Check for updates

J. Appl. Probab. 1–17 (2026)
*doi:*10.1017/jpr.2026.10109

NO FREE LUNCH FOR STOCHASTIC GRADIENT LANGEVIN DYNAMICS

NATESH PILLAI,* ** *Harvard University*
AARON SMITH ,*** **** *University of Ottawa*
AZEEM ZAMAN,* ***** *Harvard University*

Abstract

As sample sizes grow, scalability has become a central concern in the development of Markov chain Monte Carlo (MCMC) methods. One general approach to this problem, exemplified by the popular stochastic gradient Langevin dynamics (SGLD) algorithm, is to use a small random subsample of the data at every time step. This paper shows that this approach often fails: while decreasing the sample size increases the speed of each MCMC step, for typical datasets this is balanced by a matching decrease in accuracy. This result complements recent work on lower bounds for subsampling MCMC by focusing on actual errors for SGLD, including non-reversible algorithms, and by giving assumptions that are comparatively easy to verify.

Keywords: Stochastic gradient Langevin dynamics; Markov chain Monte Carlo; subsampling; total variation distance; exponential family

2020 Mathematics Subject Classification: Primary 65C05
Secondary 60J22

1. Introduction

It is well known that MCMC algorithms often perform poorly when applied to large datasets. The simplest cause of this poor scaling is the fact that many popular MCMC algorithms require a computation involving every data point at every time step of the algorithm. This suggests that the per-step computational cost of MCMC scales at least linearly in the size n of the data set. It is natural to try to avoid this problem by looking only at a subsample of the data at every time step. The goal of this paper (together with the longer companion paper [10]) is to sketch limits on what performance gains can be achieved with simple subsample-based algorithms.

The main results of the current paper, Theorem 1 and its variants Theorems 2 and 3, give a strong sense in which improvement is impossible. Informally, they show that, under certain conditions, the error of a popular subsampling algorithm cannot be much smaller than the full-sample algorithm it was based on.

Received 18 August 2025 UTC; accepted 15 May 2026 UTC.
* Postal address: Department of Statistics, Harvard University, 33 Oxford Street, Cambridge, MA, USA.
** Email address: pillai@fas.harvard.edu
*** Postal address: Department of Mathematics and Statistics, University of Ottawa, 150 Louis Pasteur, Ottawa, ON, Canada. Email address: asmi28@uottaw.aca
**** Tutte Institute for Mathematics and Computing, 1929 Ogilvie Road, Ottawa, ON, Canada
***** Email address: azeem.zaman@gmail.com

© The Author(s), 2026. Published by Cambridge University Press on behalf of Applied Probability Trust. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

For comparison with upper bounds, our theorems only give lower bounds on the error of subsampling algorithms. Very roughly, Theorem 1 says that, for the error of a subsampling algorithm that looks at M out of N datapoints, every step must be large until the number of steps T is at least $T \gtrsim N/M$; that is, it is *necessary* for the total number of points used in the algorithm to be comparable to the total number of datapoints. For a complementary bound, Theorem 4.1 of [18] implies that the analogous algorithm with *no* subsampling converges in O(1) steps under conditions that are often satisfied for the examples in this paper. In particular, for the algorithm without subsampling to have low error, it is *sufficient* for the total number of points used in the algorithm to be comparable to the total number of datapoints. The conditions in both our paper and [18] are slightly stronger than is necessary, so this qualitative comparison may not hold for every possible example discussed in both papers. See e.g. [9] for qualitatively similar upper bounds under weaker conditions.

Our paper gives only *lower* bounds on the error, but effectively matching upper bounds can be found in [7] and the references therein.

The current paper differs from the companion [10] in two main ways. The most important is that the companion paper did not give strong bounds for non-reversible chains, although many of the most popular subsampling chains are non-reversible. The current paper closes this gap, showing that the same qualitative conclusions hold for the simplest and most popular non-reversible algorithm when applied to classical statistical models. Second, by focusing on a specific algorithm, the current paper can give assumptions that are much easier to read and verify than the very generic assumptions in [10].

1.1. Related work

Early work such as [11], and later work such as [8] and [17], showed promising empirical results for subsampling MCMC. This was accompanied by theoretical work such as [2], [19], and [8] itself. These theoretical papers gave quantitative error bounds on subsampling algorithms, including proofs that the methods could be used to obtain consistent estimates. However, as explained in [2], these early theoretical results seemed to fall short of what was desired: bounds showing that subsampling MCMC was *better* than 'naive' MCMC for a wide range of realistic statistical examples, rather than merely showing that it was *not much worse.*

This paper was inspired by Nagapetyan et al. [12], who presented some evidence that this falling-short reflected a real problem in the underlying algorithm, not merely a limitation in the proof techniques. They studied a natural approach to analyzing a popular subsampling algorithm, called stochastic gradient Langevin dynamics (SGLD). The main conclusion was that, after appropriate counting of computational costs, these error bounds did not depend on the size of the subsample used. In other words, the error bounds could not be used to show that subsampling was an effective strategy. However, Nagapetyan et al. [12] left open the possibility that this was an artifact of their analysis of the algorithm; perhaps subsampling was effective, even if their error bounds could not show it.

In [10], some of the present authors showed that this was not the case for a wide variety of algorithms: subsampling MCMC really could not offer substantial speedups. However, [10] still had a number of important gaps. The most important was related to reversibility: the strongest conclusions in [10] only apply to reversible chains. Since the SGLD algorithm studied in [12] is non-reversible, this means that there was still an open question as to whether SGLD could result in substantial improvement in some regime.

The main contributions of this paper are short proofs that this does not happen: subsampling cannot improve certain important aspects of the performance of SGLD. Informally, Theorem 1

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

says the following: there are many posterior distributions for which *any* SGLD algorithm will return extremely poor samples *until* it has seen each data point at least once on average. Theorems 2 and 3 give analogous results in other regimes: SGLD where you have access to a 'good' control variate, and SGLD over a much longer running time.

This paper is a short note aimed at closing an important gap left by [10] and [12], and so we make no effort to give results that are as general as possible. For readers who might be interested in obtaining similar results for other non-reversible chains, we note that the key technical idea was to use the coupling construction given in Section 3.1 on a 'reasonable' forward-mapping representation of SGLD as given in Example 1. Exactly the same coupling construction can be used for *any* non-reversible chain with the same type of driving randomness, and this can be used to obtain small perturbations for the underlying Markov chain. Perturbation bounds for the Markov chain can then be combined with perturbation bounds for specific models, as described in the companion paper [10].

We note that there are many papers with *positive* results about the convergence of SGLD, saying that subsampling really does speed up computations in certain situations; see e.g. [4] and [13] for strong examples. At first glance these theorems may appear to be in conflict with ours, but in fact they are describing different phenomena. We certainly acknowledge that there are situations in which subsampling is unambiguously helpful. For two such examples, it is very easy to find situations in which subsampling vastly reduces the initial 'burn-in' period of an MCMC algorithm, and [4] makes it clear that using control variates can genuinely speed up SGLD. In other situations, the comparison between our results and those of [4] and [13] is more subtle. See Section 4.2 for a more detailed discussion.

Finally and most importantly, this paper is not the first to provide non-convergence results for SGLD. We particularly emphasize [4], which shows that the stationary distribution of SGLD is *extremely* far from the target posterior for a selection of reasonable models. Our analysis is closer to that of [12], focusing on the dynamics rather than the stationary distribution. Our analysis also gives non-trivial error lower bounds in situations where SGLD does converge to the correct distribution as n goes to infinity, either due to the choice of control variate (see Section 3.3) or minibatch size and step size (see Section 4.2).

2. Notation and generic bounds

We introduce generic notation and bounds for MCMC algorithms. These will be applied to specific target distributions later in the paper.

2.1. Basic notation

For two probability distributions μ, ν on a common Polish probability space $(\Omega, \mathcal{F})$, the total variation distance is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

For $n \in \mathbb{N}$, define $[n] = \{1, 2, \ldots, n\}$. By a small abuse of notation, we will sometimes treat a vector as a set (writing, for example, $x \in \nu$ rather than $x \in \{\nu_1, \ldots, \nu_n\}$) when there is no possible ambiguity. We will frequently consider datasets $\{X_1, \ldots, X_n\}$, and use $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ to denote the full dataset.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

2.2. Forward-mapping representation of subsampling MCMC

We consider the usual setup for Bayesian inference. We have a parameter space $(\Theta, \mathcal{F}_\Theta)$, a family of models $\{p(\cdot \mid \theta)\}_{\theta \in \Theta}$, a prior p, and a dataset $\mathcal{X} = \{X_1, \ldots, X_n\}$. For the entire paper, we assume that (Θ, d) is a Polish space and that $\mathcal{F}_\Theta$ is the usual Borel σ-algebra.

We now define a family of 'subsampling' MCMC algorithms, which we will show contains many popular algorithms as special cases.

Fix a pair of probability spaces $(\mathbb{A}, \mathcal{F}_\mathbb{A})$ and $(\Omega, \mathcal{F}_\Omega)$ and a maximum 'subsample size' $m \in \mathbb{N}$. Let $F\colon \Theta \times \mathbb{A} \times [n]^m \times \Omega \mapsto \Theta \times \mathbb{A}$ be a measurable function, let μ be a probability measure on $[n]^m$, and let η be a probability measure on Ω. The space Ω is used to capture all of the random variables sampled during a step of the underlying algorithm, *except* the choice of subsample. Typically, we can assume without loss of generality that this is of the form $[0, 1]^k$ with the usual Borel σ-algebra. These choices together define a forward-mapping representation for a Markov chain $\{\theta_t\}_{t \geq 0}$ via Algorithm 1.

Algorithm 1: Sampling with explicit randomness.

1: **Input:** Starting point $(\theta_0, r_0) \in \Theta \times \mathbb{A}$, number of iterations $T \in \mathbb{N}$.
2: **For** $t = 1, \ldots T$ **do**
3: Sample $E_t \sim \mu$, $U_t \sim \eta$.
4: Update state $(\theta_t, r_t) \leftarrow F(\theta_{t-1}, r_{t-1}, E_t, U_t)$
5: **end for**
6: Return the samples $\{\theta_1, \ldots, \theta_T\}$

One can check that this generic algorithm covers many popular subsampling algorithms, including stochastic gradient Langevin dynamics (see [21]) and subsampling pseudo-marginal Metropolis–Hastings (see [16]). The auxiliary variable r_t allows us to include popular algorithms such as stochastic gradient Hamiltonian Monte Carlo [6] or the pseudo-marginal algorithm [1] in the form of Algorithm 1.

As a concrete example, the following gives a representation of the usual SGLD algorithm on $\mathbb{R}^d$.

Example 1. Fix dimension $d \in \mathbb{N}$, step size $\epsilon > 0$, and subsample size $m = m(\theta) \leq M$. Then SGLD (with the usual Euler–Maruyama discretization) can be written in the form of Algorithm 1 with the choices

$$\mathbb{A} = \emptyset, \quad \Omega = \mathbb{R}^d, \quad \eta = N(0, \mathbf{1}_d), \quad \mu = \mathrm{Unif}([n]^M), \tag{1}$$

and

$$F(\theta, r, E, U) = \theta + \frac{\epsilon}{2}\left(\nabla \log p(\theta) + \frac{n}{m(\theta)} \sum_{j=1}^{m(\theta)} \nabla \log p(X_{E[j]} \mid \theta)\right) + \sqrt{\epsilon}\, U. \tag{2}$$

The main constraint comes from the restriction to independent and identically distributed (i.i.d.) selection of subsamples, which excludes some algorithms with complex subsample selection mechanisms (see e.g. [17] for a survey of approaches). We will show in later sections how some algorithms that do this can still be placed in the framework of this paper (see Remark 1).

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

2.3. Perturbations of posterior distributions of exponential families

We focus on exponential families with likelihoods of the form

$$p(x \mid \theta) = h(x) \exp(\theta R(x) - A(\theta)), \tag{3}$$

where $x \in \mathbb{R}$, h and R are real-valued functions, and A is the associated log-normalizing constant. Fixing a prior p on Θ, the associated posterior density is given by

$$p(\theta \mid \mathcal{X}_n) = \frac{\exp(\theta S - nA(\theta))p(\theta)}{\int \exp(\theta' S - nA(\theta'))p(\theta')\,\mathrm{d}\theta'},$$

where $S = \sum_{i=1}^{n} R(X_i)$ is the sufficient statistic.

We then consider the 'perturbed' posterior associated with a 'perturbed' sufficient statistic $S + n\delta$:

$$p_\delta(\theta \mid \mathcal{X}_n) = \frac{\exp(\theta(S + n\delta) - nA(\theta))p(\theta)}{\int \exp(\theta'(S + n\delta) - nA(\theta'))p(\theta')\,\mathrm{d}\theta'}.$$

We will use the following perturbation assumption on the likelihood when analyzing MCMC on 'small' time intervals.

Assumption 1. There exists $\gamma > 0$ so that for all sequences $c_n \to \infty$, all datasets $X_1, X_2, \ldots$, and all sequences $|\delta_n| > c_n/\sqrt{n}$,

$$\liminf_{n\to\infty} \|p(\theta \mid \mathcal{X}_n) - p_{\delta_n}(\theta \mid \mathcal{X}_n)\|_{\mathrm{TV}} \geq \gamma.$$

This condition says that the posterior distribution is somewhat sensitive to large changes in the data. This assumption holds with $\gamma = 1$ when the posterior distribution is close to Gaussian, as well as in many other situations. See Appendix A.1 for a simple condition under which Assumption 1 holds with high probability, and which is based only on the first two moments of the posterior. Appendix A.1 also contains an application of this sufficient condition to show that the Gaussian model satisfies this condition with $\gamma = 1$.

Assumption 1 is relevant to small time scales. On large time scales, we consider smaller perturbations, leading to the following assumption.

Assumption 2. There exists $\gamma > 0$ so that for all sequences $c_n \to 0$, all datasets $X_1, X_2, \ldots$, and all sequences $\delta_n < c_n/\sqrt{n}$,

$$\liminf_{n\to\infty} \frac{\|p(\theta \mid \mathcal{X}_n) - p_{\delta_n}(\theta \mid \mathcal{X}_n)\|_{\mathrm{TV}}}{\delta_n} \geq \gamma.$$

Again, this assumption is easy to verify for posterior distributions that are near-Gaussian, and it holds in substantial generality. See Appendix A.2 for a quick proof and related calculations.

3. Perturbation lower bounds for stochastic gradient Langevin dynamics targeting exponential families

We introduce our coupling construction, then prove our main result, Theorem 1.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

3.1. Coupling construction for SGLD

We consider the representation of SGLD given in Example 1. In this section, we construct a coupling of two Markov chains to ensure that they remain far apart with high probability. This construction represents the main observation behind our main result: it is possible to couple the choice of subsamples across two chains so that the two chains are *quite close in total variation,* but certain functionals are *somewhat far in expectation.*

We begin by constructing a distribution ν that is close (in total variation distance) to the distribution μ used in Example 1. Fix a parameter $\delta > 0$ and $M, n \in \mathbb{N}$ with $M \leq n$; n remains the size of the dataset, and we think of M as the subsample size for SGLD. Sample $E[1], \ldots, E[M] \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, 2, \ldots, n\})$, then independently sample $B[1], \ldots, B[M] \overset{\text{i.i.d.}}{\sim} \text{Bern}(\delta)$ and $E^+[1], \ldots, E^+[M] \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{\lceil n/2 \rceil, \ldots, n\})$. Finally, for $i \in \{1, 2, \ldots, M\}$, set

$$D[i] = \begin{cases} E[i], & B[i] = 0, \\ E^+[i], & B[i] = 1. \end{cases}$$

Let $\nu = \nu_{M,\delta}$ denote the distribution of the vector D and let $\mu = \mu_M$ be the distribution of the vector E.

Recall that sampling $E_t \sim \mu_M \equiv \text{Unif}([n]^M)$ gives the SGLD algorithm in equations (1) to (2), targeting the posterior $p(\theta \mid x)$. If we were to replace μ_M with $\nu_{M,\delta}$ in the algorithm, this construction turns out to give another valid SGLD algorithm; but targeting a certain 'weighted' posterior distribution. More precisely, we have the following result.

Lemma 1. (Weighted gradient estimation.) *Fix* $0 < \delta < 1$, *sample the set* $D \sim \nu_{M,\delta}$, *and define* $w_i = \mathbb{P}(D[1] = i)$ *for* $i \in \{1, \ldots, n\}$. *Then*

$$\mathbb{E}\left[\frac{n}{M} \sum_{j \in D} \nabla \log p(X_j \mid \theta)\right] = \sum_{i=1}^{n} n\, w_i \nabla \log p(X_i \mid \theta).$$

Proof. For all j, we have by the definition of w_i that

$$\mathbb{E}[n \nabla \log p(X_{D[1]} \mid \theta)] = \sum_{i=1}^{n} n\, w_i \nabla \log p(X_i \mid \theta).$$

The result follows from linearity of expectation. □

We interpret this as saying that

$$\frac{n}{M} \sum_{j \in D} \nabla \log p(X_j \mid \theta)$$

is an unbiased estimator for the gradient of the log-likelihood for the weighted model

$$\prod_{i=1}^{n} p(X_i \mid \theta)^{n w_i}.$$

For later reference, if $D \sim \nu_{M,\omega}$, $0 < \omega < 1$, and $s = \lceil n/2 \rceil$, then the weights are

$$w_i^{(n,\omega)} = \frac{1 - \omega}{n} + \frac{\omega}{n - s + 1} \mathbf{1}_{\{i \geq s\}}.$$

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

Define

$$\Pi^{\mathrm{wt}}_{n,\omega}(\theta \mid \mathcal{X}_n) \propto p(\theta) \prod_{i=1}^{n} p(X_i \mid \theta)^{n w_i^{(n,\omega)}}. \tag{4}$$

For the exponential family (3), this is the same as

$$\Pi^{\mathrm{wt}}_{n,\omega}(\cdot \mid \mathcal{X}_n) = p_{\delta_n^{\mathrm{wt}}(\omega)}(\cdot \mid \mathcal{X}_n), \quad \delta_n^{\mathrm{wt}}(\omega) = \omega\left(\frac{1}{n-s+1}\sum_{i=s}^{n} R(X_i) - \frac{1}{n}\sum_{i=1}^{n} R(X_i)\right). \tag{5}$$

We can now couple samples from the original and 'perturbed' SGLD algorithm.

Lemma 2. *There exists a universal constant C with the following property. Fix*

$$M \in \mathbb{N}, \quad 0 < \alpha < \frac{1}{2}, \quad 0 < \delta < \frac{1}{2}\min\left(1, \frac{C\alpha^2}{\sqrt{M}}\right).$$

Then

$$\|\mu_M - \nu_{M,\delta}\|_{TV} \le \alpha.$$

Proof. We use a calculation that is very similar to the proof in Section 1.3 of [5]. Let $D \sim \mu_M$ and $E \sim \nu_{M,\delta}$. Let $s = \lceil n/2 \rceil$, let $p = P[E[1] \ge s] - \frac{1}{2}$ and $q = \frac{1}{2} - P[E[1] < s]$ be the 'bias' induced by $\nu_{M,\delta}$. Let $D^+ = \{i \in [M] \colon D[i] \ge s\}$ and $E^+ = \{i \in [M] \colon E[i] \ge s\}$.

We are interested in counting the number of times that points with large index are chosen. Let f be the density of the distribution $\mathcal{L}(|D^+|)$ with respect to $\mathcal{L}(|E^+|)$. We then have

$$f(D) = \left(\frac{1}{2}+p\right)^{|D^+|}\left(\frac{1}{2}-q\right)^{M-|D^+|} = \left[\left(\frac{1}{2}+p\right)\left(\frac{1}{2}-q\right)\right]^{M/2}\left(\frac{1+2p}{1-2q}\right)^{|D^+|-M/2}.$$

Thus

$$\|\mu_M - \nu_{M,\delta}\|_{TV} \le \mathbb{E}\left[\left(1 - \left[\left(\frac{1}{2}+p\right)\left(\frac{1}{2}-q\right)\right]^{M/2}\left(\frac{1+2p}{1-2q}\right)^{D^+-M/2}\right)_+\right].$$

Note that D^+ is a binomial random variable. Applying Chebyshev's inequality to D^+, the same calculation as in Section 1.3 of [5] says that there exists a universal constant $C > 0$ so that

$$\|\mu_M - \nu_{M,\delta}\|_{TV} \le C\sqrt{\alpha}.$$

This completes the proof. □

Remark 1. (*Complex selection measures.*) Note that the proof of Lemma 2 depended entirely on bounding the density f. In particular, the same argument can be used with minimal changes even when the driving measure μ_M is not uniform. We do not attempt to give universal estimates here, as the literature on SGLD contains a large variety of driving measures with rather different behaviors.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

We now wish to establish that this small perturbation of the driving randomness corresponds to a similarly sized perturbation of the sufficient statistics in the effective model. We have the following result.

Lemma 3. *Fix $\delta > 0$, $M \in \mathbb{N}$ and a non-decreasing sequence $X_1 \le X_2 \le \cdots \le X_n$. Let $s = \lceil n/2 \rceil$. Let $E \sim \mu_M$ and $D \sim \nu_{M,\delta}$. Then*

$$\mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M} R(X_{D[j]}) - \frac{1}{M}\sum_{j=1}^{M} R(X_{E[j]})\right] = \delta\left(\frac{1}{n-s+1}\sum_{i=s}^{n} R(X_i) - \frac{1}{n}\sum_{i=1}^{n} R(X_i)\right).$$

Proof. This follows immediately from the definition of μ_M, $\nu_{M,\delta}$. □

We add the (very weak) assumption that this is large with high probability.

Assumption 3. A likelihood of the form (3) and parameter θ_0 have the following property.

Fix $n \in \mathbb{N}$, let $s = \lceil n/2 \rceil$, let $Y_1, Y_2, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta_0)$, and let $X_1 \le X_2 \le \cdots \le X_n$ be the same points put in ascending order. Then

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{1}{n-s+1}\sum_{i=s}^{n} R(X_i) - \frac{1}{n}\sum_{i=1}^{n} R(X_i)\right| < \eta\right] = 0.$$

This assumption is quite weak, and is straightforward to check for popular functions R. To give a concrete example, when $R(x) = x$, this follows immediately from Chebyshev's inequality.

3.2. Lower bounds for short runs

We set notation for the main result. Fix a prior p and model $p(\cdot \mid x)$. For fixed dataset X and sample size M, let $K_{X,M}$ denote the transition kernel associated with the SGLD algorithm in Example 1. Similarly, for fixed $\delta > 0$, let $K_{X,\delta,M}$ denote the transition kernel associated with the SGLD algorithm in Example 1, with μ_M replaced by $\nu_{M,\delta}$.

Finally, we set sequences of integers T_n and (possibly time-inhomogeneous) batch sizes $m_{n,1}, \ldots, m_{n,T_n}$ and perturbation sizes ω_n, and define the total number of sampled datapoints

$$G_n = \sum_{t=1}^{T_n} m_{n,t}. \tag{6}$$

In the following theorem, we will assume that

$$\lim_{n\to\infty} \frac{G_n}{n} = 0 \tag{7}$$

and

$$\lim_{n\to\infty} \omega_n \sqrt{G_n} = 0, \qquad \lim_{n\to\infty} \omega_n \sqrt{n} = \infty. \tag{8}$$

Our main result is as follows.

Theorem 1. *Fix a prior p, likelihood $p(\cdot \mid \theta)$ and parameter value θ_0 satisfying Assumptions 1 and 3. Let $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta_0)$ be data sampled from this. For each n, let $X_1 \le \cdots \le X_n$ be $Y_1, \ldots, Y_n$ put in ascending order. Let $T_n \in \mathbb{N}$, step sizes $\epsilon_{n,1}, \ldots, \epsilon_{n,T_n} > 0$, perturbation*

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

sizes ω_n and batch sizes $m_{n,1}, \dots, m_{n,T_n} \in \{1, \dots, n\}$ be deterministic (depending only on n), and let G_n be as in (6). Assume equations (7) and (8) hold.

Let $Z_1^{(n)}, Z_2^{(n)}, \dots, Z_{T_n}^{(n)}$ be generated by Algorithm 1 with the SGLD update maps of Example 1, using $\epsilon_t = \epsilon_{n,t}$ and using only the first $m_{n,t}$ entries of the selection variable E_t at time t. Similarly, let $\tilde{Z}_1^{(n)}, \dots, \tilde{Z}_{T_n}^{(n)}$ be generated in the same way, but with the selection variable E_t sampled from the perturbed law corresponding to ω_n (i.e. with μ_M replaced by ν_{M,ω_n} in the construction of E_t) for every datapoint query. Then, for all $a > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big[\big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - p(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} + \big\|\mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} \geq (1-a)\gamma\Big] = 1.$$

Remark 2. Our notation suppresses certain dependencies, such as the step size and the distribution of the first point. This should be read as shorthand for the following: the suppressed parameters of the algorithm should be taken to be a function of *only* the size n of the data, but they may be *any* such function. In particular, this means that, for example, the starting distribution for the chain $Z_0^{(n)}$ in the theorem statement could be the target $p(\cdot \mid \mathcal{X}_n)$ for each $n \in \mathbb{N}$; but then the starting distribution for the chain $\tilde{Z}_0^{(n)}$ would also be the target $p(\cdot \mid \mathcal{X}_n)$ for each $n \in \mathbb{N}$. This is a natural assumption to make, as tuning parameters are often strongly dependent on the size n of the dataset, and we do not wish to restrict the allowed tuning parameters.

Proof. For fixed n, let $\mathcal{E}_n$ denote the event that

$$\left| \frac{1}{n-s+1} \sum_{i=s}^{n} R(X_i) - \frac{1}{n} \sum_{i=1}^{n} R(X_i) \right| \geq \eta,$$

where $s = \lceil n/2 \rceil$.

We first check that the target distributions of our algorithms are far. Combining Lemma 1, Lemma 3, and (5), on the event $\mathcal{E}_n$ we have

$$\liminf_{n\to\infty} \|p(\cdot \mid \mathcal{X}_n) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\|_{\mathrm{TV}} \geq \gamma. \tag{9}$$

We next check that the samples of our algorithms are close. Let $\mathbf{E}^{(n)}$ denote the concatenation of all indices actually queried by the first run, i.e. the vector in $[n]^{G_n}$ obtained by concatenating $E_t[1], \dots, E_t[m_{n,t}]$ over $t = 1, \dots, T_n$, and define $\tilde{\mathbf{E}}^{(n)}$ analogously for the perturbed run. Then $\mathbf{E}^{(n)} \sim \mu_{G_n}$ and $\tilde{\mathbf{E}}^{(n)} \sim \nu_{G_n,\omega_n}$. By Lemma 2 and (8), we may couple $\mathbf{E}^{(n)}$ and $\tilde{\mathbf{E}}^{(n)}$ so that $\mathbb{P}[\mathbf{E}^{(n)} \neq \tilde{\mathbf{E}}^{(n)}] \to 0$. Coupling also the Gaussian noises $\{U_t\}$ identically across the two runs, the forward-mapping representation implies that the full trajectories coincide with probability tending to 1:

$$\lim_{n\to\infty} \mathbb{P}\Big[\big(Z_1^{(n)}, \dots, Z_{T_n}^{(n)}\big) = \big(\tilde{Z}_1^{(n)}, \dots, \tilde{Z}_{T_n}^{(n)}\big)\Big] = 1. \tag{10}$$

Next, by the triangle inequality,

$$\begin{aligned}\Delta_n &\equiv \big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - p(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} + \big\|\mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} \\ &\geq \|p(\cdot \mid \mathcal{X}_n) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\|_{\mathrm{TV}} - \big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - \mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big)\big\|_{\mathrm{TV}}.\end{aligned}$$

Applying (10) and (9) to the terms on the right-hand side of this inequality, we have for all $a > 0$

$$\lim_{n\to\infty} \mathbb{P}[\Delta_n < (1-a)\gamma] \leq \lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n^c].$$

By Assumption 3, this last limit is 0, completing the proof. □

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

Remark 3. (*Why does Theorem 1 bound a sum of errors?*) We note that Theorem 1 gives a lower bound on the sum of errors for two Markov chains, while it is more common to give a lower bound on the error of a single Markov chain. This is not an accident, and indeed it is completely unavoidable given our assumptions: we allow the starting measure $\mathcal{L}(X_0)$ to be the target distribution, and we allow the number of steps T to be 0, so in fact one of the errors may be exactly 0!

Of course, setting $T = 0$ is a rather degenerate situation. More broadly, we note that it is quite possible for a particular procedure to give the right answer for a particular problem essentially 'by accident'. Thus we show that while it is possible for *one* chain to get the right answer, it is not possible for *both* chains to get the right answer. In practice, a user will never know which situation they are in, and so will not be able to take advantage of any 'accidental' accuracy of this type.

Remark 4. (*Why care about weighted posteriors?*) Theorem 1 shows that *at least one* of two SGLD chains must have large error: either the original chain $\{Z_t^{(n)}\}$, *or* a chain $\{\tilde{Z}_t^{(n)}\}$ targeting the weighted posterior $\Pi^{\mathrm{wt}}_{n,\omega_n}(\cdot \mid \mathcal{X}_n)$ defined in (4). It is natural to ask: Why would one care about these weighted posteriors?

First, we must acknowledge that we would prefer to show that *all* chains have large error, rather than showing *at least one chain* has large error. Despite this, we feel that there are at least two very good reasons to care about error in the weighted posterior.

(i) These weighted posterior distributions are in fact fairly common in Bayesian statistics. We do not give a full survey of places where they occur, but two prominent families of applications are using them as pseudo-likelihoods when doing post-stratification of collected data (see e.g. Section 2.1 of [20]) and using them as computational tools for simulated tempering or other annealing-like procedures (see e.g. [3] for a situation in which non-constant weights are used). Our point is not that these weighted posteriors appear in almost all Bayesian analyses, merely that they are fairly common for a variety of reasons and so one would hope a general-purpose algorithm would be able to sample from them.

(ii) While Theorem 1 does not rule out the possibility that the error of the original chain is typically much smaller than the error of the 'weighted' chain, we feel that this would be somewhat miraculous and extremely interesting in its own right. Certainly the algorithm itself does not obviously discriminate between weights that are exactly 1 and weights that are similar to 1. We are also aware of many instances (such as the papers cited above) in which people have freely used SGLD and other gradient-based methods for weighted posteriors, and we are not aware of any situations in which it has been reported that SGLD has very different performance for unweighted and slightly weighted posterior distributions.

Finally, for $T \ll n/M$ sufficiently small, it is possible to prove a version of Theorem 1 in which we obtain a lower bound on the sums of the errors of two *different unweighted* datasets, rather than a lower bound on the sums of the errors of two *different weightings* of the same dataset. We quickly sketch the key change that would be needed in the proof. In Lemma 2, we show that it is possible to couple the driving randomness of two SGLD chains associated with two slightly different weightings of the same dataset. By essentially the same argument, it is possible to couple the driving randomness of two SGLD chains associated with two slightly different datasets. (One small difference is that we would need to change Assumption 3. Currently,

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

our assumption largely boils down to assuming that the data is not nearly constant. Proving a perturbation bound analogous to Lemma 2 would require the assumption that the data is sampled from a density that does not fluctuate very quickly.) Propagating the changes gives the desired theorem.

3.3. Lower bounds with strong control variates

Theorem 1 gave a lower bound on the error of a naive SGLD algorithm. It is known that naive SGLD can fail catastrophically (see [4] for a typical example in which the stationary distribution is very far from the ostensible target), and so in practice it is common to use control variates that are highly informative but not quite sufficient statistics. In this section, we give lower bounds on the error rate of SGLD for a toy model of this situation: a log-likelihood that is the sum of one dominant term that is used as a control variate, and a second lower-order term that is estimated from the data. In this regime, the stationary measure of SGLD will converge to the 'right' answer as n goes to infinity, and we obtain lower bounds that depend on the error in the second-order term. Informally, we view these as being similar to Theorem 1: the accuracy of SGLD cannot improve upon some natural baseline until a large fraction of the data has been used.

Let $\theta = (\theta^{(1)}, \theta^{(2)}) \in \Theta \subseteq \mathbb{R}^2$ and consider the two-parameter exponential family

$$p(x \mid \theta) = \exp\big(\theta^{(1)} R_1(x) + \theta^{(2)} R_2(x) - A(\theta)\big) p_0(x).$$

For data $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, define sufficient statistics

$$S_1 = \sum_{i=1}^{n} R_1(X_i), \quad S_2 = \sum_{i=1}^{n} R_2(X_i),$$

scaling parameter α_n, and pseudo-posterior

$$p(\theta \mid \mathcal{X}_n) \propto \exp\big(\theta^{(1)} S_1 + \alpha_n \, \theta^{(2)} S_2 - n A(\theta)\big) p(\theta).$$

We consider an SGLD-like algorithm in which S_1 is available exactly for use as a control variate, while S_2 is estimated by subsampling.

Example 2. (*Simple SGLD with control variates.*) Fix step sizes $\epsilon_t > 0$ and batch sizes $m_t \in \{1, \ldots, n\}$. At time t, sample indices $E_t[1], \ldots, E_t[m_t]$ i.i.d. from Unif($\{1, \ldots, n\}$) and define

$$\widehat{S}_{2,t} := \frac{n}{m_t} \sum_{j=1}^{m_t} R_2(X_{E_t[j]}).$$

With i.i.d. $U_t \sim N(0, \mathbf{1}_d)$, update

$$\theta_t = \theta_{t-1} + \frac{\epsilon_t}{2} (\nabla \log p(\theta_{t-1}) + (S_1, \alpha_n \widehat{S}_{2,t}) - n \nabla A(\theta_{t-1})) + \sqrt{\epsilon_t} \, U_t.$$

3.3.1. *Short-run lower bound.* For $\delta \in \mathbb{R}$, define the 'second-statistic perturbed posterior'

$$p^{(2)}_{\delta}(\theta \mid \mathcal{X}_n) \propto \exp\big(\theta^{(1)} S_1 + \alpha_n \, \theta^{(2)} (S_2 + n\delta) - n A(\theta)\big) p(\theta).$$

We use the following analog of Assumption 1 for perturbations in the R_2-statistic.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

Assumption 4. There exist constants c, $\gamma_2 > 0$, and a deterministic sequence $\beta_n \downarrow 0$ so that for all datasets $\mathcal{X}_n$ and all sequences $\delta_n > c/\sqrt{n}$,

$$\liminf_{n\to\infty} \frac{\left\| p(\cdot \mid \mathcal{X}_n) - p^{(2)}_{\delta_n}(\cdot \mid \mathcal{X}_n) \right\|_{\mathrm{TV}}}{\beta_n} \geq \gamma_2.$$

We also use the obvious analog of Assumption 3, again applied to R_2.

Assumption 5. There exists $\eta > 0$ such that, letting $s = \lceil n/2 \rceil$ and letting $X_1 \leq \cdots \leq X_n$ denote the ordered sample from $p(\cdot \mid \theta_0)$, we have

$$\lim_{n\to\infty} \mathbb{P}\left[\left| \frac{1}{n-s+1} \sum_{i=s}^{n} R_2(X_i) - \frac{1}{n} \sum_{i=1}^{n} R_2(X_i) \right| < \eta \right] = 0.$$

Fix sequences T_n and batch sizes $m_{n,1}, \ldots, m_{n,T_n}$ and define $G_n = \sum_{t=1}^{T_n} m_{n,t}$. Consider two runs of Example 2 with identical step sizes and Gaussian noises, but with different index laws:

(i) the *uniform* run draws all queried indices i.i.d. from Unif($\{1, \ldots, n\}$);

(ii) the *perturbed* run draws all queried indices i.i.d. from the one-sample mixture corresponding to parameter ω_n.

Define the (data-dependent) perturbation size, where $s = \lceil n/2 \rceil$,

$$\delta_n^{(2)} := \omega_n \left(\frac{1}{n-s+1} \sum_{i=s}^{n} R_2(X_i) - \frac{1}{n} \sum_{i=1}^{n} R_2(X_i) \right).$$

We are ready to state our analog of Theorem 1. We emphasize that there is one substantial difference: the lower bound scales like β_n rather than as a constant. This should be unsurprising. Since S_2 contains less information than S_1, even MCMC based only on the known value of S_1 would result in an error that goes to 0 with n, at a rate that depends on β_n. Our result shows that one needs to see a substantial fraction of the data to improve on this control-variate-only estimate.

Theorem 2. *Fix notation as in the statement of Theorem 1, except for the following changes. Replace Assumptions 1 and 3 with Assumptions 4 and 5. Replace the conditions on G_n, ω_n with the assumption that $G_n/n \to 0$, $\omega_n\sqrt{G_n} \to 0$, $\liminf_{n\to\infty} \omega_n\sqrt{n} > 0$ and $\omega_n^{1/2}(G_n)^{1/4} = o(\beta_n)$. Let $Z^{(n)}_{T_n}$ be the output of the uniform run and $\tilde{Z}^{(n)}_{T_n}$ the output of the perturbed run.*

Then, for all $a > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left\| \mathcal{L}\left(Z^{(n)}_{T_n}\right) - p(\cdot \mid \mathcal{X}_n) \right\|_{\mathrm{TV}} + \left\| \mathcal{L}\left(\tilde{Z}^{(n)}_{T_n}\right) - p^{(2)}_{\delta_n^{(2)}}(\cdot \mid \mathcal{X}_n) \right\|_{\mathrm{TV}} \geq (1-a)\gamma_2\beta_n \right] = 1.$$

Proof. The proof is the same as the proof of Theorem 1, with R replaced by R_2 and with p_δ replaced by $p^{(2)}_\delta$. We sketch the two main changes.

(i) *Targets are far.* By Lemma 3 applied to R_2 and the definition of $\delta_n^{(2)}$, on the event in Assumption 5 we have $|\delta_n^{(2)}| \geq \omega_n\eta$. Since $\liminf_{n\to\infty} \omega_n\sqrt{n} > 0$, this implies $|\delta_n^{(2)}| > c/\sqrt{n}$, and hence Assumption 4 yields a total-variation separation of at least $\gamma_2\beta_n$ between $p(\cdot \mid \mathcal{X}_n)$ and $p^{(2)}_{\delta_n^{(2)}}(\cdot \mid \mathcal{X}_n)$ with probability tending to 1.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

(ii) *Samples are close.* By Lemma 2, together with the assumption $\omega_n^{1/2}(G_n)^{1/4} = o(\beta_n)$, we may couple the concatenated index vectors so that they agree with probability $1 - o(\beta_n)$. Coupling the Gaussian noises identically and using the forward-mapping representation then implies that the two trajectories agree with probability $1 - o(\beta_n)$. The triangle inequality completes the argument exactly as in the proof of Theorem 1. □

Remark 5. In typical near-Gaussian settings, we expect the best choice of β_n in Assumption 4 to scale proportionally to α_n for perturbations of size $\delta_n \asymp 1/\sqrt{n}$. In that case, the additional condition $\omega_n^{1/2}(G_n)^{1/4} = o(\beta_n)$ simply requires that the coupling error in the subsample indices is negligible at the scale of the S_2-induced posterior shift.

3.4. Lower bounds for long runs

We give results that are very similar to those of Theorem 1, but for longer time scales. We use the same notation as that immediately preceding the statement of Theorem 1, but replace conditions (7) and (8) with

$$\lim_{n\to\infty} \frac{G_n}{n} = \infty \tag{11}$$

and

$$\lim_{n\to\infty} \omega_n \sqrt{G_n} = 0. \tag{12}$$

Lemma 4. *There exists a universal constant $C_{\mathrm{coup}} > 0$ such that the following holds. Fix $G \in \mathbb{N}$ and $0 < \delta < \frac{1}{2}$. Let $\mathbf{E} \sim \mu_G$ and $\mathbf{D} \sim \nu_{G,\delta}$. Then*

$$\|\mu_G - \nu_{G,\delta}\|_{\mathrm{TV}} \le C_{\mathrm{coup}}\, \delta^{1/2}\, G^{1/4}.$$

In particular, there exists a coupling of $\mathbf{E}$ and $\mathbf{D}$ such that $\mathbb{P}[\mathbf{E} \neq \mathbf{D}] \le C_{\mathrm{coup}}\, \delta^{1/2}\, G^{1/4}$.

Proof. This is an immediate corollary of Lemma 2 by taking $M = G$ and choosing the parameter α there proportional to $\delta^{1/2} G^{1/4}$. The second statement follows from the standard coupling characterization of total variation distance. □

The proof is quite similar to the proof of Theorem 1.

Theorem 3. *Fix notation as in the statement of Theorem 1, with the following changes. Replace Assumption 1 with Assumption 2. Replace the conditions on ω_n, G_n with equations (11) and (12).*

Then, for all $a > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big[\big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - p(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} + \big\|\mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\big\|_{\mathrm{TV}} \ge (1-a)\gamma\Big] = 1.$$

Proof. For fixed n, let $\mathcal{E}_n$ denote the event that

$$\left| \frac{1}{n-s+1} \sum_{i=s}^{n} R(X_i) - \frac{1}{n} \sum_{i=1}^{n} R(X_i) \right| \ge \eta,$$

where $s = \lceil n/2 \rceil$.

We first check that the target distributions of our algorithms are far. Combining Lemma 1, Lemma 3, and (5), on the event $\mathcal{E}_n$ we have

$$\liminf_{n\to\infty} \omega_n^{-1} \|p(\cdot \mid \mathcal{X}_n) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\|_{\mathrm{TV}} \ge \eta\, \gamma. \tag{13}$$

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

We next check that the samples of our algorithms are close.

Let $\mathbf{E}^{(n)} \in [n]^{G_n}$ denote the concatenation of all indices queried by the original run (i.e. the concatenation of the first $m_{n,t}$ entries of E_t over $t = 1, \ldots, T_n$), and let $\mathbf{D}^{(n)} \in [n]^{G_n}$ denote the analogous concatenation for the perturbed run. By Lemma 4, we may couple $\mathbf{E}^{(n)}$ and $\mathbf{D}^{(n)}$ so that

$$\mathbb{P}[\mathbf{E}^{(n)} \neq \mathbf{D}^{(n)}] \leq C_{\text{coup}}\, \omega_n^{1/2}\, G_n^{1/4}.$$

Coupling also the Gaussian noises identically across the two runs and using the forward-mapping representation, it follows that under this coupling

$$\mathbb{P}\big[\big(Z_1^{(n)}, \ldots, Z_{T_n}^{(n)}\big) \neq \big(\tilde{Z}_1^{(n)}, \ldots, \tilde{Z}_{T_n}^{(n)}\big)\big] \leq C_{\text{coup}}\, \omega_n^{1/2}\, G_n^{1/4}, \tag{14}$$

where C_{coup} is the constant from Lemma 4. Next, by the triangle inequality,

$$\begin{aligned}\Delta_n &\equiv \big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - p(\cdot \mid \mathcal{X}_n)\big\|_{\text{TV}} + \big\|\mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big) - \Pi^{\text{wt}}_{n,\omega_n}(\cdot \mid \mathcal{X}_n)\big\|_{\text{TV}} \\ &\geq \|p(\cdot \mid \mathcal{X}_n) - \Pi^{\text{wt}}_{n,\omega_n}(\cdot \mid \mathcal{X}_n)\|_{\text{TV}} - \big\|\mathcal{L}\big(Z_{T_n}^{(n)}\big) - \mathcal{L}\big(\tilde{Z}_{T_n}^{(n)}\big)\big\|_{\text{TV}}.\end{aligned}$$

Applying inequalities (14) and (13) to the terms on the right-hand side of this inequality, we have for all $a > 0$

$$\lim_{n\to\infty} \mathbb{P}\big[\Delta_n < \eta\gamma\, \omega_n - C_{\text{coup}}\, \omega_n^{1/2} (G_n)^{1/4}\big] \leq \lim_{n\to\infty} \mathbb{P}[\mathcal{E}_n^c].$$

By Assumption 3, this last limit is 0, so for all $a > 0$

$$\lim_{n\to\infty} \mathbb{P}\big[\Delta_n < (1-a)\eta\gamma\omega_n + \big(a\eta\gamma\omega_n - C_{\text{coup}}\omega_n^{1/2}(G_n)^{1/4}\big)\big] = 0.$$

By assumption (12),

$$\liminf_{n\to\infty} \big(a\eta\gamma\omega_n - C_{\text{coup}}\omega_n^{1/2}(G_n)^{1/4}\big) \geq 0,$$

so this completes the proof. □

4. Discussion

We have shown that, under certain common conditions, SGLD is not much more efficient than the 'full' gradient Langevin dynamics (usually called ULA in the MCMC literature). We comment on a few questions about subsampling MCMC that are left open by this work.

4.1. Filling in the blanks

Assumption 3 is quite a strong assumption about the form of the underlying model, and clearly we expect a similar phenomenon to hold in much greater generality. We make two comments.

(i) *One sufficient step.* Inspecting the proof of Theorem 1, we see that Assumption 3 is only used when checking inequality (9). Inequality (9) itself is a basic fact about the target distribution: that 'typical' small tweaks to the distribution of the observed data result in 'similarly sized' changes to the posterior. Checking this fact in great generality is somewhat difficult and so beyond the scope of a short paper. However, we expect this

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

fact to be true fairly generically. Most of our companion paper [10] is spent on proving analogous (if more complicated) facts in much more general settings. In addition, it is straightforward to estimate the size of the perturbation in inequality (9) from data in low-dimensional problems, so in practice this assumption can be checked for specific problems of interest.

(ii) *Practical conclusions.* There is a natural concern that a theoretical result proved only for simple distributions (such as Gaussians) may not extend to more reasonable problems. This skepticism is very reasonable for 'positive' results, such as showing that an estimator converges *quickly*: a method might be taking advantage of the special properties of the simple distribution. However, the main result in this paper is 'negative': it shows that an estimator is *slow* for some class of simple distributions. We think it would be quite surprising if generic methods such as SGLD happened to fail only on a small class of simple examples, and we are not aware of any empirical work suggesting that this phenomenon happens.

4.2. Lower bounds under other scalings

The lower bounds in [4] and [12] focus on the setting of constant step size, and show that (under certain conditions) SGLD does not converge quickly to a distribution that is close to the target. However, various positive theoretical results on SGLD show that the algorithm *does* converge *eventually* to the intended target upon appropriate rescaling [7, Theorem 4]. In this section, we show that our qualitative conclusions hold even in this setting: Theorem 1 implies that SGLD does not give samples that are close to the intended target until roughly n datapoints have been observed, even if the algorithm scales so that it converges to the target.

Mimicking the notation of [7], we fix exponents α, β and constant $c > 0$ and set a constant step size and minibatch size

$$h_n = c\, n^{-\alpha}, \quad b_n = \lceil n^{\beta} \rceil.$$

Let $\{Z_t^{(n)}\}$ denote an SGLD chain with step size and minibatch size as above, and let π_n denote the intended target. Under the assumptions stated in [7, Theorem 4], the algorithm has error

$$W_2\big(\mathcal{L}\big(Z_{K_n}^{(n)}\big), \pi_n\big) = \mathrm{O}\big(n^{-\alpha/2}\big) + \mathrm{O}\big(n^{(1-\alpha-\beta)/2}\big),$$

after

$$K_n \asymp n^{\alpha-1} \log n$$

steps. In particular, under the condition that $\alpha \geq 1$ and $\alpha + \beta > 1$, we have

$$\lim_{n\to\infty} W_2\big(\mathcal{L}\big(Z_{K_n}^{(n)}\big), \pi_n\big) = 0. \tag{15}$$

We now compare this to our results. Fix any $\beta \in (0, 1)$ and take $\alpha = 1$. Run $T_n \asymp n^{1-\beta} / \log n$ iterations with $m_{n,t} \equiv b_n = \lceil n^{\beta} \rceil$, so that the total number of datapoint queries is

$$G_n = \sum_{t \leq T_n} m_{n,t} = T_n b_n \approx \frac{n}{\log n}.$$

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

Since $T_n \to \infty$, equation (15) applies. On the other hand, because $G_n = o(n)$, if we choose $\omega_n = n^{-1/2}(\log n)^{1/4}$, we satisfy the limits $\omega_n\sqrt{G_n} \to 0$ and $\omega_n\sqrt{n} \to \infty$ required by Theorem 1. Applying Theorem 1 yields

$$\lim_{n\to\infty} \mathbb{P}\left[\left\|\mathcal{L}\left(Z_{T_n}^{(n)}\right) - p(\cdot \mid \mathcal{X}_n)\right\|_{\mathrm{TV}} + \left\|\mathcal{L}\left(\tilde{Z}_{T_n}^{(n)}\right) - \Pi_{n,\omega_n}^{\mathrm{wt}}(\cdot \mid \mathcal{X}_n)\right\|_{\mathrm{TV}} \geq \frac{1}{2}\gamma\right] = 1.$$

Informally, running this algorithm until you observe $G_n \approx n/\log(n)$ datapoints still gives the lower bound above with $\omega_n = n^{-1/2}(\log n)^{1/4}$, even though (by (15)) the algorithm does converge as the running time and dataset size become large.

Appendix. Verifying perturbation assumptions

We give some concrete calculations for verifying the perturbation assumptions in certain special cases. Qualitatively similar results are known for many other distributions, but (to our knowledge) the required calculations are longer and more abstract. See Chapter 3 of [15] for our favorite introduction to such bounds.

A.1. Verifying Assumption 1

Assumption 1 is unfamiliar but seems to hold for a wide variety of popular models. In many cases, including Gaussians, this can be easily verified using simple moment bounds.

We recall the following sharp moment condition found in Theorem 1 of [14]. For any distributions P, Q on $\mathbb{R}$ with distinct means μ_P, μ_Q and variances σ_P^2, σ_Q^2,

$$\|P - Q\|_{\mathrm{TV}} \geq \frac{(\mu_P - \mu_Q)^2}{(\sigma_P + \sigma_Q)^2 + (\mu_P - \mu_Q)^2}.$$

In the special case that $p(x \mid \theta)$ is a Gaussian with mean θ and fixed variance 1 and the prior p is uninformative, this bound gives

$$\|p(\theta \mid (x_1, \ldots, x_n)) - p_{\delta_n}(\theta \mid (x_1, \ldots, x_n))\|_{\mathrm{TV}} \geq \frac{\delta_n^2}{4n^{-1} + \delta_n^2}.$$

Since we are interested in the regime that $\delta_n \gg 1/\sqrt{n}$, this lower bound converges to 1 as n goes to infinity.

A.2. Verifying Assumption 2

We first check that Assumption 2 holds for Gaussians. Let μ_1, μ_2 denote two Gaussians with means $\theta_1 < \theta_2$ and identical variance 1. It is straightforward to check that μ_1 has higher density than μ_2 exactly on the set

$$A = \left\{x \colon x < \frac{\theta_1 + \theta_2}{2}\right\}.$$

Thus, letting Φ denote the CDF of a standard Gaussian, we have

$$\begin{aligned}\|\mu_1 - \mu_2\|_{\mathrm{TV}} &= 2(\mu_1(A) - \mu_2(A)) \\ &= 2\left(\Phi\left(\frac{\theta_2 - \theta_1}{2}\right) - \Phi\left(-\frac{\theta_2 - \theta_1}{2}\right)\right) \\ &= \frac{2}{\sqrt{2\pi}}(\theta_2 - \theta_1) + O((\theta_2 - \theta_1)^2).\end{aligned}$$

Thus, with an uninformative prior, Assumption 2 is satisfied for any $0 < \gamma < 2/\sqrt{2\pi}$.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press

Funding information

AS thanks NSERC for support via DG RGPIN-2022-03012.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37**, 697–725.
[2] Bardenet, R., Doucet, A. and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *J. Mach. Learn. Res.* **18**, 1515–1557.
[3] Bon, J. J., Lee, A. and Drovandi, C. (2021). Accelerating sequential Monte Carlo with surrogate likelihoods. *Statist. Comput.* **31**, 62.
[4] Brosse, N., Durmus, A. and Moulines, E. (2018). The promises and pitfalls of stochastic gradient Langevin dynamics. In *Advances in Neural Information Processing Systems 31,* ed. S. Bengio et al. Curran Associates.
[5] Chatterjee, S. (2019). A general method for lower bounds on fluctuations of random variables. *Ann. Prob.* **47**, 2140–2171.
[6] Chen, T., Fox, E. B. and Guestrin, C. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning* (Proceedings of Machine Learning Research **32**), ed. E. P. Xing and T. Jebara, pp. 1683–1691. PMLR.
[7] Dalalyan, A. S. and Karagulyan, A. (2019). User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Process. Appl.* **129**, 5278–5311.
[8] De Sa, C., Chen, V. and Wong, W. (2018). Minibatch Gibbs sampling on large graphical models. In *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research **80**), ed. J. Dy and A. Krause, pp. 1165–1173. PMLR.
[9] Durmus, A. and Moulines, É. (2022). On the geometric convergence for MALA under verifiable conditions. Available at https://arxiv.org/abs/2201.01951.
[10] Johndrow, J. E., Pillai, N. S. and Smith, A. (2025+). No free lunch for approximate MCMC. Available at https://arxiv.org/abs/2010.12514.
[11] Korattikara, A., Chen, Y. and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning* (Proceedings of Machine Learning Research **32**), ed. E. P. Xing and T. Jebara, pp. 181–189. PMLR.
[12] Nagapetyan, T., Duncan, A. B., Hasenclever, L., Vollmer, S. J., Szpruch, L. and Zygalakis, K. (2017). The true cost of stochastic gradient Langevin dynamics. Available at https://arxiv.org/abs/1706.02692.
[13] Nemeth, C. and Fearnhead, P. (2021). Stochastic gradient Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **116**, 433–450.
[14] Nishiyama, T. (2023+). Lower bounds for the total variation distance given means and variances of distributions. Available at https://arxiv.org/abs/2212.05820.
[15] Pollard, D. (2005). *Asymptopia*. Available at http://www.stat.yale.edu/pollard/Books/.
[16] Quiroz, M., Villani, M. and Kohn, R. (2019). Speeding up MCMC by efficient data subsampling. *J. Amer. Statist. Assoc.* **114**, 831–843.
[17] Quiroz, M., Villani, M., Kohn, R., Tran, M.-N. and Dang, K.-D. (2018). Subsampling MCMC: An introduction for the survey statistician. *Sankhyā A* **80**, 1–37.
[18] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**, 341–363.
[19] Rudolf, D. and Schweizer, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24**, 2610–2639.
[20] Savitsky, T. D. and Toth, D. (2016). Bayesian estimation under informative sampling. *Electron. J. Statist.* **10**, 1677–1708.
[21] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, ed. L. Getoor and T. Scheffer, pp. 681–688. Omnipress.

https://doi.org/10.1017/jpr.2026.10109 Published online by Cambridge University Press