} 0 & 0 \\ 0 & \Gamma \end{smallmatrix}\right]$. This yields a combined parameter update formula of

$$\theta_{k+1}^{(n)} = \theta_k^{(n)} + \frac{h^{(n)}}{2} M^{-1} \tilde{\theta}_k^{(n)}, \qquad \psi_{k+1}^{(n)} = \left(I - \frac{h^{(n)}\Gamma}{2} M^{-1}\right) \psi_k^{(n)} + \frac{h^{(n)}}{2} \hat{G}_k^{(n)} + \sqrt{\frac{h^{(n)}}{\beta^{(n)}} \Gamma}\, \xi_k. \tag{15}$$

The corresponding limiting process is $\mathrm{d}\tilde{\vartheta}_t = -\frac{1}{2}\tilde{B}\tilde{\vartheta}_t \mathrm{d}t + \sqrt{\tilde{A}}\mathrm{d}\tilde{W}_t$, where $\tilde{W}_t$ is a $2d$-dimensional standard Brownian motion, and the drift and diffusion matrices are, respectively, $\tilde{B} = c_h \left[\begin{smallmatrix} 0 & -M^{-1} \\ \mathcal{J}_\star & \Gamma M^{-1} \end{smallmatrix}\right]$ and $\tilde{A} = \mathbb{I}_{[\mathfrak{b}+\mathfrak{h}\leq\mathfrak{t}]} \frac{c_h^2 \overline{c_b}}{4c_b} \left[\begin{smallmatrix} 0 & 0 \\ 0 & \mathcal{I}_\star \end{smallmatrix}\right] + \mathbb{I}_{[\mathfrak{t}\leq\mathfrak{b}+\mathfrak{h}]} \frac{c_h}{c_\beta} \left[\begin{smallmatrix} 0 & 0 \\ 0 & \Gamma \end{smallmatrix}\right]$.

### A.2 Extension to control variates

SGLD methods with control variates aim improve the reliability of SGLD as an MCMC method to reduce the variance caused by mini-batching by introducing a “zero variance control variate” [Baker et al., 2019, Nagapetyan et al., 2017]. 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 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. Hence, the spatial scaling can always be chosen so that the noise from Gaussian innovations persists in the limit by taking $\mathfrak{w} = \mathfrak{t}/2$, where the corresponding limiting process takes the form of Eq. (7) with $B = c_h \Gamma \mathcal{J}_\star$ and $A = \frac{c_h}{c_\beta}\Lambda$.

### A.3 Extension to constrained parameter spaces

If $\Theta \subsetneq \mathbb{R}^d$, 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 been without adjusting for the constraint. Moreover, these conditions both satisfy that the distance between iterates is constrained by what the distance would have been without adjusting for the constraint. In Appendix I we consider boundary dynamics satisfying a generalized version of this property. When $\Theta \subsetneq \mathbb{R}^d$ and $\theta_\star \in \text{interior}(\Theta)$ the proof is essentially the same because, intuitively, the assumption that $\vartheta^{(n)}(0) \rightsquigarrow \vartheta(0)$ ensures that the processes we consider all start near $\theta_\star$ and away from the boundary of $\Theta$, and thus the spatial scaling drives the boundary of $\Theta$ outside any bounded set.

# B 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.

### B.1 Bernstein-von Mises under misspecification

**Definition 1.** *The first and second order* Fisher information matrices*, $\mathcal{I}$ and $\mathcal{J}$ respectively, are defined for a log-likelihood function $\ell$ and probability distribution $P$ by*

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

Let $\mathcal{X}$ be a Polish space with $\sigma$-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 $\mathbf{X}^{(\mathbb{N})} := (X_i)_{i\in\mathbb{N}} \sim P^{\otimes\mathbb{N}}$. Let $n \in \mathbb{N}$ denote a sample size, let $[n] := \{1, \dots, n\}$, and let $\mathbf{X}^{(n)} := (X_i)_{i\in[n]} \sim P^{\otimes n}$ be an I.I.D. sample of size $n$ from $P$.

Let $\Theta \subseteq \mathbb{R}^d$ be open and nonempty, let $Q$ be a regular conditional distribution from $\Theta$ 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[6].

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 $\mathrm{d}Q_\theta/\mathrm{d}\mu$ for each $\theta \in \Theta$. Let $\ell(\theta; x) := \log q_\theta(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$. We consider $\mathbb{M} := \{Q_\theta \mid \theta \in \Theta\}$ to be a *model* for $P$. The model is *well-specified* when $P \in \mathbb{M}$, and is *misspecified* otherwise. The *pseudo-true parameter* of the model is defined as $\theta_\star := \arg\max_{\theta \in \Theta} \underset{X \sim P}{\mathbb{E}} \ell(\theta; X)$. If $\mu \ll P$ then

$$\theta_\star = \arg\max_{\theta \in \Theta} \underset{X \sim P}{\mathbb{E}} \ell(\theta; X) = \arg\min_{\theta \in \Theta} \mathrm{KL}\left(P \| Q_\theta\right).$$

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

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

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

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

where for a random variable or $\sigma$-field $G$, an expectation operator $\mathbb{E}$ and a random variable $Y$, $\mathbb{E}^G(Y)$ is the conditional expectation of $Y$ given $G$. The posterior $\Pi^{(\mathbf{X}^{(n)})}$ can be viewed as a probability kernel from $\mathcal{X}^n$ to $\Theta$.

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

$$\pi^{(\mathbf{X}^{(n)})}(\theta) \propto \pi^{(0)}(\theta) \prod_{i \in [n]} q_\theta(X_i) = \pi^{(0)}(\theta) \exp\left(\sum_{i \in [n]} \ell(\theta; X_i)\right). \tag{16}$$

Let $\widehat{\theta}^{(n)} := \arg\max_{\theta \in \Theta} \sum_{i \in [n]} \ell(\theta; X_i)$ denote the maximum likelihood estimator (MLE) of $\theta_\star$ given the data $\mathbf{X}^{(n)}$. Posterior distributions have a general tendency to concentrate around the MLE as the sample size increases. Therefore, we will often reparameterize the model by considering a *local parametrization*, where to each parameter $\theta \in \Theta$ we associate a *local parameter*, $\vartheta \in \sqrt{n}\left(\Theta - \widehat{\theta}^{(n)}\right)$ based on the identification

$$\vartheta = \sqrt{n}\left(\theta - \widehat{\theta}^{(n)}\right)$$

and the *local model* is given by

$$\mathbb{M}^{(\mathbf{X}^{(n)})} := \left\{ Q_{\widehat{\theta}^{(n)} + \frac{1}{\sqrt{n}}\vartheta} \mid \vartheta \in \sqrt{n}\left(\Theta - \widehat{\theta}^{(n)}\right) \right\}.$$

The random localization map is given by

$$\mathrm{loc}_{\mathbf{X}^{(n)}} : \theta \mapsto \sqrt{n}\left(\theta - \widehat{\theta}^{(n)}\right)$$

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

**Proposition 2** (BvM under model misspecification, Kleijn and Van der Vaart [2012]). *Under regularity conditions,*

$$\left\| [\mathrm{loc}_{\mathbf{X}^{(n)}}]_\sharp \Pi^{(\mathbf{X}^{(n)})} - \Phi \right\|_{\mathrm{TV}} \xrightarrow{P} 0.$$

*with* $\theta_\star = \arg\max_{\theta \in \Theta} \underset{X \sim P}{\mathbb{E}} \ell(\theta; X)$, $\mathcal{J}_\star = -\underset{X \sim P}{\mathbb{E}}\left[\nabla^{\otimes 2} \ell(\theta_\star;\ X)\right]$, *and* $\Phi = \mathrm{N}\left(0, \mathcal{J}_\star^{-1}\right)$.

[6] $\Theta$ is equipped with the Borel $\sigma$-field inherited from $\mathbb{R}^d$

## B.2 Convergence modes of measures and operators

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

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

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

- $\mu_n$ converges in *total variation* to $\mu$, denoted by $\mu_n \overset{\text{TV}}{\to} \mu$, *if and only if*

$$\sup_{f\in B(\mathcal{A})} \frac{|\mu_n f - \mu f|}{\|f\|_\infty} \to 0.$$

- if $\mathcal{A}$ is also a topological space and the $\sigma$-field on $\mathcal{A}$ is the Borel $\sigma$-field, then $\mu_n$ converges *in distribution* (also called *weakly*) to $\mu$, denoted by $\mu_n \rightsquigarrow \mu$, *if and only if* for all $f \in \overline{C}(\mathcal{A})$, $|\mu_n f - \mu f| \to 0$.

Clearly

$$\left(\mu_n \overset{\text{TV}}{\to} \mu\right) \implies \left(\mu_n \overset{\text{s}}{\to} \mu\right) \implies (\mu_n \rightsquigarrow \mu)$$

while the converses do not hold in general.

For a Banach Space $L$ with norm $\|\cdot\|$ denote its dual space (the space of all bounded linear operators on $L$) by $L'$. $L'$ is a Banach space with norm $\|y\| := \sup_{x\in L\setminus\{0\}} |fx| / \|x\|$ for all $f \in L'$. Denote the set of bounded linear operators from $L$ to itself by $\mathcal{B}(L)$. $\mathcal{B}(L)$ is also a Banach space with norm given by $\|T\| = \sup_{x\in L\setminus\{0\}} \|Tx\| / \|x\|$.

For a sequence of bounded linear operators, $\{T_n\}_{n\in\mathbb{N}}$, and a bounded linear operator, $T$, all mapping a Banach Space $L$ to itself, we have the following modes of convergence:

- $T_n$ converges *in norm* to $T$ if and only if

$$\|T_n - T\| = \sup_{(x,y)\in L\times L'} \frac{|\langle y,\ (T_n - T)x\rangle|}{\|x\|\,\|y\|} \to 0 \tag{17}$$

- $T_n$ converges *strongly* to $T$, denoted $T_n \overset{\text{s}}{\to} T$ if and only if for all $x \in L$

$$\sup_{y\in L'} \frac{|\langle y,\ (T_n - T)x\rangle|}{\|y\|} \to 0 \tag{18}$$

Clearly

$$(\|T_n - T\| \to 0) \implies \left(T_n \overset{\text{s}}{\to} T\right)$$

while the converse does not hold in general.

## B.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 *infinitesimal generator* (or just *generator*, for brevity) of the semigroup $T$ is the (possibly unbounded) linear operator defined by

$$Af = \lim_{t \to 0^+} \frac{T_t f - f}{t}$$

for $f \in \mathrm{dom}(A) = \{f \in L \mid \lim_{t\to 0^+} (T_t f - f)/t \text{ exists}\}$. Let

$$\hat{C}(\mathbb{R}^d) = \left\{ f \in C(\mathbb{R}^d) \mid \forall \epsilon > 0\ \exists K_{f,\epsilon} \subset \mathbb{R}^d \text{ compact with } \sup_{\theta \notin K_{f,\epsilon}} |f(\theta)| \leq \epsilon \right\}$$

Then $\hat{C}(\mathbb{R}^d)$ is a Banach space under the norm $\|f\|_\infty = \sup_{\theta\in\mathbb{R}^d} |f(\theta)|$. The dual space of $\hat{C}(\mathbb{R}^d)$ is the space of bounded signed measures under the *total variation norm*

$$\|\mu\|_{\mathrm{TV}} = \sup_{\substack{f \in \hat{C}(\mathbb{R}^d) \\ \|f\|_\infty \leq 1}} \left| \int f(\theta)\mu(d\theta) \right|.$$

We will work with $(L, \|\cdot\|) = \left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$. A semigroup on $\left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$ is *positive* if

v) $f \geq 0 \implies Tf \geq 0$.

A semigroup on $\left(\hat{C}(\mathbb{R}^d), \|\cdot\|_\infty\right)$ is *Feller* if it is strongly continuous, contractive, and positive.

Semigroups naturally model the *forward operators* of Markov processes in continuous time. If $X_t$ is a Markov process with transition kernels $k_t(\cdot,\cdot)$ then the forward operator corresponding to the Markov process (equivalently, corresponding to its transitio kernels) is defined by

$$T_t f(x) = \mathbb{E}_x f(X_t) = \int f(y) k_t(x, dy) \tag{19}$$

where $\mathbb{E}_x$ denotes expectation under the law of the Markov process given when $X(0) = x$ almost surely. The semigroup property is then equivalent to the Kolmogorov forward equation.

The generator, $A$, of a Feller semigroup $T$ has a dense domain; $\mathrm{dom}(A)$ is dense in $\hat{C}(\mathbb{R}^d)$. A Markov process for which the corresponding forward operators form a a Feller semigroup is called a *Feller process*. Feller processes have a richly developed theory; see, for example, Ethier and Kurtz [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 $t > 0$, $\lim_{s\to t^-} X(s)$ exists and $\lim_{s \to t^+} X_t$. Second for each $I \in \{[0,T] \mid T > 0\} \cup \{\mathbb{R}_+\}$, the collection of all càdlàg functions from $I$ to $\mathbb{R}^d$ is a separable and complete metric space under the *Skorohod metric* [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 reference. 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 $\overline{C}(\mathbb{R}^d)$.

**Proposition 3** (Approximation of Markov Chains (compiled from Ethier and Kurtz [2009])**.** *Let $A : C_c^\infty(\mathbb{R}^d) \to \overline{C}(\mathbb{R}^d)$ be linear and suppose that the closure of the graph of $A$ (with respect to the graph norm defined by $\|f\|_A = \|f\|_\infty + \|Af\|_\infty$ for all $f \in L$) generates a Feller semigroup $T$ on $\mathbb{R}^d$. Let $(\vartheta_t)_{t\in\mathbb{R}_+}$ be a Markov process with forward operator semigroup $T$. Let $\left((\theta_k^{(n)})_{k\in\mathbb{N}\cup\{0\}}\right)_{n\in\mathbb{N}}$ be a sequence of (discrete-time) Markov chains on $\mathbb{R}^d$ with respective transition kernels $(U^{(n)})_{n\in\mathbb{N}}$. Suppose that $0 < \alpha^{(n)} \to \infty$, and let*

$$A^{(n)} = \alpha^{(n)}\left(U^{(n)} - I\right) \qquad T_t^{(n)} = \left(U^{(n)}\right)^{\lfloor \alpha^{(n)}\ t \rfloor} \qquad \vartheta_t^{(n)} = \theta^{(n)}_{\lfloor \alpha^{(n)}\ t\rfloor}.$$

*If $\left\|A^{(n)}f - Af\right\|_\infty \to 0$ for all $f \in C_c^\infty(\mathbb{R}^d)$, then* *(a)* $T_t^{(n)} \xrightarrow{s} T_t$ *for each* $t > 0$*, and*

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

*Proof of Proposition 3.* (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].

### B.4 Miscellaneous notation and definitions

**Definition 2** (Convergence in Probability to a constant)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space, let* $(\mathcal{X}, \tau)$ *be a topological space endowed with the* $\sigma$*-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* $X_n \xrightarrow{p} x$*, when for every neighbourhood* $x \in U \in \tau$ *we have*

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

**Lemma 1.** *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* $n_{m_k}$ *such that* $X_{n_{m_k}} \to x$ *almost surely as* $k \to \infty$ *then* $X_n \xrightarrow{p} x$*.*

*If* $(\mathcal{X}, \tau)$ *is first-countable then the converse also holds; if* $X_n \xrightarrow{p} x$ *then for every sub-sequence* $n_m$ *there is a sub-sub-sequence* $n_{m_k}$ *such that* $X_{n_{m_k}} \to x$ *almost surely as* $k \to \infty$*.*

The proof of this result is the same as in Durrett [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.

## C 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 2.

**Theorem 2** (Scaling Limits of SGD/SGLD/LD (Full))**.** *Suppose that* $(\theta_k^{(n)})_{k\in\mathbb{N}}$ *evolves according to the gradient-based algorithm in Eq. (35) with step size* $h^{(n)} = c_h n^{-\mathfrak{h}}$*,* $b^{(n)} = \lfloor c_b n^{\mathfrak{b}} \rfloor$*,* $\beta^{(n)} = c_\beta n^{\mathfrak{t}}$*, all other tuning parameters constant in* $n$*. Let* $\theta_\star \in \mathbb{R}^d$*. Let* $\mathbf{X}^{(\mathbb{N})} = (X_i)_{i\in\mathbb{N}} \sim P^{\otimes\mathbb{N}}$*, and* $\widehat{\theta}^{(n)}$ *be a critical point of the log-likelihood function* $\sum_{i=1}^n \ell(\cdot, X_i)$ *for each* $n \in \mathbb{N}$*; that is* $\sum_{i=1}^n \nabla\ell(\widehat{\theta}^{(n)}, X_i) = 0$ *for all* $n \in \mathbb{N}$*.*

*Let* $\vartheta_t^{(n)} = w^{(n)}\left(\theta_{\lfloor\alpha(n)t\rfloor}^{(n)} - \widehat{\theta}^{(n)}\right)$*, where* $w^{(n)} = n^{\mathfrak{w}}$*,* $\alpha^{(n)} = n^{\mathfrak{a}}$*,* $\mathfrak{w} \in (0,1)$*,*

$$\mathfrak{a} = \min\left\{\mathfrak{h},\ (\mathfrak{t} + \mathfrak{h} - 2\mathfrak{w}),\ (\mathfrak{b} + 2\mathfrak{h} - 2\mathfrak{w})\right\}.$$

*If Assumptions 1 to 5 all hold,* $\mathfrak{a} > 0$*, and* $\vartheta^{(n)}(0) \rightsquigarrow \vartheta(0)$ *then*

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

$$d\vartheta_t = -\frac{c_{\mathfrak{d}}}{2}\Gamma\mathcal{J}(\theta_\star)\vartheta_t dt + \sqrt{c_{\mathfrak{g}}\Lambda + c_{\mathfrak{mb}}\Gamma\mathcal{I}(\theta_\star)\Gamma'}\, dW_t,$$

*with*

$$c_{\mathfrak{d}} = \begin{cases} c_h & \mathfrak{a} = \mathfrak{h} \\ 0 & \mathfrak{a} < \mathfrak{h} \end{cases}, \qquad c_{\mathfrak{g}} = \begin{cases} \frac{c_h}{c_\beta} & \mathfrak{a} = \mathfrak{h} + \mathfrak{t} - 2\mathfrak{w} \\ 0 & \mathfrak{a} < \mathfrak{h} + \mathfrak{t} - 2\mathfrak{w} \end{cases}$$

*and*

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

2. *If $T^{(n)}$ and $T$ are defined as in Proposition 3, then under the conditions above, every subsequence of $\left(T^{(n)}\right)_{n\in\mathbb{N}}$, then $\left(T^{(n_m)}\right)_{m\in\mathbb{N}}$, has a further sub-subsequence, $\left(T^{(n_{m_k})}\right)_{k\in\mathbb{N}}$, such that with probability 1, $T_t^{(n_{m_k})} \xrightarrow{s} T_t$ for all $t > 0$.*

Before beginning the proof of this result, Theorem 2, we require the following lemma, which is used to turn the moment conditions in our assumptions into bounds on the magnitudes of certain random variables that hold all but finitely often with probability 1.

**Lemma 2.** *Let $\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ be non-decreasing, right continuous with left limits, with $\alpha(0) = 0$, and $\lim_{t\to\infty} \alpha_t = \infty$. Let $Z_i \sim \mu$ for all $i \in \mathbb{N}$ (possibly not independent) with $Z_1 \geq 0$ almost surely such that $\mathbb{E}[\alpha(Z_1)] < \infty$. Let $\alpha^+ : u \mapsto \inf\{t \geq 0 \text{ s.t. } \alpha_t \geq u\}$ be the generalized inverse of $\alpha$. Then*

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

*Proof of Lemma 2.* 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

$$\infty > \mathbb{E}\left[(\alpha(Z_1))\right] = \int_0^\infty \mathbb{P}(W_1 > t)dt \geq \sum_{n=1}^\infty \mathbb{P}(W_1 > n) = \sum_{n=1}^\infty \mathbb{P}(W_n > n)$$

Therefore, from the Borel–Cantelli lemma $\mathbb{P}(W_n > n \quad \text{i.o.}) = 0$, and equivalently $\mathbb{P}(W_n \leq n \quad \text{a.b.f.o.}) = 1$. Now, whenever $W_n \leq n$ for all but finitely many $n$, then there exists $K \in \mathbb{N}$ and $I_1, \dots I_K \in \mathbb{N}$ with $W_n \leq n$ for all $n \in \mathbb{N} \setminus \{I_j : j \in [K]\}$. Therefore, for all $n \geq \max_{j\leq K} W_{I_j}$, $\max_{i\leq n} W_i \leq n$. Therefore $\mathbb{P}(\max_{i\leq n} W_i \leq n \quad \text{a.b.f.o.}) = 1$, and equivalently $\mathbb{P}(\max_{i\leq n} W_i > n \quad \text{i.o.}) = 0$. Finally, $W_i > n$ if and only if $Z_i > \alpha^+(n)$, hence

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

## C.1 Proof of Theorem 2

Let $\mathcal{J}_\star = \mathcal{J}(\theta_\star)$ and $\mathcal{I}_\star = \mathcal{I}(\theta_\star)$.

The proof proceeds in the following stages. In Appendix C.1.1, we will reduce the problem of weak convergence in the Skorohod topology in probability to one of weak convergence in the Skorohod topology almost-surly along subsequences and construct appropriate such subsequences. In Appendix C.1.2 we introduce notation that will be useful in the remainder of the proof. In Appendix C.1.3 we discuss what is needed to apply Proposition 3 to establish the processes converge weakly in the Skorohod topology almost-surely. This amounts to showing that the difference between the approximate generator and limiting generator evaluated a smooth test function with compact support vanishes uniformly. We will examine this difference in two regimes. First, in Appendix C.1.4, we will consider arguments sufficiently far from the support of the test function. Then, in Appendix C.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 C.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 C.1.7 to C.1.9. The third type of term is the remainder term, which is handled in Appendix C.1.10. Putting all of this together allows us to apply Proposition 3 along our subsequences, establishing the main result.

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

Let

$$\begin{aligned}\Upsilon^{(n)} &= \max\left(\Upsilon_1^{(n)}, \Upsilon_2^{(n)}, \Upsilon_3^{(n)}\right), \\ \Upsilon_1^{(n)} &= n^{q_3}\left\|\widehat{\theta}^{(n)} - \theta_\star\right\|, \\ \Upsilon_2^{(n)} &= \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star)\right\|, \\ \Upsilon_3^{(n)} &= \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)}\left\|\widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}(\theta_\star)\right\|.\end{aligned}$$

Each of the $\Upsilon$ 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, $\Upsilon^{(n)} \xrightarrow{\text{p}} 0$. Then, by Lemma 1, for every subsequence $(n_m)_{m\in\mathbb{N}}$ there is a further sub-subsequence $(n_{m_k})_{k\in\mathbb{N}}$ so that this convergence is almost sure. Along an arbitrary such sub-subsequence, we will verify that $(\vartheta^{(n_{m_k})})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ in the Skorohod topology almost surely. Since weak convergence is metrizable (e.g., by the Levi–Prokhorov metric, and hence corresponds to a topology on probability distributions), and since for any subsequence $(n_m)_{m\in\mathbb{N}}$ we will have shown a further subsequence $(n_{m_k})_{k\in\mathbb{N}}$ such that $(\vartheta_t^{(n_{m_k})})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ a.s., by Lemma 1 it must hold that $(\vartheta_t^{(n)})_{t\in\mathbb{R}_+} \rightsquigarrow (\vartheta_t)_{t\in\mathbb{R}_+}$ in probability.

Now, let $(n_m)_{m\in\mathbb{N}}$ be an arbitrary subsequence[7] of $\mathbb{N}$ such that $\Upsilon^{(n_m)} \xrightarrow{\text{a.s.}} 0$. Let $\Omega$ denote the underlying probability space. Let

$$\begin{aligned}\Omega^{(0)} &= \bigcap_{i=1}^{3}\Omega^{(i)}, \\ \Omega^{(1)} &= \left\{\Upsilon^{(n_m)} \to 0\right\}, \\ \Omega^{(2)} &= \left\{\max_{i\in[n]}\|\nabla\ell(\theta_\star; X_i)\| \le n^{1/p_2} \quad \text{a.b.f.o}\right\}, \\ \Omega^{(3)} &= \left\{\max_{i\in[n]}\left\|\nabla^{\otimes 2}\ell(\cdot; X_i)\right\|_\infty \le n^{1/p_3} \quad \text{a.b.f.o.}\right\}.\end{aligned}$$

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

### C.1.2 Additional notation used in the proof

We notate the increments of the localized iterative algorithms (given that $\vartheta_0^{(n)} = \vartheta$) due to the Gaussian innovation ($\xi$), 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

$$\begin{aligned}\Delta_\xi^{(n)} &:= w^{(n)}\sqrt{h\beta^{-1}\Lambda}\;\xi_1, \\ \Delta_{\pi^{(0)}}^{(n)}(\vartheta) &:= \frac{hw^{(n)}\Gamma}{2n}\nabla\log\pi^{(0)}\left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1}\vartheta\right), \\ \Delta_\ell^{(n)}(\vartheta) &:= \frac{hw^{(n)}\Gamma}{2b^{(n)}}\sum_{j\in\left[b^{(n)}\right]}\nabla\ell\left(\widehat{\theta}^{(n)} + (w^{(n)})^{-1}\vartheta;\; X_{I_1^{(n)}(j)}\right), \text{ and} \\ \Delta^{(n)}(\vartheta) &:= \Delta_\xi^{(n)} + \Delta_{\pi^{(0)}}^{(n)}(\vartheta) + \Delta_\ell^{(n)}(\vartheta).\end{aligned}$$

We define the sequence of operators $A^{(n)}$ by

$$[A^{(n)}f](\vartheta) = \alpha^{(n)}\left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[f(\vartheta + \Delta^{(n)}(\vartheta))\right] - f(\vartheta)\right). \qquad (20)$$

[7] Since every sub-subsequence is itself a subsequence, we can simplify our notation from here onward.

for all $n \in \mathbb{N}$, and all $f \in C_c^\infty(\mathbb{R}^d)$, where $\alpha^{(n)} = n$. The generator of the (presumed, at this point) limiting OU process is given by

$$[Af](\vartheta) = -\left\langle \frac{c_{\mathtt{d}}}{2}\Gamma\mathcal{J}_\star\vartheta,\ \nabla f(\vartheta)\right\rangle + \frac{1}{2}\left(c_{\mathtt{g}}\Lambda + c_{\mathtt{mb}}\Gamma\mathcal{I}_\star\Gamma'\right) : \nabla^{\otimes 2} f(\vartheta) \tag{21}$$

### C.1.3 How Proposition 3 is applied

Consider a single realization of $\mathbf{X}^{(\mathbb{N})} \in \Omega^{(0)}$. Our goal, now, is to apply Proposition 3, treating $\mathbf{X}^{(\mathbb{N})}$ as fixed. To do so, it suffices to show that for each $f \in C_c^\infty(\mathbb{R}^d)$ we have

$$\lim_{m\to\infty} \sup_{\vartheta \in \mathbb{R}^d} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

For an arbitrary test function, $f \in C_c^\infty(\mathbb{R}^d)$, with compact support $K_0$, we will show this in two parts. First we will identify a compact extension, $K_1 \supset K_0$ to the compact support of $f$ such that

$$\lim_{m\to\infty} \sup_{\vartheta \in K_1^c} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

Then we will separately show that

$$\lim_{m\to\infty} \sup_{\vartheta \in K_1} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| = 0.$$

### C.1.4 Convergence away from the test function support

For all $\vartheta \in K_0^c$, $f(\vartheta) = 0$, $\nabla f(\vartheta) = 0$, and $\nabla^{\otimes 2} f(\vartheta) = 0$. Therefore, for any $K_1 \supset K_0$,

$$\begin{aligned}
&\sup_{\vartheta \in K_1^c} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| \\
&\qquad \le \alpha^{(n_m)} \|f\|_\infty \sup_{\vartheta \in K_1^c} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left[\vartheta + \Delta^{(n_m)}(\vartheta) \in K_0\right].
\end{aligned} \tag{22}$$

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

$$c_0 = \frac{c_h \|\Gamma\|}{2}\left(3 + \left\|\nabla \log \pi^{(0)}(\theta_\star)\right\|\right) + \sqrt{c_h / c_\beta}\,\|\Lambda\|.$$

Then, using Eq. (22) and $\Delta^{(n_m)}(\vartheta) = \Delta_\xi^{(n_m)}(\vartheta) + \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta) + \Delta_\ell^{(n_m)}(\vartheta)$,

$$\begin{aligned}
&\sup_{\vartheta \in K_1^c} \left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| \\
&\quad \le \alpha^{(n_m)} \|f\|_\infty \sup_{\|\vartheta\| > 2R_0 + 2c_0} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left[\left\|\vartheta + \Delta^{(n_m)}(\vartheta)\right\| \le R_0\right] \\
&\quad \le \alpha^{(n_m)} \|f\|_\infty \sup_{\|\vartheta\| > 2R_0 + 2c_0} \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left[\left\|\Delta_\xi^{(n_m)}\right\| \ge \|\vartheta\| - \left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\| - \left\|\Delta_\ell^{(n_m)}(\vartheta)\right\| - R_0\right].
\end{aligned} \tag{23}$$

For $\vartheta \in K_1^c$, using the assumption that $\nabla \log \pi^{(0)}$ is $L_0$-Lipschitz and $h^{(n)} = c_h n^{\mathfrak{h}}$ and $w^{(n)} = n^{\mathfrak{w}}$,

$$\begin{aligned}
\left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\| &\le \frac{h^{(n_m)} w^{(n_m)} \|\Gamma\|}{2n_m} \left\|\nabla \log \pi^{(0)}\left(\widehat{\theta}^{(n_m)} + (w^{(n_m)})^{-1}\vartheta\right)\right\| \\
&\le \frac{h^{(n_m)} w^{(n_m)} \|\Gamma\|}{2n_m} \left(\left\|\nabla \log \pi^{(0)}(\theta_\star)\right\| + L_0 \left\|\widehat{\theta}^{(n_m)} - \theta_\star\right\| + \frac{L_0 \|\vartheta\|}{w^{(n_m)}}\right) \\
&\le \frac{c_h n_m^{\mathfrak{w} - \mathfrak{h} - 1} \|\Gamma\|}{2} \left(\left\|\nabla \log \pi^{(0)}(\theta_\star)\right\| + L_0 \left\|\widehat{\theta}^{(n_m)} - \theta_\star\right\| + \frac{L_0 \|\vartheta\|}{n_m^{\mathfrak{w}}}\right),
\end{aligned}$$

and similarly

$$\begin{aligned}
&\left\|\Delta_{\ell}^{(n_m)}(\vartheta)\right\| \\
&\quad\leq \frac{h^{(n_m)}w^{(n_m)}\left\|\Gamma\right\|}{2b^{(n_m)}}\left\|\sum_{j\in\left[b^{(n_m)}\right]}\nabla\ell\left(\widehat{\theta}^{(n_m)}+(w^{(n_m)})^{-1}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)\right\| \\
&\quad\leq \frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}}\left\|\Gamma\right\|}{2b^{(n_m)}}\sum_{j\in\left[b^{(n_m)}\right]}\left(\left\|\nabla\ell\left(\theta_\star;\ X_{I_1^{(n_m)}(j)}\right)\right\|+L(X_{I_1^{(n_m)}(j)})\left\|\widehat{\theta}^{(n_m)}-\theta_\star\right\|+\frac{L(X_{I_1^{(n_m)}(j)})\left\|\vartheta\right\|}{{n_m}^{\mathfrak{w}}}\right) \\
&\quad\leq \frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}}\left\|\Gamma\right\|}{2}\left(L_\star(\mathbf{X}^{(n_m)})+L(\mathbf{X}^{(n_m)})\left\|\widehat{\theta}^{(n_m)}-\theta_\star\right\|+L(\mathbf{X}^{(n_m)})\frac{\left\|\vartheta\right\|}{{n_m}^{\mathfrak{w}}}\right)
\end{aligned}$$

where we define the (random) Lipschitz constants $L(X_i)$, $L_\star(\mathbf{X}^{(n_m)})$, and $L(\mathbf{X}^{(n_m)})$ by:

$$\begin{aligned}
L(X_i) &:= \left\|\nabla^{\otimes 2}\ell(\cdot;X_i)\right\|_\infty, \\
L_\star(\mathbf{X}^{(n_m)}) &:= \max_{i\leq n_m}\left\|\nabla\ell\left(\theta_\star;\ X_i\right)\right\|,\ \text{and} \\
L(\mathbf{X}^{(n_m)}) &:= \max_{i\leq n_m} L(X_i).
\end{aligned}$$

Using that $\mathbf{X}^{(\mathbb{N})}\in\Omega^{(0)}$, so that $\Upsilon^{(n_m)}\to 0$ etc., if $m$ is large enough that all of the following hold:

$$\begin{aligned}
\sup_{m'\geq m}\Upsilon^{(n_m)} &\leq \min(1, L_0^{-1}), \\
1 &\geq \sup_{m'\geq m}\frac{L_\star(\mathbf{X}^{(n_{m'})})}{n_{m'}^{1/p_2}}, \\
n_m &\geq \max((2c_h\left\|\Gamma\right\|)^{1/(1/p_3-\mathfrak{h})}, (2c_h L_0\left\|\Gamma\right\|)^{\frac{1}{\mathfrak{h}+1-\mathfrak{a}-\mathfrak{w}}}), \qquad \text{and} \\
1 &\geq \sup_{m'\geq m}\frac{L(\mathbf{X}^{(n_{m'})})}{n_{m'}^{1/p_3}};
\end{aligned}$$

then, using that $0<\mathfrak{w}<1$,

$$\left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\| \leq \frac{c_h\left\|\Gamma\right\|}{2}\left(\left\|\nabla\log\pi^{(0)}\left(\theta_\star\right)\right\|+1\right)+\frac{1}{4}\left\|\vartheta\right\|,$$

and

$$\begin{aligned}
\left\|\Delta_{\ell}^{(n_m)}(\vartheta)\right\| &\leq \frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}}\left\|\Gamma\right\|}{2}\left(n_m^{1/p_2}+n_m^{1/p_3}\Upsilon^{(n_m)}+n_m^{1/p_3-\mathfrak{w}}\left\|\vartheta\right\|\right) \\
&\leq \frac{c_h\left\|\Gamma\right\|}{2}\left(n_m^{1/p_2-\mathfrak{h}+\mathfrak{w}}+n_m^{1/p_3-\mathfrak{h}+\mathfrak{w}}\Upsilon^{(n_m)}+n_m^{1/p_3-\mathfrak{h}}\left\|\vartheta\right\|\right), \\
&\leq c_h\left\|\Gamma\right\|+\frac{1}{4}\left\|\vartheta\right\|.
\end{aligned}$$

Therefore, for $\vartheta\in K_1^c$ (and hence $\left\|\vartheta\right\|>2R_0+2c_0$),

$$\begin{aligned}
&\left\|\vartheta\right\|-\left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\|-\left\|\Delta_{\ell}^{(n_m)}(\vartheta)\right\|-R_0 \\
&\quad\geq \frac{1}{2}\left\|\vartheta\right\|-\frac{c_h\left\|\Gamma\right\|}{2}\left(3+\left\|\nabla\log\pi^{(0)}\left(\theta_\star\right)\right\|\right)-R_0 \\
&\quad\geq \sqrt{c_h/c_\beta\left\|\Lambda\right\|}.
\end{aligned}$$

Therefore, combining this with Eq. (23) and the definition of $\Delta_\xi^{(n_m)}(\vartheta)$,

$$\begin{aligned}\lim_{m\to\infty}\sup_{\vartheta\in K_1^c}\left|[A^{(n_m)}f](\vartheta)-[Af](\vartheta)\right| &\le \lim_{m\to\infty}\alpha^{(n_m)}\left\|f\right\|_\infty \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left(\|\xi_1\|\ge n_m^{\mathfrak{h}/2+\mathfrak{t}/2-\mathfrak{w}}\right)\\ &\le \lim_{m\to\infty}\alpha^{(n_m)}\left\|f\right\|_\infty d\ \mathbb{P}^{\mathbf{X}^{(\mathbb{N})}}\left(|\xi_{1,1}|\ge \frac{1}{\sqrt{d}}n_m^{\mathfrak{h}/2+\mathfrak{t}/2-\mathfrak{w}}\right)\\ &\le \lim_{m\to\infty} 2n_m^{\mathfrak{a}}\left\|f\right\|_\infty d\exp(-n_m^{\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}}/2d)\\ &= 0.\end{aligned}$$

since $\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}\ge\mathfrak{a}>0$.

#### C.1.5 Taylor expansion near the test function support

Recalling the definition of $A^{(n_m)}$ in Eq. (20), using the definition of the time-scaling factor $\alpha^{(n)}=n^{\mathfrak{a}}$, taking a second-order Taylor expansion of the test function $f\in C_c^\infty$, and applying the decomposition $\Delta^{(n_m)}(\vartheta)=\Delta_\xi^{(n_m)}(\vartheta)+\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)+\Delta_\ell^{(n_m)}(\vartheta)$,

$$\begin{aligned}&[A^{(n_m)}f](\vartheta)\\ &\quad=\alpha^{(n_m)}\left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[f\left(\vartheta+\Delta^{(n_m)}(\vartheta)\right)\right]-f(\vartheta)\right)\\ &\quad=\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla f(\vartheta),\ \Delta_\xi^{(n_m)}\right\rangle}_{[1.\xi]^{(n_m)}(\vartheta)=0}+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla f(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle}_{\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)}+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla f(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta)\right\rangle}_{[1.\ell]^{(n_m)}(\vartheta)}\\ &\qquad+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_\xi^{(n_m)},\ \Delta_\xi^{(n_m)}\right\rangle}_{[2.\xi\xi]^{(n_m)}(\vartheta)}+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla^{\otimes 2}f(\vartheta)\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_\xi^{(n_m)}\right\rangle}_{\left[2.\pi^{(0)}\xi\right]^{(n_m)}(\vartheta)=0}\\ &\qquad+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla^{\otimes 2}f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\xi^{(n_m)}\right\rangle}_{[2.\ell\xi]^{(n_m)}(\vartheta)=0}+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle}_{\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)}\\ &\qquad+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\nabla^{\otimes 2}f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}\right\rangle}_{\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)}+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle\frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta)\right\rangle}_{[2.\ell\ell]^{(n_m)}(\vartheta)}\\ &\qquad+\underbrace{n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[\frac{1}{6}\left[\nabla^{\otimes 3}f(\vartheta+S\Delta^{(n_m)}(\vartheta))\right]\left(\Delta^{(n_m)}(\vartheta),\Delta^{(n_m)}(\vartheta),\Delta^{(n_m)}(\vartheta)\right)\right]}_{[3.R]^{(n_m)}(\vartheta)}\end{aligned}$$

for some $S\in[0,1]$ depending on $f,\vartheta,\Delta^{(n_m)}(\vartheta)$, where $\nabla^{\otimes 3}f(\vartheta)$ is the trilinear from of third order partials of $f$ at $\vartheta$ (and hence is linear in each of its three arguments). Terms that are linear in $\Delta_\xi^{(n_m)}$ have mean 0 and can be eliminated outright, as indicated in their corresponding underbraces. Terms are labelled by the order of the term, followed by the increments that appear in the term; for example $[2.\ell\xi]^{(n_m)}(\vartheta)$ is the second order term involving a likelihood increment and a Gaussian noise (innovation) increment. The $R$ in $[3.R]^{(n_m)}(\vartheta)$ denotes that it is the *remainder*.

Recall that

$$[Af](\vartheta)=\underbrace{-\left\langle\frac{c_{\mathtt{d}}}{2}\Gamma\mathcal{J}_\star\vartheta,\ \nabla f(\vartheta)\right\rangle}_{[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)}+\underbrace{\frac{c_{\mathtt{g}}}{2}\Lambda:\nabla^{\otimes 2}f(\vartheta)}_{[\mathrm{II}.\Lambda](\vartheta)}+\underbrace{\frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma':\nabla^{\otimes 2}f(\vartheta)}_{[\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)}.$$

We have similarly labelled these terms, with the roman numeral denoting the order and the subsequent symbol denoting the coefficient matrix (up to scaling factors). Thus, after eliminating terms which are linear in $\Delta_{\xi}^{(n_m)}$, and thus have mean 0, the difference of approximate and limiting generator applied to the test function can be expressed as

$$\begin{aligned}
&\left|[A^{(n_m)}f](\vartheta) - [Af](\vartheta)\right| \\
&\quad \le \left|\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| + \left|\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| + \left|\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\qquad + \left|[1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)\right| \\
&\qquad + \left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\mathrm{II}.\Lambda](\vartheta)\right| \\
&\qquad + \left|[2.\ell\ell]^{(n_m)}(\vartheta) - [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right| \\
&\qquad + \left|[3.R]^{(n_m)}(\vartheta)\right|.
\end{aligned}$$

We will show that each of these seven terms vanish uniformly on $K_1$. The first three terms listed above, those non-remainder terms with no corresponding term in the limiting generator, will be handled first. Then we will handle each of the terms which corresponds to part of the limiting generator, and lastly we will handle the remainder term.

### C.1.6 Terms that do not contribute to the limit

$$\begin{aligned}
\left|\left[1.\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| &= n_m^{\mathfrak{a}}\left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla f(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\le \frac{c_h n_m^{\mathfrak{a}-\mathfrak{h}+\mathfrak{w}-1}\left\|\Gamma\right\|}{2}\left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla f(\vartheta),\ \nabla\log\pi^{(0)}\left(\widehat{\theta}^{(n_m)} + n_m^{-\mathfrak{w}}\vartheta\right)\right\rangle\right| \\
&\le \frac{c_h n_m^{\mathfrak{a}-\mathfrak{h}+\mathfrak{w}-1}\left\|\Gamma\right\|}{2}\left\|\nabla f\right\|_\infty\left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\left(\Upsilon^{(n_m)} + \frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right)\right),
\end{aligned}$$

which vanishes uniformly on $K_1$, since $\mathfrak{a}+\mathfrak{w}-\mathfrak{h}-1 \le \mathfrak{w}-1 < 0$.

$$\begin{aligned}
&\left|\left[2.\pi^{(0)}\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\quad = \left|n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\quad \le n_m^{\mathfrak{a}}\left\|\nabla^{\otimes 2}f\right\|_\infty\left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}-1}\left\|\Gamma\right\|}{2}\right)^2 \\
&\qquad \times\left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\left\|\widehat{\theta}^{(n_m)}-\theta_\star\right\| + L_0\frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right)^2
\end{aligned}$$

which vanishes uniformly since $\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-2 \le (2\mathfrak{w}-2)-h < 0$ (which follows from $\mathfrak{h} \ge \mathfrak{a}$ and $\mathfrak{w} < 1$).

$$\begin{aligned}
&\left|\left[2.\ell\pi^{(0)}\right]^{(n_m)}(\vartheta)\right| \\
&\quad = \left|n_m^{\mathfrak{a}}2\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_{\ell}^{(n_m)}(\vartheta),\ \Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\rangle\right| \\
&\quad \le 2n_m^{\mathfrak{a}}\left\|\nabla^{\otimes 2}f\right\|_\infty\left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}-1}\left\|\Gamma\right\|}{2}\right)\left(\frac{c_h n_m^{\mathfrak{w}-\mathfrak{h}}\left\|\Gamma\right\|}{2}\right) \\
&\qquad \times\left(\left\|\nabla\log\pi^{(0)}(\theta_\star)\right\| + L_0\Upsilon^{(n_m)} + L_0\frac{2R_0+2c_0}{n_m^{\mathfrak{w}}}\right) \\
&\qquad \times\left(n_m^{1/p_2} + n_m^{1/p_3}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{w}}\right)
\end{aligned}$$

which vanishes uniformly due to the assumptions of the relationship between $\mathfrak{h}, \mathfrak{a}, \mathfrak{w}, p_3, p_2$ under each assumption.

### C.1.7 Convergence of the drift term

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

$$
\begin{aligned}
&[1.\ell]^{(n_m)}(\vartheta) \\
&\quad = n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla f(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta) \right\rangle \\
&\quad = \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \nabla f(\vartheta),\ \frac{c_h n_m^{\mathfrak{a}+\mathfrak{w}-\mathfrak{h}} \Gamma}{2b^{(n_m)}} \sum_{j\in\left[b^{(n_m)}\right]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right) \right\rangle \\
&\quad = \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ n_m^{\mathfrak{a}+\mathfrak{w}-\mathfrak{h}-1} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) \right\rangle \\
&\quad = \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{\mathfrak{a}-\mathfrak{h}-1} \sum_{i\in[n_m]} \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds\right) \right\rangle
\end{aligned}
\tag{24}
$$

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

$$
\begin{aligned}
&\left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds\ + \mathcal{J}_\star\right) \vartheta \right\rangle \right| \\
&\quad \leq c_h \left\|\Gamma\right\| \left\|\nabla f\right\|_\infty (R_0 + c_0) \left\| \int_0^1 \left[ n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) + \mathcal{J}_\star \right] ds \right\| \\
&\quad \leq c_h \left\|\Gamma\right\| \left\|\nabla f\right\|_\infty (R_0 + c_0) \cdot \Upsilon^{(n_m)},
\end{aligned}
$$

and thus vanishes uniformly on $K_1$.

When $\mathfrak{a} > \mathfrak{h}$, so $c_{\mathsf{d}} = 0$ and hence $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) = 0$ (where $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)$ is the drift term appearing in the definition of the limiting generator $A$ in Eq. (21)), then the drift term will be inactive in the limit. We show this by using the fact that $[1.\ell]^{(n_m)}(\vartheta)$ is a vanishing distance from a sequence that vanishes:

$$
\begin{aligned}
&\left| [1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) \right| \\
&\quad \leq n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds\ + \mathcal{J}_\star\right) \vartheta \right\rangle \right| \\
&\qquad + n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \mathcal{J}_\star \vartheta \right\rangle \right|;
\end{aligned}
$$

and hence vanishes uniformly on $K_1$.

When $\mathfrak{a} = \mathfrak{h}$, then the drift term is active in the limit, and we show that $[1.\ell]^{(n_m)}(\vartheta)$ converges to the drift term from the limiting process $[\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta)$:

$$
\begin{aligned}
&\left| [1.\ell]^{(n_m)}(\vartheta) - [\mathrm{I}.\Gamma\mathcal{J}_\star](\vartheta) \right| \\
&\quad = n_m^{\mathfrak{h}-\mathfrak{a}} \left| \left\langle \frac{c_h \Gamma^\dagger}{2} \nabla f(\vartheta),\ \left(\int_0^1 n_m^{-1} \sum_{i\in[n_m]} \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds\ + \mathcal{J}_\star\right) \vartheta \right\rangle \right|
\end{aligned}
$$

vanishes uniformly on $K_1$.

### C.1.8 Convergence of the diffusion term corresponding to Gaussian noise

$$\begin{aligned}&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\text{II}.\Lambda](\vartheta)\right| \\ &\quad = \left|n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_\xi^{(n_m)},\ \Delta_\xi^{(n_m)}\right\rangle - \frac{c_h}{2c_\beta}\Lambda : \nabla^{\otimes 2}f(\vartheta)\right|\end{aligned}$$

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

$$\begin{aligned}&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\text{II}.\Lambda](\vartheta)\right| \\ &\quad \leq \frac{c_h}{2c_\beta}\left|n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla^{\otimes 2}f(\vartheta)\sqrt{\Lambda}\xi_1,\ \sqrt{\Lambda}\xi_1\right\rangle - \Lambda : \nabla^{\otimes 2}f(\vartheta)\right| \\ &\quad = 0\end{aligned}$$

If $\mathfrak{a} + 2\mathfrak{w} - \mathfrak{h} - \mathfrak{t} < 0$ then the corresponding diffusion term is inactive in the limit, and so $c_{\mathtt{g}} = 0$ and so $[\text{II}.\Lambda](\vartheta) = 0$. In that case we show that $[2.\xi\xi]^{(n_m)}(\vartheta)$ vanishes uniformly.

$$\begin{aligned}&\left|[2.\xi\xi]^{(n_m)}(\vartheta) - [\text{II}.\Lambda](\vartheta)\right| \\ &\quad \leq \frac{c_h}{2c_\beta}n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}}\left|\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \nabla^{\otimes 2}f(\vartheta)\sqrt{\Lambda}\xi_1,\ \sqrt{\Lambda}\xi_1\right\rangle\right| \\ &\quad = \frac{c_h}{2c_\beta}n_m^{\mathfrak{a}+2\mathfrak{w}-\mathfrak{h}-\mathfrak{t}}\left\|\Lambda\right\|_F\left\|\left\|\nabla^{\otimes 2}f\right\|_F\right\|_\infty,\end{aligned}$$

which vanishes uniformly.

### C.1.9 Convergence of the diffusion term corresponding to minibatch noise

$$\begin{aligned}&\left|[2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right| \\ &\quad = \left|n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left\langle \frac{1}{2}\nabla^{\otimes 2}f(\vartheta)\Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta)\right\rangle - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' : \nabla^{\otimes 2}f(\vartheta)\right| \\ &\quad = \frac{1}{2}\left|\left[n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[\left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\right] : \nabla^{\otimes 2}f(\vartheta) - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma' : \nabla^{\otimes 2}f(\vartheta)\right]\right| \\ &\quad \leq \frac{\left\|\nabla^{\otimes 2}f_F\right\|_\infty}{2}\left\|\left[n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[\left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\right] - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma'\right]\right\|_F \\ &\quad \leq \sqrt{d}\frac{\left\|\nabla^{\otimes 2}f_F\right\|_\infty}{2}\left\|\left[n_m^{\mathfrak{a}}\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}}\left[\left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2}\right] - \frac{c_{\mathtt{mb}}}{2}\Gamma\mathcal{I}_\star\Gamma'\right]\right\|\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m^{\mathfrak{a}} \left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2} \\
&= \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2} \Gamma \left( \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right)^{\otimes 2}\right)\Gamma' \\
&\quad + \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2} \Gamma \Bigg( \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j)}\right) \\
&\qquad\qquad \otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j')}\right)\Bigg)\Gamma' \\
&= \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4b^{(n_m)}} \Gamma \left( \frac{1}{n_m} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)\Gamma' \\
&\quad + \frac{c_h^2 n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}}}{4(b^{(n_m)})^2} \Gamma \Bigg( \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j)}\right) \\
&\qquad\qquad \otimes\nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{\vartheta}{n_m^{\mathfrak{w}}};\ X_{I_1^{(n_m)}(j')}\right)\Bigg)\Gamma'
\end{aligned}$$

If the mini-batches are drawn with replacement, then

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right) \otimes \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j')}\right) \\
&= \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m^2} \sum_{i\in[n_m]} \sum_{i'\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) \otimes \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right) \\
&= b^{(n_m)}(b^{(n_m)}-1) \left( \frac{1}{n_m} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2} \\
&= b^{(n_m)}(b^{(n_m)}-1) \left( \frac{1}{n_m} \sum_{i\in[n_m]} \int_0^1 \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds \frac{1}{n_m^{\mathfrak{w}}}\vartheta\right)^{\otimes 2}
\end{aligned}$$

Thus, if $\mathfrak{a} + 2\mathfrak{w} - 2\mathfrak{h} - \mathfrak{b} = 0$, so that $c_{\mathtt{mb}} \neq 0$ and the corresponding term is active in the limit, and the minibtaches are drawn with replacement, then combining the past several equations gives:

$$\begin{aligned}
&\left| [2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta) \right| \\
&\leq \frac{\sqrt{d}\,\|\Gamma\|^2 \left\|\nabla^{\otimes 2} f_F\right\|_\infty}{2} \left\| \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} \left( \frac{1}{n_m} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right) - c_{\mathtt{mb}}\mathcal{I}_\star \right\| \\
&\quad + \frac{\sqrt{d} c_h^2 \,\|\Gamma\|^2 \left\|\nabla^{\otimes 2} f_F\right\|_\infty n_m^{-2\mathfrak{w}}}{8} \left\| \left( \frac{1}{n_m} \sum_{i\in[n_m]} \int_0^1 \nabla^{\otimes 2}\ell\left(\widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) ds\ \vartheta\right)^{\otimes 2} \right\|
\end{aligned}$$

For $\vartheta \in K_1$, and for all $n_m$ large enough that $r_{\mathcal{J},n_m} \geq R_0 + c_0$

$$\begin{aligned}
&n_m^{-2\mathfrak{w}} \left\| \left( \frac{1}{n_m} \sum_{i \in [n_m]} \int_0^1 \nabla^{\otimes 2} \ell \left( \widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) ds\ \vartheta \right)^{\otimes 2} \right\| \\
&\quad = n_m^{-2\mathfrak{w}} \left\| \frac{1}{n_m} \sum_{i \in [n_m]} \int_0^1 \nabla^{\otimes 2} \ell \left( \widehat{\theta}^{(n_m)} + \frac{s}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right) ds\ \vartheta \right\|^2 \\
&\quad \leq \frac{\left(2R_0 + 2c_0\right)^2}{n_m^{2\mathfrak{w}}} \left( \|\mathcal{J}_\star\| + \Upsilon^{(n_m)} \right)^2,
\end{aligned}$$

which vanishes uniformly.

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

$$\begin{aligned}
&\left\| \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right)^{\otimes 2} \right) - c_{\mathtt{mb}} \mathcal{I}_\star \right\| \\
&\quad \leq \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} \left\| \left( \frac{1}{n_m} \sum_{i \in [n_m]} \nabla \ell \left( \widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}} \vartheta;\ X_i \right)^{\otimes 2} \right) - \mathcal{I}_\star \right\| \\
&\qquad + \left| \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} - \frac{c_h^2}{4c_b} \right| \|\mathcal{I}_\star\| \\
&\quad \leq \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} \Upsilon^{(n_m)} + \left| \frac{c_h^2}{4c_b} \frac{c_b n_m^{\mathfrak{b}}}{\lfloor c_b n_m^{\mathfrak{b}} \rfloor} - \frac{c_h^2}{4c_b} \right| \|\mathcal{I}_\star\| .
\end{aligned}$$

And, if $\mathfrak{a} + 2\mathfrak{w} - 2\mathfrak{h} - \mathfrak{b} < 0$ and the mini-batches are drawn with replacement, so that $c_{\mathtt{mb}} = 0$, and the corresponding diffusion term is inactive in the limit and $[\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma']\,(\vartheta) = 0$, then

$$\begin{aligned}
&\left| [2.\ell\ell]^{(n_m)}\,(\vartheta) - [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma']\,(\vartheta) \right| \\
&\quad \leq \left| \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m^{\mathfrak{a}} \left( \Delta_\ell^{(n_m)}(\vartheta) \right)^{\otimes 2} - n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-\mathfrak{b}} \frac{c_h^2}{4c_b} \mathcal{I}_\star \right| + n_m^{\mathfrak{a}+2\mathfrak{w}-2\mathfrak{h}-\mathfrak{b}} \left| \frac{c_h^2}{4c_b} \mathcal{I}_\star \right|
\end{aligned}$$

which vanishes uniformly by the previous arguments.

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

$$\left| [2.\ell\ell]^{(n_m)}\,(\vartheta) \text{ - } [\mathrm{II}.\Gamma\mathcal{I}_\star\Gamma']\,(\vartheta) \right|$$

vanishes uniformly on $K_1$.

If the mini-batches are drawn without replacement.

$$
\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \sum_{j\in\left[b^{(n_m)}\right]} \sum_{j'\in\left[b^{(n_m)}\right]\setminus\{j\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j)}\right) \otimes \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{I_1^{(n_m)}(j')}\right) \\
&= \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)} \sum_{i\in[n_m]} \sum_{i'\in[n_m]\setminus\{i\}} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) \otimes \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right) \\
&= \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)} \sum_{i\in[n_m]} \sum_{i'\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right) \otimes \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_{i'}\right) \\
&\qquad - \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2} \\
&= b^{(n_m)}(b^{(n_m)}-1)\frac{n_m}{n_m-1}\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2} \\
&\qquad - \frac{b^{(n_m)}(b^{(n_m)}-1)}{n_m(n_m-1)} \sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2},
\end{aligned}
$$

and so,

$$
\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} n_m \left(\Delta_\ell^{(n_m)}(\vartheta)\right)^{\otimes 2} \\
&\quad = \frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1}\Gamma\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right)\Gamma' \\
&\qquad\quad + \frac{c_h^2}{4(b^{(n_m)})^2}\Gamma\left(b^{(n_m)}(b^{(n_m)}-1)\frac{n_m}{n_m-1}\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)\right)^{\otimes 2}\right)\Gamma'
\end{aligned}
$$

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

$$
\begin{aligned}
&\left\|\frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1}\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right) - c_{\mathtt{mb}}\mathcal{I}_\star\right\| \\
&\quad \leq \frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1}\left\|\left(\frac{1}{n_m}\sum_{i\in[n_m]} \nabla\ell\left(\widehat{\theta}^{(n_m)} + \frac{1}{n_m^{\mathfrak{w}}}\vartheta;\ X_i\right)^{\otimes 2}\right) - \mathcal{I}_\star\right\| \\
&\qquad + \left|\frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1} - c_{\mathtt{mb}}\right|\left\|\mathcal{I}_\star\right\| \\
&\quad \leq \frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1}\Upsilon^{(n_m)} + \left|\frac{c_h^2}{4b^{(n_m)}}\frac{n_m - b^{(n_m)}}{n_m-1} - c_{\mathtt{mb}}\right|\left\|\mathcal{I}_\star\right\|,
\end{aligned}
$$

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

$$
\left|[2.\ell\ell]^{(n_m)}(\vartheta) \text{ - } [\text{II}.\Gamma\mathcal{I}_\star\Gamma'](\vartheta)\right|
$$

vanishes uniformly on $K_1$.

### C.1.10 Convergence of the Remainder Term

$$\begin{aligned}
&\left|[3.R]^{(n_m)}(\vartheta)\right| \\
&\quad = n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left[\frac{1}{6}\left[\nabla^{\otimes 3} f(\vartheta + S\Delta^{(n_m)}(\vartheta))\right]\left(\Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta), \Delta^{(n_m)}(\vartheta)\right)\right] \\
&\quad \leq \frac{n_m^{\mathfrak{a}}}{6}\left\|\nabla^{\otimes 3} f\right\|_\infty \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta^{(n_m)}(\vartheta)\right\|^3 \\
&\quad \leq \frac{27 n_m^{\mathfrak{a}}}{6}\left\|\nabla^{\otimes 3} f\right\|_\infty \left(\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta_\xi^{(n_m)}\right\|^3 + \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\|^3 + \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta_\ell^{(n_m)}(\vartheta)\right\|^3\right),
\end{aligned}$$

Now

$$\begin{aligned}
\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta_\xi^{(n_m)}\right\|^3 &\leq \left(\frac{c_h}{2c_\beta} n_m^{-\mathfrak{h}-\mathfrak{t}+2\mathfrak{w}}\left\|\Lambda\right\|\right)^{3/2} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\xi_1\right\|^3 \\
&= n_m^{-3/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w})}\left(\frac{c_h}{2c_\beta}\left\|\Lambda\right\|\right)^{3/2} 2^{3/2}\frac{\mathit{\Gamma}\left(\frac{d+3}{2}\right)}{\mathit{\Gamma}\left(\frac{d}{2}\right)},
\end{aligned}$$

where $\mathit{\Gamma}$ is the gamma function. Note that $\alpha - 3/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}) \leq -1/2\,(\mathfrak{h}+\mathfrak{t}-2\mathfrak{w}) \leq -\mathfrak{a}/2 < 0$

Second,

$$\left\|\Delta_{\pi^{(0)}}^{(n_m)}(\vartheta)\right\|^3 \leq \left(\frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}-1}\left\|\Gamma\right\|}{2}\right)^3\left(\left\|\nabla \log \pi^{(0)}(\theta_\star)\right\| + L_0\left\|\widehat{\theta}^{(n_m)} - \theta_\star\right\| + L_0\frac{2R_0 + 2c_0}{n_m^{\mathfrak{w}}}\right)^3.$$

Note that $\mathfrak{a} - 3\mathfrak{h} + 3\mathfrak{w} - 3 \leq -2\mathfrak{h} - 3(1-\mathfrak{w}) < 0$.

Third,

$$\begin{aligned}
&\mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\|\Delta_\ell^{(n_m)}(\vartheta)\right\|^3 \\
&\leq \left(\frac{c_h n_m^{-\mathfrak{h}+\mathfrak{w}}\left\|\Gamma\right\|}{2}\right)^3\left(n_m^{1/p_2} + n_m^{1/p_3}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{w}}\right)^3 \\
&\leq \left(\frac{c_h\left\|\Gamma\right\|}{2}\right)^3\left(n_m^{1/p_2-\mathfrak{h}+\mathfrak{w}} + n_m^{1/p_3-\mathfrak{h}+\mathfrak{w}}\Upsilon^{(n_m)} + n_m^{1/p_3-\mathfrak{h}}\right)^3
\end{aligned}$$

Therefore, $\left|[3.R]^{(n_m)}(\vartheta)\right|$ vanishes uniformly.

# D Proof of Corollary 2

*Proof of Corollary 2.* 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 $\nu^{(n_{m_k})}$ has a sub-sub-subsequence $\nu^{(n_{m_{k_j}})}$ converging weakly to $\nu$ almost surely. Since weak convergence of probability measures is metrizable, then applying Lemma 1 yields the desired result.

By the second part of Theorem 2, every sub-subsequence of $\left(T^{(n_m)}\right)_{m\in\mathbb{N}}$, $\left(T^{(n_{m_k})}\right)_{k\in\mathbb{N}}$, has a further sub-sub-subsequence, $\left(T^{(n_{m_{k_j}})}\right)_{j\in\mathbb{N}}$, such that with probability 1, $T_t^{(n_{m_{k_j}})} \xrightarrow{\text{s}} T_t$ on $\overline{C}(\mathbb{R}^d)$ for all $t > 0$.

Applying Ethier and Kurtz [2009, Part 4, Theorem 9.10], we have that every weak limit of $\left\{\nu^{(n_{m_{k_j}})}\right\}_{j\in\mathbb{N}}$ is stationary for $T$. As a consequence of the assumption that the spectrum of $\Gamma\mathcal{J}(\theta_\star)$ is a subset of $\{x \in \mathbb{C} \text{ s.t. } \Re(x) > 0\}$, $T$ has a unique stationary distribution (see, for example, Karatzas and Shreve [2014]), $\nu = \mathrm{N}(0, Q_\infty)$. Thus every weak limit of $\left\{\nu^{(n_{m_{k_j}})}\right\}_{j\in\mathbb{N}}$ must be $\nu$.

Since $\left\{\nu^{(n_m)}\right\}_{m\in\mathbb{N}}$ is assumed to be tight, then all of its sub-subsequences have a weakly converging sub-sub-subsequence, concluding the proof.

# E Sufficient conditions for Assumption 4 and Assumption 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 4** (Sufficient conditions for Assumption 4)**.** *Each of the following imply Assumption 4.*

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

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

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

$$\mathbb{E}\left[ \sup_{\theta \in B_{\delta_1}(\theta_\star)} \frac{\left\| \nabla^{\otimes 2} \ell(\theta; X_1) - \nabla^{\otimes 2} \ell(\theta_\star; X_1) \right\|}{\|\theta - \theta_\star\|} \right] < \infty,$$

*Proof of Proposition 4.*

a) Let $r_{\mathcal{J},n} = \delta_1 n^{\mathfrak{w}/2}/2$. Then $B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right) \subseteq B\left(\widehat{\theta}^{(n)}, \delta_1/2\right)$.

Given that $\widehat{\theta}^{(n)} \xrightarrow{\text{p}} \theta_\star$, any subsequence of indices $n_m$ has a further sub-subsequence of indices $n_{m_k}$ where both $\widehat{\theta}^{(n_{m_k})} \to \theta_\star$ and

$$\sup_{\theta \in B_{\delta_1}(\theta_\star)} \left\| \frac{1}{n_{m_k}} \sum_{i \in \left[n_{m_k}\right]} \nabla^{\otimes 2} \ell(\theta; X_i) + \mathcal{J}(\theta) \right\| \to 0 \text{ a.s.}$$

Then there is a $k_0$ such that if $k \geq k_0$ then $\left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| \leq \delta_1/2$. Therefore if $k \geq k_0$ then $B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right) \subseteq B\left(\theta_\star, \delta_1\right)$.

Thus, for $k \geq k_0$,

$$\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r/n_{m_k}^{\mathfrak{w}}\right)} \|\mathcal{J}(\theta) - \mathcal{J}(\theta_\star)\| \\
&\quad \leq \sup_{\theta \in B(\theta_\star, \delta_1)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1/n_{m_k}^{\mathfrak{w}/2}\right)} \|\mathcal{J}(\theta) - \mathcal{J}(\theta_\star)\| \\
&\quad \leq \sup_{\theta \in B(\theta_\star, \delta_1)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}(\theta) \right\| + \sup_{\theta \in B\left(\theta_\star, \left\|\widehat{\theta}^{(n_{m_k})} - \theta_\star\right\| + \delta_1/n_{m_k}^{\mathfrak{w}/2}\right)} \|\mathcal{J}(\theta) - \mathcal{J}(\theta_\star)\| \\
&\quad \overset{\text{a.s.}}{\to} 0.
\end{aligned}$$

Therefore, every subsequence of $S_n = \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\|$ has a further sub-subsequence converging almost surely to 0, and hence $S_n$ converges in probability to 0.

b) Equicontinuity implies there is a function $\rho_{\mathcal{J}_\star} : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{t \to 0} \rho_{\mathcal{J}_\star}(t) = 0$, and

$$\sup_{x \in \mathcal{X}} \sup_{\vartheta \in B_\delta(\theta_\star)} \left\| \nabla^{\otimes 2} \ell(\vartheta; x) - \nabla^{\otimes 2} \ell(\theta_\star; x) \right\| \leq \rho_{\mathcal{J}_\star}(\delta).$$

Let $r_{\mathcal{J},n} = n^{\mathfrak{w}/2}$. Then

$$\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \widehat{\mathcal{J}}^{(n)}(\theta_\star) \right\| + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \sup_{\theta \in B\left(\theta_\star, \left\|\widehat{\theta}^{(n)} - \theta_\star\right\| + n^{-\mathfrak{w}/2}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \widehat{\mathcal{J}}^{(n)}(\theta_\star) \right\| + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \leq \rho_{\mathcal{J}_\star}\left( \left\| \widehat{\theta}^{(n)} - \theta_\star \right\| + n^{-\mathfrak{w}/2} \right) + \left\| \widehat{\mathcal{J}}^{(n)}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\quad \xrightarrow{\mathrm{p}} 0.
\end{aligned}$$

In the last step we used that the first term vanishes in probability because $\widehat{\theta}^{(n)} \xrightarrow{\mathrm{p}} \theta_\star$, and the second term vanishes in probability by the weak law of large numbers.

c) Let

$$\begin{aligned}
Q_n &= \frac{1}{n} \sum_{i \in [n]} \left[ \sup_{\theta \in B_{\delta_1}(\theta_\star)} \frac{\left\| \nabla^{\otimes 2} \ell(\theta; X_i) - \nabla^{\otimes 2} \ell(\theta_\star; X_i) \right\|}{\|\theta - \theta_\star\|} \right], \text{ and} \\
q &= \mathbb{E}\left[ \sup_{\theta \in B_{\delta_1}(\theta_\star)} \frac{\left\| \nabla^{\otimes 2} \ell(\theta; X_1) - \nabla^{\otimes 2} \ell(\theta_\star; X_1) \right\|}{\|\theta - \theta_\star\|} \right].
\end{aligned}$$

By the weak law of large numbers, $Q_n \xrightarrow{\mathrm{p}} q$ and $\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \xrightarrow{\mathrm{p}} \mathcal{J}(\theta_\star)$. Let $r_{\mathcal{J},n} = \delta_1 n^{\mathfrak{w}/2}/2$. As in part a), given that $\widehat{\theta}^{(n)} \xrightarrow{\mathrm{p}} \theta_\star$, any subsequence of indices $n_m$ has a further sub-subsequence of indices $n_{m_k}$ where both $\widehat{\theta}^{(n_{m_k})} \to \theta_\star$, $Q_{n_{m_k}} \to q$, and $\widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \to \mathcal{J}(\theta_\star)$ almost surely. Then there is a $k_0$ such that if $k \geq k_0$ then $\left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| \leq \delta_1/2$. Therefore if $k \geq k_0$ then $B\left( \widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}} \right) \subseteq B(\theta_\star, \delta_1)$. Thus, for $k \geq k_0$,

$$\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, r_{\mathcal{J},n}/n_{m_k}^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \mathcal{J}_\star \right\| \\
&\leq \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)} \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \right\| \\
&\leq \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\qquad + \left( \left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2 \right) \sup_{\theta \in B\left(\widehat{\theta}^{(n_{m_k})}, \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2\right)} \frac{\left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \right\|}{\|\theta - \theta_\star\|} \\
&\leq \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\qquad + \left( \left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2 \right) \sup_{\theta \in B(\theta_\star, \delta_1)} \frac{\left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta) - \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) \right\|}{\|\theta - \theta_\star\|} \\
&\leq \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| \\
&\qquad + \left( \left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2 \right) \sup_{\theta \in B(\theta_\star, \delta_1)} \frac{1}{n_{m_k}} \sum_{i \in \left[n_{m_k}\right]} \left[ \frac{\left\| \nabla^{\otimes 2} \ell(\theta; X_i) - \nabla^{\otimes 2} \ell(\theta_\star; X_i) \right\|}{\|\theta - \theta_\star\|} \right] \\
&\leq \left\| \widehat{\mathcal{J}}^{(n_{m_k})}(\theta_\star) - \mathcal{J}(\theta_\star) \right\| + \left( \left\| \widehat{\theta}^{(n_{m_k})} - \theta_\star \right\| + \delta_1 n_{m_k}^{-\mathfrak{w}/2}/2 \right) Q_{n_{m_k}} \xrightarrow{\text{a.s.}} 0
\end{aligned}$$

Therefore, every subsequence of $S_n = \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{J},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{J}}^{(n)}(\theta) - \mathcal{J}(\theta_\star) \right\|$ has a further sub-subsequence converging almost surely to 0, and hence $S_n$ converges in probability to 0.

**Proposition 5** (Sufficient conditions for Assumption 5)**.** *Each of the following imply Assumption 5.*

*a) there exists a $\delta_2 > 0$ with $\sup_{\theta \in B_{\delta_2}(\theta_\star)} \left\| \frac{1}{n} \sum_{i\in[n]} \nabla \ell(\theta; X_i)^{\otimes 2} - \mathcal{I}(\theta) \right\| \xrightarrow{\mathrm{p}} 0$ and $\mathcal{I}$ is continuous at $\theta_\star$,*

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

*c) $\mathbb{E}\left[ \left\| \nabla^{\otimes 2} \ell(\cdot; X_1) \right\|_\infty^2 \right] < \infty$,*

*Proof of Proposition 5.*

a), b) The proofs are the same as for Proposition 4 a), b).

c) Let $Q_n = \frac{1}{n}\sum_{i\in[n]} \left\| \nabla^{\otimes 2} \ell(\cdot; X_i) \right\|_\infty^2$, $q = \mathbb{E} \left\| \nabla^{\otimes 2} \ell(\cdot; X_1) \right\|_\infty^2$, and let $r_{\mathcal{I},n} = n^{\mathfrak{w}/2}$. By the weak law of large numbers, $Q_n \xrightarrow{\mathrm{p}} q$, and $\widehat{\mathcal{I}}^{(n)}(\theta_\star) \xrightarrow{\mathrm{p}} \mathcal{I}(\theta_\star)$. Starting with

$$\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}_\star \right\| \le \left\| \widehat{\mathcal{I}}^{(n)}(\theta_\star) - \mathcal{I}(\theta_\star) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left\| \widehat{\mathcal{I}}^{(n)}(\theta) - \widehat{\mathcal{I}}^{(n)}(\theta_\star) \right\|,$$

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

$$\begin{aligned}
&\left\| \widehat{\mathcal{I}}^{(n)}(\theta) - \widehat{\mathcal{I}}^{(n)}(\theta_\star) \right\| \\
&\quad \le \frac{1}{n} \sum_{i\in[n]} \left\| \left( \nabla \ell\left(\theta_\star;\ X_i\right) + \int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right)^{\otimes 2} - \nabla \ell\left(\theta_\star;\ X_i\right)^{\otimes 2} \right\| \\
&\quad \le \frac{2}{n} \sum_{i\in[n]} \left\| \nabla \ell\left(\theta_\star;\ X_i\right) \right\| \left\| \int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right\| \\
&\qquad + \frac{1}{n} \sum_{i\in[n]} \left\| \left( \int_0^1 \nabla^{\otimes 2} \ell\left(\theta_\star + s(\theta - \theta_\star);\ X_i\right) ds\ (\theta - \theta_\star) \right)^{\otimes 2} \right\| \\
&\quad \le \frac{2}{n} \sum_{i\in[n]} \left\| \nabla \ell\left(\theta_\star;\ X_i\right) \right\| \left\| \nabla^{\otimes 2} \ell(\cdot; X_i) \right\|_\infty \left\| \theta - \theta_\star \right\| + \frac{1}{n} \sum_{i\in[n]} \left\| \nabla^{\otimes 2} \ell(\cdot; X_i) \right\|_\infty^2 \left\| \theta - \theta_\star \right\|^2 \\
&\quad \le 2 \left\| \theta - \theta_\star \right\| \sqrt{\frac{1}{n} \sum_{i\in[n]} \left\| \nabla \ell\left(\theta_\star;\ X_i\right) \right\|^2} \sqrt{\frac{1}{n} \sum_{i\in[n]} L(X_i)^2} + \left\| \theta - \theta_\star \right\|^2 Q_n \\
&\quad \le 2 \left\| \theta - \theta_\star \right\| \sqrt{\operatorname{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))} \sqrt{Q_n} + \left\| \theta - \theta_\star \right\|^2 Q_n,
\end{aligned}$$

Plugging this back in,

$$\begin{aligned}
&\sup_{\theta \in B\left(\widehat{\theta}^{(n)}, r_{\mathcal{I},n}/n^{\mathfrak{w}}\right)} \left\| \widehat{\mathcal{I}}^{(n)}(\theta) - \mathcal{I}_\star \right\| \\
&\quad \le \left\| \widehat{\mathcal{I}}^{(n)}(\theta_\star) - \mathcal{I}(\theta_\star) \right\| + \sup_{\theta \in B\left(\widehat{\theta}^{(n)}, n^{-\mathfrak{w}/2}\right)} \left( 2 \left\| \theta - \theta_\star \right\| \sqrt{\operatorname{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))} \sqrt{Q_n} + \left\| \theta - \theta_\star \right\|^2 Q_n \right) \\
&\quad \le \left\| \widehat{\mathcal{I}}^{(n)}(\theta_\star) - \mathcal{I}(\theta_\star) \right\| + 2\left( \left\| \widehat{\theta}^{(n)} - \theta_\star \right\| + n^{-\mathfrak{w}/2} \right) \sqrt{\operatorname{Tr}(\widehat{\mathcal{I}}^{(n)}(\theta_\star))} \sqrt{Q_n} + \left( \left\| \widehat{\theta}^{(n)} - \theta_\star \right\| + n^{-\mathfrak{w}/2} \right)^2 Q_n \\
&\quad \xrightarrow{\mathrm{p}} 0.
\end{aligned}$$

# F Proof of Proposition 1

Recall that

$$d\vartheta_t = -\frac{1}{2}B\vartheta_t\,dt + \sqrt{A}\,dW_t, \tag{25}$$

which implies

$$\vartheta_t = \exp(-B/2)\vartheta_0 + \int_0^t \exp(-B(t-s)/2)A^{1/2}dW_s. \tag{26}$$

Assuming stationarity, $\vartheta_t \sim \mathcal{N}(0, Q_\infty)$ where $Q_\infty = \int_0^\infty \exp(-Bs/2)A\exp(-Bs/2)ds$, we have

$$\mathrm{Cov}\left(\int_0^t \vartheta_s\,ds\right) = \mathbb{E}\left(\int_0^t\int_0^t \vartheta_s\vartheta_r^T dsdr\right) = \int_0^t\int_0^s \mathbb{E}(\vartheta_s\vartheta_r^T)drds + \int_0^t\int_0^r \mathbb{E}(\vartheta_s\vartheta_r^T)dsdr. \tag{27}$$

We focus on the first term since the second term can be written similarly:

$$\begin{aligned}
\int_0^t\int_0^s \mathbb{E}(\vartheta_s\vartheta_r^T)drds &= \int_0^t\int_0^s \mathbb{E}\left[\left(\exp(-B(s-r)/2)\vartheta_r + \int_r^s \exp(-B(s-u)/2)A^{1/2}dW_u\right)\vartheta_r^T\right]drds \\
&= \int_0^t\int_0^s \exp(-B(s-r)/2)\mathbb{E}(\vartheta_r\vartheta_r^T)drds \\
&= \int_0^t\int_0^s \exp(-B(s-r)/2)Q_\infty drds \\
&= \int_0^t -2B^{-1}(\exp(-Bs/2)-1)Q_\infty ds \\
&= \left[4B^{-2}(\exp(-Bt/2)-1) + 2tB^{-1}\right]Q_\infty.
\end{aligned} \tag{28}$$

We can write $\int_0^t\int_0^r \mathbb{E}(\vartheta_s\vartheta_r^T)dsdr$ similarly and combine the two results

$$\begin{aligned}
\mathrm{Cov}\left(\bar{\vartheta}_t\right) = \frac{1}{t^2}\,\mathrm{Cov}(\int_0^t \vartheta_s\,ds) &= \frac{1}{t^2}\left[\int_0^t\int_0^s \mathbb{E}(\vartheta_s\vartheta_r^T)drds + \int_0^t\int_0^r \mathbb{E}(\vartheta_s\vartheta_r^T)dsdr\right] \\
&= \frac{4}{t}\,\mathrm{Sym}\left(B^{-1}Q_\infty\right) - \frac{8}{t^2}\,\mathrm{Sym}\left(B^{-2}\left\{I - e^{-tB/2}\right\}Q_\infty\right),
\end{aligned} \tag{29}$$

which verifies Eq. (11).

Using Taylor's theorem and the assumption that $-B$ is Hurwitz, we obtain

$$e^{-tB/2} - I = \sum_{k=1}^{\ell}\frac{1}{k!}\left(\frac{-tB}{2}\right)^k + R_\ell(t), \tag{30}$$

where $\|R_\ell(t)\| \le \frac{\|tB/2\|^{\ell+1}}{(\ell+1)!}$. Taking $\ell = 3$ yields

$$\begin{aligned}
&\frac{4}{t}B^{-1}AB^{-\top} - \frac{8}{t^2}\,\mathrm{Sym}\left(B^{-2}\left\{I - e^{-tB/2}\right\}Q_\infty\right) \\
&= \frac{4}{t}B^{-1}AB^{-\top} + \frac{8}{t^2}\left\{-\left(\frac{t}{2}\right)\mathrm{Sym}\left(B^{-1}Q_\infty\right) + \frac{1}{2}\left(\frac{t}{2}\right)^2 Q_\infty - \frac{1}{6}\left(\frac{t}{2}\right)^3\mathrm{Sym}\left(BQ_\infty\right) + \mathrm{Sym}\left(B^{-2}R_3(t)Q_\infty\right)\right\} \\
&= Q_\infty - \frac{t}{6}A + \tilde{R}_3(t),
\end{aligned} \tag{31}$$

where $\left\|\tilde{R}_3(t)\right\| \le \frac{t^2}{48}\|B\|^4\left\|B^{-2}Q_\infty\right\|$, and we have used that $\mathrm{Sym}(B^{-1}Q_\infty) = B^{-1}AB^{-\top}$ and $\mathrm{Sym}\left(BQ_\infty\right) = A$, and that $B^{-1}$ and $R_\ell(t)$ commute.

For any $t > 0$, we have

$$\begin{aligned}
\left\|\frac{8}{t^2}\,\mathrm{Sym}\left(B^{-2}\left\{I - e^{-tB/2}\right\}Q_\infty\right)\right\| &= \left\|\frac{8}{t^2}\,\mathrm{Sym}\left(\left\{I - e^{-tB/2}\right\}B^{-2}Q_\infty\right)\right\| \\
&\le \frac{8}{t^2}\left\|\mathrm{Sym}\left(B^{-2}Q_\infty\right)\right\| \le \frac{8}{t^2}\left\|B^{-2}Q_\infty\right\|,
\end{aligned} \tag{32}$$

which is small when $t \gg 3 \left\| B^{-2} Q_\infty \right\|^{1/2}$.

## G Proof of Corollary 3

*Proof.* For Eq. (13), we have

$$\begin{aligned}
\bar{Q}_k^{(n)} &= \operatorname{Cov}\left(\bar{\theta}^{(n)}_{\lfloor mn/b^{(n)} \rfloor}\right) \approx \frac{1}{(w^{(n)})^2} \operatorname{Cov}\left(\bar{\vartheta}_{mn/(b^{(n)}\alpha^{(n)})}\right) \\
&= \frac{4}{m} \frac{\alpha^{(n)} b^{(n)}}{n(w^{(n)})^2} \operatorname{Sym}\left(\left\{c_h \Gamma \mathcal{J}_\star\right\}^{-1} Q_\infty\right) \\
&\quad - \frac{8}{m^2} \frac{(\alpha^{(n)} b^{(n)})^2}{(n w^{(n)})^2} \operatorname{Sym}\left(\left\{c_h \Gamma \mathcal{J}_\star\right\}^{-2} \left\{I - \exp\left[-\frac{c_h m n}{2 b^{(n)} \alpha^{(n)}} \Gamma \mathcal{J}_\star\right] Q_\infty\right\}\right).
\end{aligned}$$

Now, given $\mathfrak{b} + \mathfrak{h} \leq \mathfrak{t}$,

$$\begin{aligned}
\lim_{n \to \infty} n \bar{Q}_k^{(n)} = \frac{4 c_b}{m} \operatorname{Sym}\left(\left\{c_h \Gamma \mathcal{J}_\star\right\}^{-1} Q_\infty\right) \\
- \mathbb{I}_{[\mathfrak{b} + \mathfrak{h} = 1]} \frac{8 c_b^2}{m^2} \operatorname{Sym}\left(\left[c_h \Gamma \mathcal{J}_\star\right]^{-2} \left[I - e^{-\frac{c_h m}{2 c_b} \Gamma \mathcal{J}_\star}\right] Q_\infty\right) \Bigg\}
\end{aligned}$$

The rest follows by combining this with Proposition 1 and the simplifications following it, and by noting that since $\mathfrak{h} + \mathfrak{b} \leq 1$ and $\mathfrak{h} > 0$ we must have $\mathfrak{b} < 1$, and hence $\overline{c_b} = 1$.

## H 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 $\nabla \ell\left(\theta;\ X_I\right)$ is replaced by $\nabla \ell\left(\theta;\ X_I\right) - \nabla \ell\left(\theta_\star;\ X_I\right)$. 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 $[2.\ell\ell]^{(n_m)}(\vartheta)$ from the proof of Theorem 1 in Appendix C is vanishing under any scaling limit where the drift term $[1.\ell]$ does not vanish.

$$
\begin{aligned}
&\underbrace{n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left\langle \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) \Delta_\ell^{(n_m)}(\vartheta),\ \Delta_\ell^{(n_m)}(\vartheta) \right\rangle}_{[2.\ell\ell]^{(n_m)}\,(\vartheta)} \\
&\quad = n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \frac{1}{2} \nabla^{\otimes 2} f(\vartheta) : \left( \Delta_\ell^{(n_m)}(\vartheta) \right)^{\otimes 2} \\
&\quad = n_m^{\mathfrak{a}} \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\vartheta) : \left( \frac{h w^{(n)} \Gamma}{2 b^{(n)}} \sum_{j \in \left[b^{(n)}\right]} \left( \nabla \ell \left( \widehat{\theta}^{(n)} + (w^{(n)})^{-1} \vartheta;\ X_{I_1^{(n)}(j)} \right) - \nabla \ell \left( \widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)} \right) \right) \right)^{\otimes 2} \\
&\quad = \frac{c_h^2}{c_b^2} n_m^{\mathfrak{a} - 2\mathfrak{h} + 2\mathfrak{w} - 2\mathfrak{b}} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\vartheta) \Gamma^\top \\
&\qquad\quad : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left( \sum_{j \in \left[b^{(n)}\right]} \left( \nabla \ell \left( \widehat{\theta}^{(n)} + (w^{(n)})^{-1} \vartheta;\ X_{I_1^{(n)}(j)} \right) - \nabla \ell \left( \widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)} \right) \right) \right)^{\otimes 2} \\
&\quad \approx \frac{c_h^2}{c_b^2} n_m^{\mathfrak{a} - 2\mathfrak{h} + 2\mathfrak{w} - 2\mathfrak{b}} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\vartheta) \Gamma^\top : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left( \sum_{j \in \left[b^{(n)}\right]} \nabla^{\otimes 2} \ell \left( \widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)} \right) (w^{(n)})^{-1} \vartheta \right)^{\otimes 2} \\
&\quad = \frac{c_h^2}{c_b^2} n_m^{\mathfrak{a} - 2\mathfrak{h} - 2\mathfrak{b}} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\vartheta) \Gamma^\top : \mathbb{E}^{\mathbf{X}^{(\mathbb{N})}} \left( \sum_{j \in \left[b^{(n)}\right]} \nabla^{\otimes 2} \ell \left( \widehat{\theta}^{(n)};\ X_{I_1^{(n)}(j)} \right) \vartheta \right)^{\otimes 2} \\
&\quad \approx n_m^{\mathfrak{a} - 2\mathfrak{h} - 2\mathfrak{b}} \frac{1}{2} \Gamma \nabla^{\otimes 2} f(\vartheta) \Gamma^\top : \left[ b^{(n)} (b^{(n)} - 1) \mathcal{J}_\star \vartheta \vartheta^\top \mathcal{J}_\star + b^{(n)} K(\theta_\star; \vartheta) \right]
\end{aligned}
$$

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

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

# I Sketch Proof for constrained parameter spaces

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

(i) $\mathcal{P}$ is *faithful* to $\Theta$, meaning that if $\mathrm{Conv}(\theta, \theta + \Delta_{\pi^{(0)}} + \Delta_\ell + \Delta_\xi) \subset \Theta$ then

$$\mathcal{P}(\theta, \Delta_{\pi^{(0)}}, \Delta_\ell, \Delta_\xi) = \theta + \Delta_{\pi^{(0)}} + \Delta_\ell + \Delta_\xi, \tag{33}$$

where $\mathrm{Conv}(\theta_1, \theta_2)$ 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 $(\theta, \Delta_{\pi^{(0)}}, \Delta_\ell, \Delta_\xi) \in \Theta \times (\mathbb{R}^d)^3$

$$\|\mathcal{P}(\theta, \Delta_{\pi^{(0)}}, \Delta_\ell, \Delta_\xi) - \theta\| \leq c_{\mathcal{P}} \left( \|\Delta_{\pi^{(0)}}\| + \|\Delta_\ell\| + \|\Delta_\xi\| \right). \tag{34}$$

We will consider the iterative algorithm on $\Theta$ given by

$$\theta_{k+1}^{(n)} = \mathcal{P}\left( \theta_k^{(n)},\ \frac{h\Gamma}{2n} \nabla \log \pi^{(0)} \left( \theta_k^{(n)} \right),\ \frac{h\Gamma}{2} \frac{1}{b} \sum_{j \in [b]} \nabla \ell \left( \theta_k^{(n)};\ X_{I_k^{(n)}(j)} \right),\ \sqrt{h\beta^{-1}\Lambda}\ \xi_k \right). \tag{35}$$

The key idea is that, if $\theta_\star \in \mathrm{interior}(\Theta)$, there is a $r > 0$ with $\theta_\star \in B(\theta_\star, r) \subset \mathrm{interior}(\Theta)$, and for any compactly supported test function $f$ and compact extension of its support, $K_1$, for sufficiently large sample sizes $n$, $K_1 \subseteq B(0, w^{(n)} r)$. In the proof of the $\Theta = \mathbb{R}^d$ case we found that, along sub-sequences $(n_{m_k})$, the increments from the log-likelihood and from the prior vanish uniformly within a sufficiently large extension of the support of $f$. Combining this with faithfulness of $\mathcal{P}$ (defined in Appendix A.3) and an application of the Lebesgue dominated convergence theorem to handle truncation of the Gaussian increments shows that the $A_{n_m} f \to Af$ uniformly within the extension of the support of $f$ when $\Theta \neq \mathbb{R}^d$. Moreover, the local property of the boundary condition (defined in Appendix A.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.

# J 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

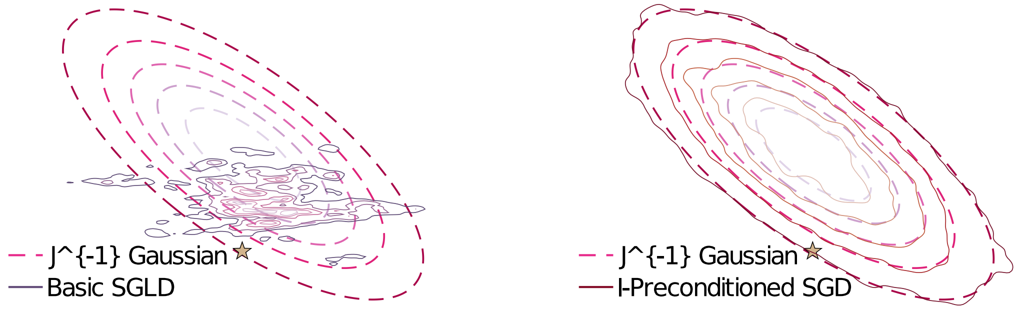
$$\begin{aligned}\mathrm{SpecGap}_\zeta := \inf \Bigg\{ \frac{\pi[f^2] - \langle f,\ Pf\rangle_{L^2(\pi)}}{\pi[f^2] - \zeta/2} \ \Bigg| \\ f \in L^2(\pi),\ \pi f = 0,\ \pi[f^2] > \zeta,\ \|f\|_{L^2(\pi)} < \infty \Bigg\}.\end{aligned} \tag{36}$$

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.

# K Additional Details for Experimental Results

| | Experiment 1 | Experiment 2 | Experiment 3 |
|---|---|---|---|
| true distribution | $\mathrm{N}_{10}\left(0, \frac{1}{2}I + \frac{1}{2}\mathbf{11}'\right)$ | unknown | unknown |
| log-likelihood $\ell(\cdot;\theta)$ | $\sum_{i=1}^{10} \frac{(x_i-\theta_i)^2}{\sqrt{i}}$ | $yx^\top\theta - \log(1+e^{x^\top\theta})$ | $yx^\top\theta - \exp(x^\top\theta)$ |
| log-prior $\log \pi^{(0)}(\theta)$ | 0 | 0 | 0 |
| sample size $n$ | 1000 | 1000000 | 150000 |
| batch size $b$ | 1 | 1000 | 250 |
| number of steps $k$ | $10000n/b$ | $1000n/b$ | $1000n/b$ |
| step size (SGD) $h$ | $4b/n$ | $4b/n$ | $4b/n$ |
| step size (SGLD) $h$ | $2b/n$ | $b/n$ | $2b/n$ |
| inv. temp. (SGLD) $\beta$ | 2 | 1 | 2 |

Table 3: 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.



(a) SGLD without Preconditioning

(b) SGD Preconditioned by $\mathcal{I}_\star$

Figure 4: Joint results of experiment 2: Parameters 1 and 4

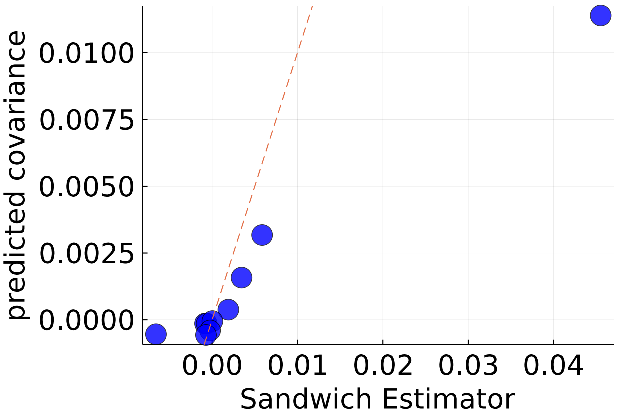


Figure 5: Further result for experiment 2 comparing the scaled sandwich covariance estimator Eq. (13) to the predicted values variance-covariance matrix based upon Eq. (14) for iterate averages when $\mathfrak{h} + \mathfrak{b} = 1$. We see that the higher order correction is material in this case, as expected based upon the theoretical results.