Journal of Machine Learning Research 22 (2021) 1-30 Submitted 4/19; Revised 3/20; Published 7/21

Implicit Langevin Algorithms for Sampling From Log-concave Densities

Liam Hodgkinson LIAM.HODGKINSON@BERKELEY.EDU
Department of Statistics, UC Berkeley, Berkeley, CA, 94720, USA
International Computer Science Institute, Berkeley, CA, 94704, USA

Robert Salomone ROBERT.SALOMONE@QUT.EDU.AU
Centre for Data Science, Queensland University of Technology, Brisbane, QLD, 4001, Australia

Fred Roosta FRED.ROOSTA@UQ.EDU.AU
School of Mathematics and Physics, The University of Queensland, St Lucia QLD 4067, Australia
International Computer Science Institute, Berkeley, CA 94704, USA

Editor: Frank Wood

Abstract

For sampling from a log-concave density, we study implicit integrators resulting from θ-method discretization of the overdamped Langevin diffusion stochastic differential equation. Theoretical and algorithmic properties of the resulting sampling methods for $\theta \in [0, 1]$ and a range of step sizes are established. Our results generalize and extend prior works in several directions. In particular, for $\theta \geq 1/2$, we prove geometric ergodicity and stability of the resulting methods for all step sizes. We show that obtaining subsequent samples amounts to solving a strongly-convex optimization problem, which is readily achievable using one of numerous existing methods. Numerical examples supporting our theoretical analysis are also presented.

Keywords: sampling, MCMC, implicit integrators, Bayesian regression

1. Introduction

Effectively sampling from arbitrary unnormalized probability distributions is a fundamental aspect of the Monte Carlo method, and is central in Bayesian inference. The most common cases involve probability densities π with support on all of $\mathbb{R}^d$, which can be written in the unnormalized form as

$$\pi(\boldsymbol{x}) \propto \exp(-f(\boldsymbol{x})), \qquad \boldsymbol{x} \in \mathbb{R}^d.$$

The sampling problem concerns the construction of a set of points $\{\boldsymbol{X}_k\}$ whose empirical distribution approaches π in some appropriate sense. A standard approach is *Markov Chain Monte Carlo* (MCMC), in which approximate sampling from π is accomplished by simulating a π-ergodic Markov chain. By the ergodic theorem, this provides consistent Monte Carlo estimators for expectations involving the density π. The most popular approach to generate such a set of points is the *Metropolis-Hastings* algorithm (Hastings, 1970), which constructs a π-ergodic Markov chain by generating a proposal from a given transition kernel and implements an acceptance criterion for these proposals; see Robert and Casella (1999) for an overview of such methods. While geometric rates of convergence (geometric ergodicity) can be guaranteed in a wide variety of settings, performance is highly susceptible to the underlying proposal. However, the effectiveness of Metropolis-Hastings methods diminish in higher dimensions, as step sizes must be scaled inversely with dimension, making rapid exploration of the space unlikely; see for example, Roberts and Rosenthal (2001).

Many of these issues lie with the steadfast requirement of consistency: that the sample empirical distribution should *asymptotically* be the same as π. Ensuring this requirement in turn can result in incurring serious penalty to the mixing rate of the chain. However, when seeking a fixed (finite) number of samples, which is almost always the case in practice, consistency is not necessarily a decisive property. Therefore, it has recently become popular to consider rapidly converging Markov chains whose stationary distributions are only *approximations* to π with a bias of adjustable size (Dalalyan, 2017a,b; Wibisono, 2018; Cheng et al., 2018a; Cheng and Bartlett, 2018). While the resulting Monte Carlo estimator is no longer consistent, it will often have dramatically smaller variance. This is an example of a *bias-variance tradeoff*, where a biased method can require significantly less computational effort to reach the same mean-squared error as an asymptotically unbiased Metropolis-Hastings chain (Korattikara et al., 2014).

The most studied of these methods is the *unadjusted Langevin algorithm* (ULA), seen in Roberts and Tweedie (1996), which is constructed by considering the overdamped Langevin diffusion equation, given by the stochastic differential equation (SDE)

$$\boldsymbol{L}_0 \sim \pi_0, \qquad \mathrm{d}\boldsymbol{L}_t = -\frac{1}{2}\nabla f(\boldsymbol{L}_t)\mathrm{d}t + \mathrm{d}\boldsymbol{W}_t, \tag{1}$$

and employing the forward Euler integrator, also known as Euler–Maruyama approximation (Kloeden and Platen, 2013), to obtain iterates of the form

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \frac{h}{2}\nabla f(\boldsymbol{X}_k) + \sqrt{h}\boldsymbol{Z}_k. \tag{2}$$

Here, $\boldsymbol{W}_t$ denotes d-dimensional standard Brownian motion, π_0 is some arbitrary (possibly deterministic) initial distribution, each $\boldsymbol{Z}_k$ is an independent standard d-dimensional normal random vector, and h is the *step size* parameter representing the temporal mesh size of the Euler method. Since (2) is explicitly defined, it is often referred to as *explicit* Euler scheme. This main interest in (1) lies in the well known fact that, under certain mild conditions and regardless of π_0, for any $t > 0$ the distribution of $\boldsymbol{L}_t$ is absolutely continuous (so we may consider its density π_t on $\mathbb{R}^d$), and $\boldsymbol{L}_t$ is an ergodic Markov process with limiting distribution π, that is, $\pi_t(\boldsymbol{x}) \to \pi(\boldsymbol{x})$ as $t \to \infty$ for all $\boldsymbol{x} \in \mathbb{R}^d$ (Kolmogorov, 1937).

However, unlike the Langevin SDE, the distribution of samples obtained from ULA (2) will, generally speaking, not converge to π as $t \to \infty$. More precisely, ULA is an asymptotically *biased* sampling algorithm, with corresponding bias proportional to step size (temporal mesh size). Despite this, in situations where MCMC fails to perform well, for example, high-dimensional problems, ULA can provide approximate samples from the target density with acceptable accuracy (Durmus and Moulines, 2019).

The theoretical properties of ULA, including geometric ergodicity (Hansen, 2003; Roberts and Tweedie, 1996), and performance in high dimensions (Durmus and Moulines, 2019) are well understood. Of particular relevance to us is the recent work of Dalalyan (2017a), Dalalyan (2017b), and Durmus and Moulines (2017), concerning the stability of ULA. Although it does not possess a single technical definition, stability of stochastic processes is often well-understood conceptually—some common characterizations include non-evanescence and Harris recurrence (Meyn and Tweedie, 2012, p. 15). To establish stability, the aforementioned works develop theoretical guarantees in the form of error bounds on the 2-Wasserstein metric between iterates of ULA and the target distribution. Doing so gives conditions under which ULA is bounded in probability, which in turn implies non-evanescence (Meyn and Tweedie, 2012, Proposition 12.1.1), and Harris recurrence, of the corresponding Markov chain (Meyn and Tweedie, 2012, Theorem 9.2.2). The inexact case, where ∇f is approximated to within an absolute tolerance, is also considered (Dalalyan and Karagulyan, 2019). Some alternative unadjusted explicit methods have also been considered; these are usually derived using other diffusions whose stationary distributions can also be prescribed (Cheng et al., 2018b).

As a direct result of the explicit nature of the underlying discretization scheme, the main issue with ULA-type algorithms is that they are stable only up to a fixed step size, beyond which the chain is no longer ergodic. In fact, Roberts and Tweedie (1996) actively discourage the use of ULA for this reason, and show that ULA may be transient for large step sizes. Stability is an essential concept when designing and analyzing methods for the numerical integration of continuous-time differential equations (Ascher, 2008). In some cases, this step size must be taken extremely small to remain stable. This becomes a major hindrance to the performance of the method in practice. Drawing comparisons to the theory of ordinary differential equations (ODEs) by dropping the stochastic term, in these cases, the Langevin diffusion is said to be *stiff* (Ascher and Petzold, 1998). Ill-conditioned problems, such as sampling from any multivariate normal distribution with a covariance matrix possessing a large condition number (Golub and Van Loan, 2012, §2.6.2), are likely to induce a stiff Langevin diffusion (Lambert, 1991, §6.2). The negative side-effects associated with ill-conditioning as well as the restrictions on step size are often only exacerbated in high-dimensional problems.

In this light, a natural alternative to using explicit schemes with careful choice of step size is to consider *implicit* variants. From the established theory of numerical solutions of ODEs, it is well-known that implicit integrators have larger regions of stability than explicit alternatives, that is, one can take larger steps without unboundedly magnifying the underlying discretization errors (Ascher, 2008). Motivated by this, we can instead consider the θ-method scheme (Ascher, 2008, p. 84), which when applied to Langevin dynamics (1), yields general iterations of form

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \frac{h}{2}\left[\theta \nabla f(\boldsymbol{X}_{k+1}) + (1-\theta)\nabla f(\boldsymbol{X}_k)\right] + \sqrt{h}\boldsymbol{Z}_k, \tag{3}$$

for some $\theta = [0,1]$. The special cases of $\theta = 0, 1$ and $1/2$ correspond to forward, backward, and trapezoidal integrators, respectively. Of course, for $\theta = 0$, (3) reduces to the explicit Euler scheme (2). As the choice of $\theta \in (0,1]$ define the endpoint in an implicit way, such integrators are often referred to as *implicit*.

To our knowledge, there have only been a handful of efforts to study the properties of sampling algorithms obtained from such implicit schemes. A universal analysis of sampling schemes based on general Langevin diffusion was conducted in Mattingly et al. (2002). There, it was shown that the implicit Euler scheme, and other numerical methods satisfying a certain minorization condition are geometrically ergodic for sufficiently small step sizes under the assumption that f is 'essentially quadratic'. In a more focused analysis, Casella et al. (2011) investigated the ergodic properties of a few implicit schemes (including the θ-method scheme) for a restricted family of one-dimensional super-Gaussian target densities. They found that, in this setting, the θ-method results in a geometrically ergodic chain for *any* step size $h > 0$, provided that $\theta \geq 1/2$, and suggested the same might be true in higher dimensions. Under slightly weaker assumptions than strong convexity, Kopec (2014) conducted a weak backward error analysis providing error bounds on the expectation of the fully implicit Euler scheme ($\theta = 1$) with respect to suitable test functions. More recently, Wibisono (2018) considered the $\theta = 1/2$ case and provided a rate of convergence of the scheme towards its biased stationary distribution under the 2-Wasserstein metric, assuming strong convexity and small step sizes. Despite these efforts, it is still unclear how implicit schemes compare with explicit schemes more generally for large step sizes, and what the effect of θ is on the bias of the method.

The aim of this work is to study the θ-method sampling scheme (3) for all $\theta \in (0, 1]$, as it applies to the relevant case of strongly log-concave distributions, particularly when f is a strongly convex function and ∇f is Lipschitz continuous. Such distributions arise frequently in Bayesian regression problems (Bishop and Tipping, 2003), for example generalized linear models (GLMs) with a Gaussian prior (Chatfield et al., 2010).

1.1 Contributions

To those ends, the contributions of this work are as follows:

1. We show that the transition density associated with (3) has a closed form solution. Then, using this, we establish conditions for geometric ergodicity, in terms of θ, the step size h, Lipschitz continuity, tail behaviour, and semi-convexity of f (Theorem 1). By doing so, we show stability of the θ-method scheme in multivariate settings for *any* step size when $\theta \geq 1/2$ and f is strongly convex, proving the conjecture by Casella et al. (2011).

2. We provide non-asymptotic theoretical guarantees for long-time behaviour of (3), which extend those of Dalalyan and Karagulyan (2019, Theorem 1) to the general implicit case (Theorem 2).

3. As for $\theta > 0$, iterations of (3) involve solving a non-linear equation, we study the effect of inexact solutions of the underlying sub-problems. We propose practically computable termination criteria and quantify the effect of approximating each iterate on the convergence rate and long-term bias of the chain with this criterion.

4. We establish large step size asymptotics for $\theta > 0$ via a central limit theorem as $h \to \infty$ (Theorem 5). As a consequence, we develop an effective default heuristic choice of step size.

5. Finally, we demonstrate the empirical performance of the implicit θ-method scheme in a series of numerical experiments; namely sampling from high-dimensional Gaussian distributions, and the posterior density of a Bayesian logistic regression problem involving a real data set.

Proofs of all results can be found in Appendix A.

Notation. In the sequel, vectors and matrices are denoted by bold lowercase and Romanized bold uppercase letters, for example, $\boldsymbol{v}$ and $\mathbf{V}$, respectively. We denote the identity matrix by $\mathbf{I}$. Regular lower-case and upper-case letters, such as s m and M, are used to denote scalar constants. Random vectors are denoted by italicized bold uppercase letters, such as $\boldsymbol{X}$. For two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ indicates that $\mathbf{A} - \mathbf{B}$ is symmetric positive semi-definite. For vectors, we let $\| \cdot \|$ denote the Euclidean norm of appropriate dimension, and $\| \cdot \|_{L^2}$ denote the L^2 norm acting on random vectors, that is, $\|\boldsymbol{X}\|_{L^2}^2 \coloneqq \mathbb{E}\|\boldsymbol{X}\|^2$. For matrices, $\| \cdot \|_2$ denotes the spectral norm.

2. Implicit Langevin Algorithm (ILA)

In this section, we establish conditions under which the sequence of θ-method iterates (3) form a Markov chain that is geometrically ergodic. For this, we impose the following assumption on the smoothness of f, which ensures the existence of a unique solution to (1); see Ikeda and Watanabe (2014, Theorem 2.4–3.1).

Assumption 1 *The function $f \in \mathcal{C}^2$ (it is twice continuously differentiable), and ∇f is M-Lipschitz, that is,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \textit{for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Under Assumption 1, Dalalyan (2017b) shows that if $h < 4/M$, iterations of ULA (2) are bounded in probability (in other words, the algorithm is *stable*). However, this restriction is a fundamental disadvantage of ULA. If M is particularly large, as might be the case in ill-conditioned problems and in high dimensions where the ULA is commonly applied, then the step size must be taken very small, which results in slow mixing time of the chain and high autocorrelation of the samples. In sharp contrast, we now show that for appropriate choice of θ, (3) does not suffer from this restriction.

The process of obtaining samples by iterating (3) is outlined in Algorithm 1. For brevity, we henceforth refer to this procedure as the implicit Langevin algorithm (ILA). Note that (3) can be rewritten as

$$(\mathcal{I} + \tfrac{1}{2}h\theta\nabla f)(\boldsymbol{X}_{k+1}) = \boldsymbol{X}_k - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{X}_k) + \sqrt{h}\boldsymbol{Z}_k,$$

where $\mathcal{I}$ denotes the identity mapping. Assuming that $\mathcal{I} + \frac{1}{2}h\theta\nabla f$ is a strictly monotone operator, that is $\mathcal{I} + \frac{1}{2}h\theta\nabla f$ is globally invertible, (3) admits a unique solution as

$$\boldsymbol{X}_{k+1} = (\mathcal{I} + \tfrac{1}{2}h\theta\nabla f)^{-1}\left[\boldsymbol{X}_k - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{X}_k) + \sqrt{h}\boldsymbol{Z}_k\right]. \tag{4}$$

Conditions under which the procedure (4) is guaranteed to be well-defined are discussed in §2.1; see Rockafellar (1976) for a thorough treatment of monotone operators and their application in proximal point algorithms. For the time being, it is also assumed in Algorithm 1 that (3) can be solved *exactly.* The discussion of inexact solutions is relegated to §3.

2.1 Theoretical analysis of Algorithm 1

In this section, we establish sufficient conditions for the geometric ergodicity of the sequence of iterates generated from Algorithm 1. To conduct such an analysis, we require the transition kernel density $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ induced from (3). In general, this is only implicitly defined,

Algorithm 1: Implicit Langevin Algorithm (ILA)

```
Input : - Initial value $\boldsymbol{X}_0 = \boldsymbol{x}_0 \in \mathbb{R}^d$
        - Number of samples $n$
        - Step size $h > 0$
        - $\theta$-method parameter $\theta \in (0, 1]$
for $k = 0, 1, \ldots, n$ do
    Draw $\boldsymbol{Z}_k \sim \mathcal{N}(0, \mathbf{I})$
    Solve (3) to obtain $\boldsymbol{X}_{k+1}$
end
```

however, assuming $\mathcal{I} + \frac{1}{2}h\theta\nabla f$ is globally invertible, $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ is nevertheless available in closed form. Assuming that $f \in \mathcal{C}^2$, this is true whenever h is chosen so that

$$\mathbf{I} + \frac{h\theta}{2}\nabla^2 f(\boldsymbol{x}) \succ 0, \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^d. \tag{5}$$

Therefore, at the very least, for (3) to be well-defined as a sampling method, we require f to be *semi-convex*, that is, there exists some $\gamma > 0$ such that $\nabla^2 f(\boldsymbol{x}) + \gamma\mathbf{I}$ is positive-semidefinite for all $\boldsymbol{x} \in \mathbb{R}^d$. For example, under Assumption 1, f is M-semi-convex and (5) holds if $h < \frac{2}{\theta M}$. This restriction on step size can be removed entirely if f is assumed to be convex.

From (4), for fixed $\boldsymbol{x} \in \mathbb{R}^d$, note that $p(\,\cdot \,|\, \boldsymbol{x})$ is the probability density function of the random variable

$$\boldsymbol{Y} = (\mathcal{I} + \tfrac{1}{2}h\theta\nabla f)^{-1}(\boldsymbol{x} - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{x}) + \sqrt{h}\boldsymbol{Z}),$$

where $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. As this is an invertible transformation of a standard Gaussian random vector, by the change of variables theorem (see for example, Shao (2008, Proposition 1.8)), we have

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \left|\det\left(\mathbf{I} + \frac{h\theta}{2}\nabla^2 f(\boldsymbol{y})\right)\right| \phi\left(\boldsymbol{y} + \frac{h\theta}{2}\nabla f(\boldsymbol{y})\,;\, \boldsymbol{x} - \frac{h(1-\theta)}{2}\nabla f(\boldsymbol{x}),\ h\mathbf{I}\right), \tag{6}$$

where $\phi(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and 'det' denotes the determinant of a matrix.

It can be seen from (6) that increasing θ (or h when $\theta > 0$) alters the landscape of the transition density, and the shape of its level-sets. To illustrate this, Figure 1 depicts the contour plots of the transition kernel (6) for an anisotropic example problem with differing θ and initial state for the same step size. It can be seen that the case with $\theta = 0$ (ULA) results in an isotropic proposal in all situations, whereas other choices of θ (implicit methods) yield proposal densities that can better adapt to the anisotropic target density.

With an explicit expression for the transition density, we can investigate the stability of the iterates given by Algorithm 1. The most convenient way of doing this is by demonstrating geometric ergodicity of the chain induced by the transition kernel (6). Recall that a Markov chain with n-fold transition kernel $p_n(\boldsymbol{y}|\boldsymbol{x})$ is said to be *geometrically ergodic*

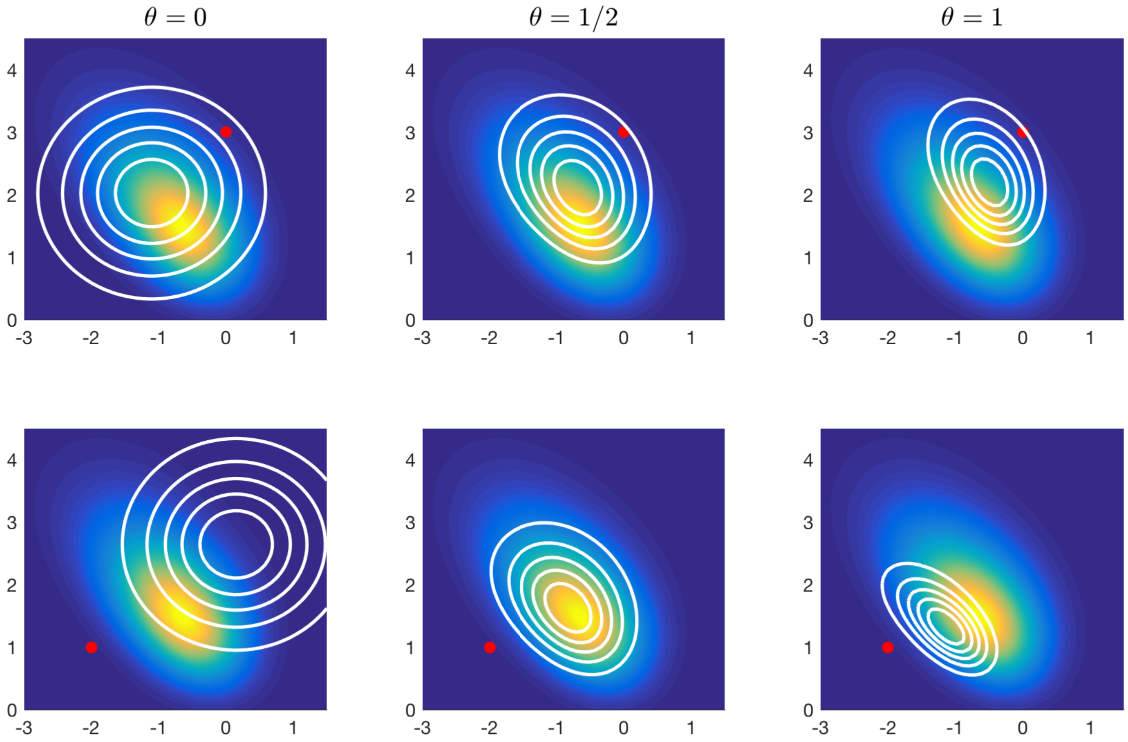

Figure 1: Transition kernel for an example density and large step size on a two–dimensional Bayesian Logistic Regression example for large h. While the traditional $\theta = 0$ case (ULA) imposes isotropic proposals, for other choices of θ, the proposal density adapts to the anisotropic target density.

toward an invariant density $\nu(\cdot)$ if there exist constants $C > 0$ and $0 < \rho < 1$ such that

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \int_{\mathbb{R}^d} |p_n(\boldsymbol{y}\,|\,\boldsymbol{x}) - \nu(\boldsymbol{y})| \mathrm{d}\boldsymbol{y} \le C\rho^n, \qquad \text{for all } n = 1, 2, \ldots.$$

Similarly, a diffusion process $\boldsymbol{X}_t$ with transition kernel density $p_t(\cdot\,|\,\cdot)$ is said to be *exponentially ergodic* towards an invariant density $\nu(\boldsymbol{x})$ if there exist constants $C, \lambda > 0$ such that

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \int_{\mathbb{R}^d} |p_t(\boldsymbol{y}\,|\,\boldsymbol{x}) - \nu(\boldsymbol{y})| \mathrm{d}\boldsymbol{y} \le Ce^{-\lambda t}, \qquad \text{for all } t > 0.$$

It was shown in Hansen (2003, Eqn. (12)) that the overdamped Langevin diffusion (1) is exponentially ergodic provided the following assumption on f holds:

Assumption 2 $m := \liminf_{\|\boldsymbol{x}\|\to\infty} \dfrac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} > 0.$

Intuitively, Assumption 2 imposes super-Gaussian tails of the target distribution. Under Assumptions 1 and 2, for any $\boldsymbol{x}$ and $\boldsymbol{y}$, there exists a constant $c(\boldsymbol{y}) \geq 0$ depending on $\boldsymbol{y}$, such that (Lemma 6)

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}),\, \boldsymbol{x} - \boldsymbol{y} \rangle \geq m\|\boldsymbol{x} - \boldsymbol{y}\|^2 - c(\boldsymbol{y}), \qquad \text{for every } \boldsymbol{x} \in \mathbb{R}^d. \tag{7}$$

Assumption 2 is not new, having also appeared in Kopec (2014), and appears to be among the weakest assumptions one can make to effectively study these implicit schemes. Clearly, (7) is a significantly weaker condition than strong convexity:

Assumption 3 *The function $f \in \mathcal{C}^2(\mathbb{R}^d)$ is m-strongly convex, that is, there exists $0 < m < \infty$ such that*

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq m\|\boldsymbol{x} - \boldsymbol{y}\|^2, \qquad \textit{for any } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

In fact, it is straightforward to show that if Assumption 2 holds, then m is necessarily common in the two inequalities. Furthermore, we remark that under Assumptions 1 and 3, the spectrum of every Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is controlled to be within $[m, M] \subset (0, \infty)$. Strong convexity is quite a natural assumption in Bayesian regression problems, as it can be guaranteed for the class of GLMs with Gaussian priors (DasGupta, 2011).

Under Assumptions 1 and 2, we can now establish geometric ergodicity of the θ-method scheme under certain conditions on θ and h (Theorem 1).

Theorem 1 *For f satisfying Assumptions 1 and 2 that is γ-semi-convex, the iterates of the θ-method scheme with associated transition kernel (6) form a geometrically ergodic chain when $h < \frac{2}{\theta\gamma}$ provided we also have either $\theta \geq 1/2$, or both $\theta < 1/2$ and $h < 4m/[M^2(1-2\theta)]$.*

While Theorem 1 establishes the geometric ergodicity of the chain towards *some* stationary distribution, in general, that distribution need not necessarily be π. Nevertheless, under Assumptions 1 and 3, we have established that the θ-method discretization of the overdamped Langevin equation is stable for *any* step size, provided $\theta \geq 1/2$. For these choices of θ, this implies that ILA is less strict about step size tuning than ULA. As will be seen in §5, this will prove to have a profound effect on the performance of ILA relative to ULA on high-dimensional problems.

2.2 Asymptotic exactness for the normal distribution

Among all values for θ, we draw special attention to the choice $\theta = 1/2$. This resulting integrator, also known as the trapezoidal scheme, is known to be *second-order accurate* when applied to ODEs; that is, for a quadratic function F, iterates of the trapezoidal scheme for solving $y' = F(y)$ yield points of the *exact* solution (Süli and Mayers, 2003, §12.4). An important consequence of this is that the global error incurred in the trapezoidal scheme is $\mathcal{O}(h^2)$ as $h \to 0$. Unfortunately, as a consequence of the Itô calculus, this property does not hold for numerical solutions of stochastic differential equations. The construction of second-order schemes which are exact for quadratic f in finite time and exhibit $\mathcal{O}(h^2)$ global error generally require careful treatment of the stochastic term—see for example Anderson and Mattingly (2011), or the Ozaki local linearization scheme (Biscay et al., 1996). However, when $\theta = 1/2$, the notion of second-order accuracy itself carries over to ILA in a rather remarkable way.

The case where f is a quadratic form corresponds to sampling from a (multivariate) normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$f(\boldsymbol{x}) = \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}). \tag{8}$$

It is easy to see that, in this particular setting, (3) becomes explicitly solvable. Indeed, letting $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$, we see that $\nabla f(\boldsymbol{x}) = \mathbf{Q}(\boldsymbol{x} - \boldsymbol{\mu})$, and so

$$\boldsymbol{X}_{k+1} = \left(\mathbf{I} + \frac{h\theta}{2}\mathbf{Q}\right)^{-1} \left[\left(\mathbf{I} - \frac{h(1-\theta)}{2}\mathbf{Q}\right)(\boldsymbol{X}_k - \boldsymbol{\mu}) + \sqrt{h}\boldsymbol{Z}_k\right] + \boldsymbol{\mu}. \tag{9}$$

Observe that if $\boldsymbol{X}_0$ is chosen to be a fixed value, all of the iterates $\boldsymbol{X}_k$ are normally distributed. As a consequence of Lévy continuity, the stationary distribution of the ILA, if it exists, must also be normally distributed. In particular, due to (9), it must have mean $\boldsymbol{m}$ and covariance $\mathbf{V}$ satisfying

$$\begin{aligned} \boldsymbol{m} - \boldsymbol{\mu} &= (\mathbf{I} + \tfrac{1}{2}h\theta\mathbf{Q})^{-1}(\mathbf{I} - \tfrac{1}{2}h(1-\theta)\mathbf{Q})(\boldsymbol{m} - \boldsymbol{\mu}) \\ \mathbf{V} &= (\mathbf{I} + \tfrac{1}{2}h\theta\mathbf{Q})^{-2}[(\mathbf{I} - \tfrac{1}{2}h(1-\theta)\mathbf{Q})^2\mathbf{V} + h\mathbf{I}]. \end{aligned}$$

Since $\mathbf{Q} \neq \mathbf{0}$, it must be the case that $\boldsymbol{m} = \boldsymbol{\mu}$. Solving for $\mathbf{V}$, the stationary distribution of the ILA is found to be

$$\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\left(\mathbf{I} + \tfrac{1}{2}h(\theta - \tfrac{1}{2})\mathbf{Q}\right)^{-1}\right).$$

Here we encounter the remarkable fact that when f is quadratic and $\theta = 1/2$, regardless of the step size chosen, ILA is *asymptotically unbiased!* To our knowledge, this was first observed in Wibisono (2018), however, as a consequence of our analysis, we can now deduce that $\theta = 1/2$ is the *only* choice of θ that yields this property. While asymptotic exactness is unlikely to hold for other sampling problems, it suggests that cases involving approximately quadratic f should see near optimal performance when $\theta = 1/2$.

3. Inexact Implicit Langevin Algorithm (i-ILA)

It is clear that the utility of ILA is dependent on the solvability of (3). Fortunately, this is made feasible by a useful reinterpretation of solutions to (3) as those of a corresponding optimization problem. Indeed, the inverse operator $(\mathcal{I} + \frac{1}{2}h\theta\nabla f)^{-1}$ is quite commonly considered in convex optimization, as it is equivalent to the proximal operator $\mathbf{prox}_{\frac{1}{2}h\theta f}$ defined by

$$\mathbf{prox}_f(\boldsymbol{v}) = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{v}\|^2 \right\}, \qquad \boldsymbol{v} \in \mathbb{R}^d.$$

This equivalence follows from that of an optimization problem and the root-finding problem for its critical values (Parikh and Boyd, 2014, Eqn. (3.4)). Therefore, (3) can be formulated as the following optimization problem (after rescaling by $2/h$):

$$\boldsymbol{X}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^d} F(\boldsymbol{x}; \boldsymbol{X}_k, \boldsymbol{Z}_k), \tag{10a}$$

where

$$F(\boldsymbol{x};\boldsymbol{y},\boldsymbol{z}) := \theta f(\boldsymbol{x}) + \frac{1}{h}\left\|\boldsymbol{x} - \left(\boldsymbol{y} - \frac{h(1-\theta)}{2}\nabla f(\boldsymbol{y}) + \sqrt{h}\boldsymbol{z}\right)\right\|^2. \tag{10b}$$

This reinterpretation of (3) was also noted in Wibisono (2018), although only the $\theta = 1/2$ case was considered. Iterations of the form (10) are often referred to as *proximal-point methods* in the optimization literature (Combettes and Pesquet, 2011; Parikh and Boyd, 2014). We remark that proximal operators were used in Pereyra (2016) in the construction of a *proximal unadjusted Langevin algorithm* (P-ULA). In fact, iterates of their P-ULA algorithm would correspond with (10) when $\theta = 1$ and if the Gaussian term $\sqrt{h}\boldsymbol{Z}_k$ were to be moved outside the proximal operator. As one might expect, this discrepancy has a significant impact on the covariance of each proposal as $h \to \infty$; it will be shown in Theorem 5 that the asymptotic covariance of these proposals is generally anisotropic.

The implementation of Algorithm 1 now hinges entirely on our ability to solve the subproblem (10). For the unadjusted Langevin algorithm where $\theta = 0$, this can be done trivially through a closed form solution. However, for $\theta > 0$, we generally have to resort to an iterative optimization scheme to solve (10). Thus far, we have assumed that the optimization problem in (10) can be solved exactly. However, more often than not this is infeasible, and one must instead consider the effects of *approximate* solutions of (10) in the overall convergence of the chain. This results in a sampling variant, which is henceforth referred to as i-ILA (for inexact ILA).

The most natural way of doing this is by measuring the error in the corresponding root-finding problem (3) via the norm of the gradient of the subproblem (10b), $\|\nabla F\|$. This is ideal because not only can it be readily computed in practice, but also the termination criterion of many iterative optimization algorithms is based on this norm falling below a given tolerance; for example, see Nocedal and Wright (2006). Furthermore, efficient algorithms for directly minimizing $\|\nabla F\|$, as a surrogate function for optimization of F, have been recently proposed, which enjoy linear, that is, geometric, convergence rates, even in the absence of smoothness or convexity of F (Roosta et al., 2018). In addition, for sampling in distributed computational environments, such as when large-scale data cannot be stored on a single machine, distributed variants of these surrogate optimization algorithms have also been recently considered (Crane and Roosta, 2019). These algorithms are particularly suitable as part of i-ILA since they are guaranteed to rapidly and monotonically decrease $\|\nabla F\|$; recall that $\|\nabla F\|$ need not be monotonically decreasing in optimization algorithms that optimize F directly.

With this in mind, we consider an inexact modification of Algorithm 1, shown in Algorithm 2, for approximate sampling from π.

3.1 Theoretical analysis of Algorithm 2

The increased stability offered by Algorithm 1 has been established in Theorem 1. However, while Theorem 1 guarantees rapid convergence towards *some* stationary distribution, closeness of the θ-method iterates to the target distribution π and the effect of increasing h on its bias as a sampling method, has yet to be established. Furthermore, Algorithm 1 and its guarantees given by Theorem 1 require exact solutions of the root-finding problem

Algorithm 2: Inexact Implicit Langevin Algorithm (i-ILA)

```
Input : - Initial value $\boldsymbol{X}_0 = \boldsymbol{x}_0 \in \mathbb{R}^d$
        - Number of samples $n$
        - Step size $h > 0$
        - $\theta$-method parameter $\theta \in [0, 1]$
        - Sub-problem inexactness tolerance $\epsilon \geq 0$
for $k = 0, 1, \dots, n$ do
    Draw $\boldsymbol{Z}_k \sim \mathcal{N}(0, \mathbf{I})$
    Find $\boldsymbol{X}_{k+1}$ satisfying $\|\nabla F(\boldsymbol{X}_{k+1}; \boldsymbol{X}_k, \boldsymbol{Z}_k)\| \leq \epsilon$, where $F$ is defined in (10b)
end
```

(3), whereas Algorithm 2 allows for such problems to be solved only inexactly. To address both of these problems, we devote this section to the development of theoretical guarantees of Algorithm 2, inspired by the techniques of Dalalyan (2017b). These guarantees come in the form of rate of convergence estimates under the 2-Wasserstein metric, defined between two probability measures ν and π by

$$W_2(\nu, \pi) = \inf_{\boldsymbol{X} \sim \nu, \boldsymbol{Y} \sim \pi} \|\boldsymbol{X} - \boldsymbol{Y}\|_{L_2}$$

where the infimum is taken over all couplings $(\boldsymbol{X}, \boldsymbol{Y})$ of ν and π, and is attained by some *optimal* coupling (Villani, 2008, Thm. 4.1). The 2-Wasserstein metric can be readily linked to other quantities of interest. For example, from the Kantorovich-Rubinstein formula (Villani, 2008, Eqn. (5.11)), for any M-Lipschitz function φ, we have that

$$|\nu(\varphi) - \pi(\varphi)| := \left| \int \varphi \, \mathrm{d}(\nu - \pi) \right| \leq M W_2(\nu, \pi).$$

Our guarantees will require the same assumptions on f seen in Dalalyan (2017b), that are Assumptions 1 and 3. Under these assumptions, the condition number of F in (10b) can be written as

$$\kappa_h := \frac{1 + \frac{1}{2}\theta h M}{1 + \frac{1}{2}\theta h m}. \tag{11}$$

Recall that the condition number (11) encodes and summarizes the curvature (the degree of relative flatness and steepness), of the graph of F. In optimization, it is well-known that a large condition number typically amounts to a more difficult problem to solve, and hence algorithms that can take such contorted curvature into account (Newton-type methods, for example), are more appropriate (Roosta-Khorasani and Mahoney, 2018; Xu et al., 2017). It is only natural to anticipate that challenges corresponding to problem ill-conditioning similarly carry over to sampling procedures as well. Indeed, large ratios of M/m, which imply increasingly anisotropic level-sets for f, can hint at more difficult sampling problems. For example, this difficulty directly manifest itself in ill-conditioning of F, which in turn results in more challenging sub-problems. Furthermore, in such situations, taking a larger step size can only exacerbate the ill-condition of F. As a result, similar to the role played by second-order methods in optimization, one can naturally expect to see implicit methods to be more appropriate for ill-conditioned sampling problems.

Under Assumptions 1 and 3, the discrepancy between the inexact variant of the θ-method given in Algorithm 2 and the target density π under the 2-Wasserstein metric is described in Theorem 2.

Theorem 2 *Suppose f satisfies Assumptions 1 and 3. Let $\theta \in (0,1]$ and let ν_t denote the distribution of the iterate $\boldsymbol{X}_t$ obtained by Algorithm 2, for each $t \geq 1$, starting from $\boldsymbol{X}_0 \sim \nu_0$. Let κ_h be as in* (11)*, and if $\theta < 1$, let*

$$h^* = \frac{(\theta - \frac{1}{2})(M+m) + \sqrt{(\theta - \frac{1}{2})^2(M+m)^2 + 4\theta(1-\theta)mM}}{\theta(1-\theta)mM}. \tag{12}$$

Furthermore,

(i) if $h \leq h^*$ or $\theta = 1$, then let

$$\rho = \frac{1 - \frac{1}{2}h(1-\theta)m}{1 + \frac{1}{2}h\theta m}, \quad \text{and} \quad C = \frac{\kappa_h}{m}; \tag{13}$$

(ii) alternatively, if $\theta < 1/2$ and $h^* < h < \frac{4}{M(1-2\theta)}$, or if $1/2 \leq \theta < 1$ and $h > h^*$, then let

$$\rho = \frac{\frac{1}{2}h(1-\theta)M - 1}{\frac{1}{2}h\theta M + 1}, \quad \text{and} \quad C = \frac{\frac{1}{2}\kappa_h^2 h}{2 + \frac{1}{2}h(2\theta - 1)M}. \tag{14}$$

Then, for any $t \in \mathbb{N}$,

$$W_2(\nu_t, \pi) \leq \kappa_h \rho^t W_2(\nu_0, \pi) + C\left(\epsilon + \min\left\{M\sqrt{hd}(2 + \sqrt{hM}), 2\sqrt{Md}\right\}\right). \tag{15}$$

Remark 3 *As $\theta \to 0$, for the transition point h^* we have $h^* \to \frac{4}{M+m}$. Moreover, at $\theta = 0$ and $\epsilon = 0$,* (15) *coincides with Dalalyan and Karagulyan (2019, Theorem 1), up to a different constant in the bias term. To see this, observe that $\theta = 0$ implies $\kappa_h = 1$, $\rho = 1 - \frac{1}{2}hm$, $C = m^{-1}$ when $h \leq \frac{4}{M+m}$, and $\rho = \frac{1}{2}hM - 1$, $C = \frac{h}{4-hM}$ when $\frac{4}{M+m} < h < \frac{4}{M}$. This gives*

$$W_2(\nu_t, \pi) \leq \begin{cases} (1 - \frac{1}{2}hm)^t W_2(\nu_0, \pi) + \frac{4M}{m}\sqrt{hd} \\ (\frac{1}{2}hM - 1)^t W_2(\nu_0, \pi) + \frac{4Mh}{4-hM}\sqrt{hd}. \end{cases}$$

Theorem 2 may thus be seen as a generalization of Dalalyan and Karagulyan (2019, Theorem 1) to arbitrary $\theta \in [0,1]$ and error ϵ.

Remark 4 *In the noteworthy case of $\theta = 1/2$, Theorem 2 implies*

$$W_2(\nu_t, \pi) \leq \frac{1 + \frac{1}{4}hM}{1 + \frac{1}{4}hm} \cdot \begin{cases} \left(\frac{1 - \frac{1}{4}hm}{1 + \frac{1}{4}hm}\right)^t W_2(\nu_0, \pi) + \frac{\epsilon + M\sqrt{hd}(2 + \sqrt{hM})}{m} & \text{if } h \leq \frac{4}{\sqrt{mM}} \\ \left(\frac{\frac{1}{4}hM - 1}{\frac{1}{4}hM + 1}\right)^t W_2(\nu_0, \pi) + \frac{1}{2}h \cdot \left(\frac{1 + \frac{1}{4}hM}{1 + \frac{1}{4}hm}\right)\left(\frac{\epsilon}{2} + \sqrt{Md}\right) & \text{otherwise.} \end{cases}$$

As Theorem 1 did for Algorithm 1, Theorem 2 shows that for $\theta \geq 1/2$, Algorithm 2 is stable for all $h > 0$. Theorem 2 does suggest that smaller values of θ will achieve faster convergence rates and smaller biases for small step sizes, although this does not appear to be the case in practice (for example, refer to §5). Observe that, for $\theta > 1/2$ and fixed h, the bias term is in the order of $\mathcal{O}(M^{-1/2})$. This implies that increasing the condition number when m is bounded below (for example the spherical Gaussian prior in Bayesian regression) results in smaller bias and faster convergence. This is in sharp contrast to ULA whose performance significantly degrades with increasing condition number in such settings.

Also, we would like to reiterate that, in stark contrast to what is observed in Roberts and Rosenthal (2001) for Metropolis-Hastings algorithms, the rate of convergence in Theorem 2 for Algorithm 2 is *not* dependent on the dimension d in any form other than through the appearances of m and M. The dimension appears in the bias term simply due to the natural expansion of the Euclidean distance with dimension. In particular, following Durmus and Moulines (2019), as the dependence on dimension is at most polynomial, this lends credence to the claim that implicit Langevin methods are well-equipped to handle high-dimensional sampling problems.

4. Asymptotics for large step size

While Theorem 2 provides an essential description of the behavior of Algorithm 2, the bounds presented there are tightest for smaller step sizes on the order of $1/M$, and are less effective when h is larger. Unfortunately, the most useful applications of ILA will occur when M is large, and so the small step size ($h \to 0$) regime will not be all that relevant. Enabled by the increased stability of ILA, we present a novel analysis of Algorithm 1 by establishing a central limit-type theorem regarding asymptotic behavior of the iterates in the $h \to \infty$ regime.

Before we begin with a formal analysis, we are able to obtain insight by considering the behavior of the subproblem (10) as $h \to \infty$. For $h \gg 1$, we have

$$\frac{1}{h}\left\|\boldsymbol{x} - \boldsymbol{x}_t + \frac{h(1-\theta)}{2}\nabla f(\boldsymbol{x}_t) + \sqrt{h}\boldsymbol{z}_t\right\|^2 = (1-\theta)\boldsymbol{x}\cdot\nabla f(\boldsymbol{x}_t) + \mathcal{O}(h^{-1/2}) + C,$$

where C does not depend on $\boldsymbol{x}$, and so does not contribute to solving (10). As a result, the iterates of the θ method in the $h \to \infty$ regime will satisfy the relations $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t^\theta$, where we let

$$\nabla f(\boldsymbol{x}_t^\theta) = \left(1 - \frac{1}{\theta}\right)\nabla f(\boldsymbol{x}_t), \qquad \boldsymbol{x} \in \mathbb{R}^d. \tag{16}$$

Letting $\boldsymbol{x}^*$ denote the unique mode of π, iterating (16) gives

$$\|\nabla f(\boldsymbol{x}_t) - \nabla f(\boldsymbol{x}^*)\| = \rho^t \|\nabla f(\boldsymbol{x}_0) - \nabla f(\boldsymbol{x}^*)\|,$$

where $\rho = \frac{1}{\theta} - 1$. Under Assumption 3, we obtain

$$\frac{m\rho^t}{M}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}^*\| \leq \frac{M\rho^t}{m}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|.$$

Therefore, the behavior of the θ-method for large h is determined according to the three regimes depicted in Table 1. The $\theta < 1/2$ case is clearly undesirable from a practical

Table 1: Asymptotic behavior of iterates of (10) as $h \to \infty$

$0 \leq \theta < 1/2$	$\rho > 1$	$\|\boldsymbol{X}_t\| \to \infty$ (unbounded in probability)
$\theta = 1/2$	$\rho = 1$	iterates oscillate about the mode
$1/2 < \theta \leq 1$	$\rho < 1$	$\boldsymbol{X}_t \to \boldsymbol{x}^*$ (collapse to the mode)

standpoint. Moreover, the collapse towards the mode seen when θ is close to one suggests enormous potential bias for large step sizes. On the other hand, the $\theta = 1/2$ case provides no damping effect whatsoever (a fact also supported by Theorem 2), making it susceptible to rare large proposals. Based on this preliminary analysis, for some small $\epsilon > 0$, a choice of $\theta = 1/2 + \epsilon$ appears to provide the safest, and potentially the most accurate of our θ-method samplers. This aligns with the rule-of-thumb used for θ-method discretization of ODEs (Ascher, 2008, p. 85). To formally extend these characterizations to the implicit θ-method scheme (3), in Theorem 5, a central limit theorem as $h \to \infty$ is obtained for a single step of the scheme about the deterministic map $\boldsymbol{x} \mapsto \boldsymbol{x}^\theta$.

Theorem 5 *Given any $f \in \mathcal{C}^2(\mathbb{R}^d)$, consider iterations given by* (10)*, where $\theta \in (0,1]$. Conditioned on $\boldsymbol{X}_k$, we have*

$$\sqrt{h}(\boldsymbol{X}_{k+1} - \boldsymbol{X}_k^\theta) \xrightarrow[h\to\infty]{\mathcal{D}} \mathcal{N}\left(\boldsymbol{0}, \frac{4}{\theta^2}\nabla^2 f(\boldsymbol{X}_k^\theta)^{-2}\right).$$

Theorem 5 implies that as $h \to \infty$, the implicit θ-method scheme behaves similarly to a Markov chain $\{\boldsymbol{W}_k\}$ with transitions

$$\boldsymbol{W}_{k+1} = \boldsymbol{W}_k^\theta + \frac{2}{\theta\sqrt{h}}\nabla^2 f(\boldsymbol{W}_k^\theta)^{-1}\boldsymbol{Z}_k,$$

whose dynamics mimic those of the map $\boldsymbol{x} \mapsto \boldsymbol{x}^\theta$, but with an additional normally-distributed noise term at each step. Furthermore, the variance of this noise term increases as the implicit component of the scheme diminishes (taking $\theta \to 0$). Despite the unusual $h \to \infty$ regime, we have found that, for typical step sizes, Theorem 5 provides a surprisingly accurate description of the transition dynamics of ILA when $\theta \geq 1/2$, and is ideal for developing heuristics.

4.1 A heuristic choice for step size

A consequence of the proof of Theorem 5 is that the covariance $\boldsymbol{\Sigma}_h(\boldsymbol{x})$ of the proposal density from the transition kernel $p(\boldsymbol{y} \,|\, \boldsymbol{x})$ behaves asymptotically as

$$\boldsymbol{\Sigma}_h(\boldsymbol{x}) \approx h\left(\mathbf{I} + \frac{h\theta}{2}\nabla^2 f(\boldsymbol{x}^\theta)\right)^{-2}, \qquad \text{as } h \to \infty.$$

Conversely, applying the inverse function theorem to (4) reveals a linear approximation about $h = 0$, and hence

$$\boldsymbol{\Sigma}_h(\boldsymbol{x}) \approx h\left(\mathbf{I} + \frac{h\theta}{2}\nabla^2 f(\boldsymbol{x})\right)^{-2}, \qquad \text{as } h \to 0.$$

These two expressions coincide when $\boldsymbol{x} = \boldsymbol{x}^\theta = \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ denotes the mode. At this point, one might expect a 'good' transition kernel to resemble the Laplace approximation of the distribution about $\boldsymbol{x}^*$, which has covariance $\nabla^2 f(\boldsymbol{x}^*)^{-1}$. This suggests a heuristic for choosing a good step size in practice, by taking h as a solution to the one-dimensional optimization problem

$$\hat{h}_\theta \coloneqq \operatorname*{arg\,min}_{h \geq 0} \left\| h \left(\mathbf{I} + \frac{h\theta}{2} \nabla^2 f(\boldsymbol{x}^*) \right)^{-2} - \nabla^2 f(\boldsymbol{x}^*)^{-1} \right\|_E, \tag{17}$$

where the norm $\|\cdot\|_E$ can be any matrix norm of choice. Solutions to (17) can be obtained using off-the-shelf methods in univariate optimization, such as golden section search (Cottle and Thapa, 2017, §9.5). We will show in the next section that, for several examples, the step size obtained from (17) with Frobenius norm tends to be an effective choice in practice, especially for $\theta = 1/2$, where it reveals itself to be *near optimal* in all of our experiments. For this choice of norm, (17) can be replaced by the equivalent problem

$$\hat{h}_\theta = \operatorname*{arg\,min}_{h \geq 0} \sum_{k=1}^{d} \left[h \left(1 + \frac{h\theta}{2} \lambda_k \right)^{-2} - \frac{1}{\lambda_k} \right]^2, \tag{18}$$

where $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$. One drawback is that solving (17) or (18) require either inversion of $\nabla^2 f(\boldsymbol{x}^*)$, or knowledge of its spectrum, respectively, both of which may be prohibitively expensive in high dimensions. In many problems, however, it is reasonable to assume a certain distribution of its spectrum; for example, that the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ of $\nabla^2 f(\boldsymbol{x}^*)$ decay exponentially:

$$\log \lambda_k \approx \left(1 - \frac{k-1}{d-1} \right) \log M + \frac{k-1}{d-1} \log m, \qquad k = 1, \ldots, d, \tag{19}$$

where m and M take the place of the smallest and largest eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$, respectively. Simplifying assumptions such as these can be justified in problems where Hessian of f is approximately low rank, in the sense that it has a small stable rank (Roosta-Khorasani and Ascher, 2015), and hence its spectrum decays fast. Under these assumptions, solving (18) becomes more tractable; we shall make use of this for Figure 3 in §5.

5. Numerical experiments

In this section, we evaluate the empirical performance of Algorithms 1 and 2 in high-dimensions as measured by a few discrepancy measures. Recall that the total variation distance between any two absolutely continuous distributions with densities p and q over $\mathbb{R}^d$ respectively is given by

$$d_{\mathrm{TV}}(p, q) \coloneqq \frac{1}{2} \int_{\mathbb{R}^d} |p(\boldsymbol{x}) - q(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}.$$

Since the total variation metric is too difficult to directly estimate in higher dimensions, we follow the standard approach in the literature (see for example Durmus and Moulines (2019) and Maire et al. (2018)) and consider instead the *mean marginal total variation* (MMTV),

$$\mathrm{MMTV}(p,q) \coloneqq \frac{1}{2d}\sum_{i=1}^{d}\int_{\mathbb{R}} |p_i(x) - q_i(x)|\mathrm{d}x,$$

which we estimate as follows. First, kernel smoothing is applied to samples for each marginal from an extended MCMC run, as well as samples obtained from a single run of each method. The total variation between these *estimated* univariate densities is then computed with high accuracy via Gauss-Kronrod quadrature (Kahaner et al., 1989).

As a weakness of MMTV is its inability to adequately compare the covariances within coordinates between the two sample sets, we also compare with a second discrepancy measure; *maximum mean discrepancy* (MMD) (Gretton et al., 2012; Muandet et al., 2017). Letting $\mathcal{H}$ denote a reproducing kernel Hilbert space over $\mathbb{R}^d$ with reproducing kernel k, MMD is defined as the integral probability metric

$$\begin{aligned}\mathrm{MMD}^2(p,q) &\coloneqq \left[\sup_{\|h\|_{\mathcal{H}}\leq 1}\int_{\mathbb{R}^d} h(x)[p(x)-q(x)]\mathrm{d}x\right]^2 \\ &= \mathbb{E}_{p,p}k(X,\tilde{X}) - 2\mathbb{E}_{p,q}k(X,Y) + \mathbb{E}_{q,q}k(Y,\tilde{Y}),\end{aligned}$$

where $X,\tilde{X}$ are independent random variables with distribution p, and $Y,\tilde{Y}$ are independent random variables with distribution q. These expectations can be estimated using samples from p and q. In our experiments, we use the Gaussian kernel:

$$k(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{y}\|^2\right),$$

where the kernel bandwidth parameter σ is chosen so that $2\sigma^2$ is the median of $(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)_{i,j=1}^n$, where $\boldsymbol{x}_i$ denotes the i-th sample taken from q.

5.1 High-dimensional Gaussian distributions

To highlight the effects of problem ill-conditioning, we once again consider sampling from a multivariate Gaussian distribution, as in (8), with explicitly computable iterates given by (9). It is easy to see that f satisfies Assumptions 1 and 3. To show efficacy in higher dimensions, we will consider $d = 1000$. Furthermore, to test the effects of ill-conditioning, we focus on three choices of $\boldsymbol{\Sigma}$ with condition numbers $\kappa \in \{1, 100, 10^8\}$. Each $\boldsymbol{\Sigma}$ is generated using the method of Bendel and Mickey (1978) to uniformly sample a correlation matrix with eigenvalues given by (19) for $m = 1$ and $M = \kappa$. For simplicity, we take $\boldsymbol{\mu} = \mathbf{0}$. MMTV and MMD discrepancies were computed between π and samples of $N = 5000$ points generated by Algorithm 1 with $\theta \in \{0, 1/2, 1\}$ and a variety of step sizes h (encompassing $4/M$ and the step size heuristics in §4). Common random numbers were used, and no burn-in period was applied. The results are shown in Figure 2. Due to the rapid explosion in magnitude of samples generated by ULA when $h \geq 4/M$, we only display discrepancies for ULA for $h < 4/M$. This critical value of is highlighted as a black solid vertical line.

In light of the large step size asymptotics, the existence of an "optimal" step size for $\theta \geq 1/2$ as evidenced in these plots is perhaps not too surprising. However, it is surprising

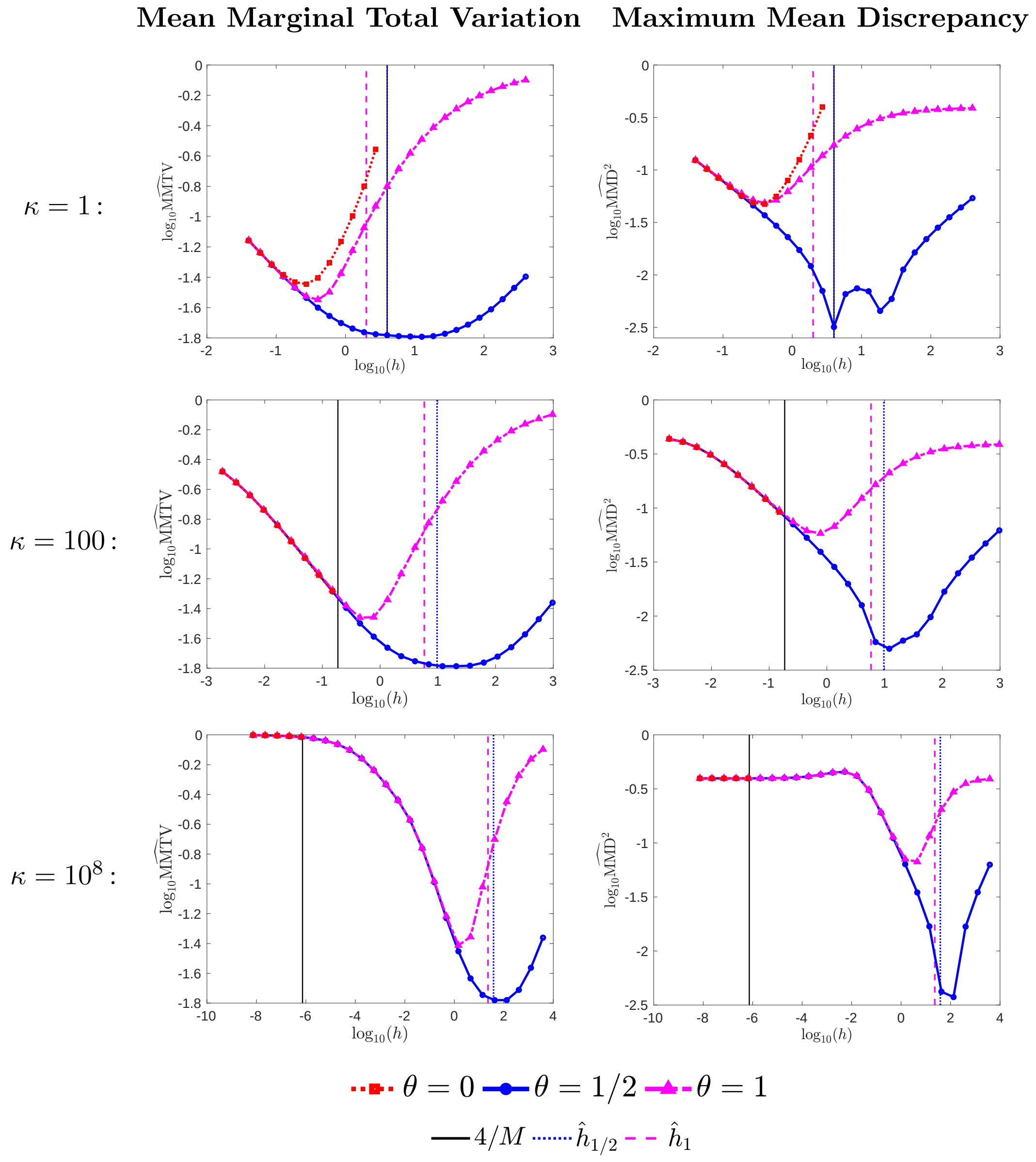

Figure 2: MMTV estimates and MMD discrepancies for 5000 samples generated by Algorithm 1 with $\theta \in \{0, 1/2, 1\}$, $h \in [\frac{4}{100M}, 100\hat{h}_{1/2}]$, and target distribution π given by (8) with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ a correlation matrix with condition number $\kappa \in \{1, 10^2, 10^8\}$.

to see that, especially for large κ, this optimum is much greater than the maximum allowed step size of $4/M$ for ULA. Most notable here is the greatly improved performance of the implicit method ($\theta = 1/2$) at this optimum over ULA for any allowable step size. Moreover, in all cases, the optimal performance of ILA for $\theta = 1/2$ exceeds that of the purely implicit method ($\theta = 1$). These two facts are not suggested by Theorem 2, implying that the large step size asymptotics should indeed play a significant role in the analysis of implicit methods moving forward.

In all cases, the step size heuristic for $\theta = 1$ performs poorly, suggesting the fully implicit case operates by a different mechanism that is currently unknown to us. For $\kappa = 1$, $\hat{h}_{1/2} = 4/M \equiv 4$, which is clearly the optimal step size, as it yields exact samples ($\boldsymbol{X}_{k+1} = \boldsymbol{Z}_k$). In fact, $\theta = 1/2$, $h = 4$ is the only choice of θ and h which results in exact samples in this scenario. According to both estimated MMTV and MMD, the step size heuristic is an almost optimal choice of h for all κ considered, even in high dimensions.

5.2 Logistic regression

We now consider sampling problems involving the Bayesian posterior densities of generalized linear models (GLM), which have log-concave likelihood functions, with Gaussian priors. For simplicity, and without loss of generality, we consider radially symmetric Gaussians. For a GLM with this choice of prior, posterior densities are proportional to $\exp(-f(\boldsymbol{x}))$, with

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(\Phi(\boldsymbol{a}_i^\top \boldsymbol{x}) - b_i \boldsymbol{a}_i^\top \boldsymbol{x} \right) + \frac{\lambda}{2} \|\boldsymbol{x}\|^2,$$

where $(\boldsymbol{a}_i, b_i),\ i = 1, 2, \cdots, n$, are the response and covariate pairs, $\boldsymbol{a}_i \in \mathbb{R}^p$, and the domain of b_i depends on the GLM. The cumulant generating function, Φ, determines the type of GLM. For example, in the case of logistic regression, $\Phi(t) = \log(1 + \exp(t))$; see McCullagh and Nelder (1989) for further details and applications. It is easy to see that

$$\nabla^2 f(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{a}_i^\top \Phi^{''}(\boldsymbol{a}_i^\top \boldsymbol{x}) + \lambda \mathbf{I} = \mathbf{A}^\top \mathbf{D} \mathbf{A} + \lambda \mathbf{I},$$

where $\mathbf{A} \in \mathbb{R}^{n \times d}$ is a matrix whose i-th row is $\boldsymbol{a}_i$, $\mathbf{D}$ is a diagonal matrix whose i-th diagonal element is $\Phi^{''}(\boldsymbol{a}^\top \boldsymbol{x})$, and λ is the precision parameter of the prior. As a result, for Assumption 3, we have

$$\lambda \leq m \leq M \leq \|\mathbf{A}\|^2 \sup_{t \in \mathbb{R}} \Phi^{''}(t) + \lambda. \tag{20}$$

For our example, we consider Bayesian logistic regression in this setting, yielding

$$f(\boldsymbol{x}) \propto \sum_{i=1}^{n} \left(\log\left(1 + \exp(\boldsymbol{a}_i^\top \boldsymbol{x})\right) - b_i \boldsymbol{a}_i^\top \boldsymbol{x} \right) + \frac{\lambda}{2} ||\boldsymbol{x}||^2, \tag{21}$$

and $\sup_{t \in \mathbb{R}} \Phi^{''}(t) \leq 1/4$. We use the `musk` (version 1) data set from the UCI repository (Dua and Graff, 2019), with the prior precision parameter $\lambda = 1$. These choices yield a target distribution which is relatively ill-conditioned, whose Hessian $\nabla^2 f(\boldsymbol{x}^*)$ about its mode $\boldsymbol{x}^*$ possesses a condition number of $\kappa > 2 \times 10^3$. We estimate the values m and M according to their lower and upper bounds in (20). For the target density given according to (21), MMTV and MMD discrepancies were computed between samples of $N = 10000$ points generated using Algorithm 2 (with $\theta \in \{0, 1/2, 1\}$, $\epsilon = 10^{-9}$ and a variety of step sizes h

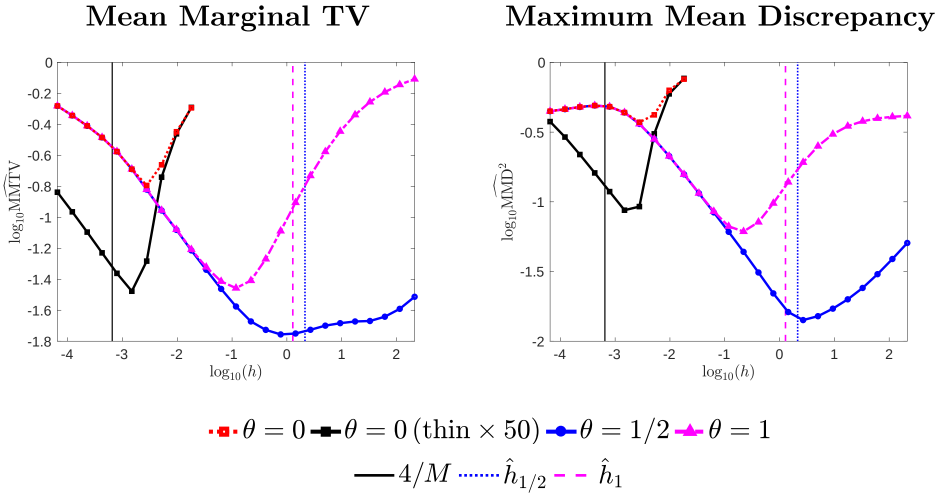

Figure 3: MMTV and MMD discrepancies between 10000 samples generated by Algorithm 2 with $\theta \in \{0, 1/2, 1\}$ and ULA with a thinning factor of 50, over $h \in [\frac{4}{10M}, 100\hat{h}_{1/2}]$, and gold standard run, for target distribution specified according to (21).

encompassing $4/M$ and the step size heuristics in §4 under the assumption that eigenvalues are distributed according to Equation 19) and a gold standard run comprised of 50,000 samples obtained from hand-tuned SMMALA (Girolami and Calderhead, 2011). Due to the large difference in computation time between ULA and ILA, for the sake of comparison, we also computed MMTV and MMD discrepancies for the ULA algorithm using the same step sizes, now with a thinning factor of 50. This factor was chosen so that the computation time of ULA became roughly equivalent to the other ILA methods. Once again, common random numbers were used, and no burn-in period was applied. The results are shown in Figure 3, and follow a similar pattern to those found in the previous example. The step size heuristic for $\theta = 1/2$ performs admirably in this case, yielding samples with smaller discrepancies to the gold standard run than ULA for any reasonable step size, even when thinned to account for the difference in computation time.

6. Conclusions

In the context of sampling from an unnormalized probability distribution, we considered a general class of unadjusted sampling algorithms that are based on implicit discretization of the Langevin dynamics. Unlike the traditional Metropolis-adjusted sampling algorithms, these unadjusted methods relax the requirement of consistency that the sample empirical distribution should asymptotically be the same as the target distribution, and hence avoid incurring serious penalty to the mixing rate of the chain. As a result, these variants generate rapidly converging Markov chains whose stationary distributions are only approximations of the target distribution with a bias that is of adjustable size. When one seeks a fixed (finite) number of samples, which is almost always the case in practice, this latter unadjusted view point can offer greatly many advantages.

In this context, we focused on the class of discretization schemes generated using θ-method in the context of smooth and strongly log-concave densities, explicitly deriving the transition kernel of the chain and establishing the corresponding sub-problems that are formulated as optimization problems. For smooth densities, the resulting implicit Langevin algorithms (ILA) have been shown to be geometrically ergodic for $\theta \geq 1/2$, irrespective of the step size. We also considered inexact variants (i-ILA) where the optimization sub-problems are solved only approximately. For this, we established non-asymptotic convergence of the sample empirical distribution to the target as measured by 2-Wasserstein metric, finding again that for $\theta > 1/2$, the resulting scheme is unconditionally stable for all step sizes. Furthermore, the growth rate in the bias term, that is shown to depend on problem's condition number, is greatly diminished for $\theta > 1/2$. Together with our numerical experiments, this suggests that the implicit methods are a more appropriate choice for ill-conditioned problems than explicit schemes. Furthermore, the case $\theta = 1/2$ appears to perform best in practice, especially when paired with our default heuristic choice of step size. The underlying reason for this is likely related to its asymptotic exactness for the normal distribution. It was suggested in Wibisono (2018) that the asymptotic bias of the $\theta = 1/2$ case could be second-order accurate, which would imply the increased performance we have observed. Unfortunately, proving this claim remains an open problem.

Although enticing, extensions of these results to non-convex cases may prove challenging due to the potential lack of unique solutions for the implicit scheme, and the relative difficulty of non-convex optimization in general. Nevertheless, one could find success in considering f that is only strongly convex outside of a compact region, as in Cheng et al. (2018a). Furthermore, although it has not been treated explicitly, we believe that implicit methods should prove effective in big data problems, that is with $f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})$ and $n \gg 1$, where it might be computationally prohibitive to evaluate f or its gradient exactly. In this regard, one can use optimization algorithms that can employ inexact oracle information; see Roosta-Khorasani and Mahoney (2018) for example. The efficacy of this approach would prove interesting for future research.

Acknowledgments

All authors have been supported by the Australian Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) under grant number CE140100049. Liam Hodgkinson acknowledges the support of an Australian Research Training Program (RTP) Scholarship. Fred Roosta was partially supported by the Australian Research Council through a Discovery Early Career Researcher Award (DE180100923). Part of this work was done while Fred Roosta was visiting the Simons Institute for the Theory of Computing.

Appendix A. Proofs

A.1 Proof of Theorem 1 (Geometric Ergodicity)

To establish geometric ergodicity, we prove the stronger Proposition 8 below. First, we connect Assumptions 1 and 2 to the lower bound (7).

Lemma 6 *The condition (7) holds under Assumptions 1 and 2.*

Proof By Assumption 1, observe that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we have

$$\begin{aligned}\frac{\langle \nabla f(\boldsymbol{x}+\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} &\leq \frac{\langle \nabla f(\boldsymbol{x}+\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} - \frac{\langle \nabla f(\boldsymbol{y}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} \\ &\leq \frac{\|\nabla f(\boldsymbol{x}+\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|^2} + \frac{\|\nabla f(\boldsymbol{y})\|\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|^2} \\ &= \frac{M\|\boldsymbol{y}\| + \|\nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x}\|}.\end{aligned}$$

Hence, we have

$$\begin{aligned}\liminf_{\|\boldsymbol{x}\|\to\infty} \frac{\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle}{\|\boldsymbol{x}-\boldsymbol{y}\|^2} &= \liminf_{\|\boldsymbol{x}\|\to\infty} \frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} \\ &\quad + \liminf_{\|\boldsymbol{x}\|\to\infty} \frac{\langle \nabla f(\boldsymbol{x}+\boldsymbol{y}) - \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} \\ &= \liminf_{\|\boldsymbol{x}\|\to\infty} \frac{\langle \nabla f(\boldsymbol{x}), \boldsymbol{x}\rangle}{\|\boldsymbol{x}\|^2} > 0.\end{aligned}$$

The result is implied by the definition of limit infimum. ■

To state the result, we recall the definition of V-uniform ergodicity, as seen in Meyn and Tweedie (2012).

Definition 7 (V-uniform ergodicity) *A ν-ergodic Markov chain $\{\boldsymbol{X}_n\}_{n=0}^{\infty}$ with Markov transition operator $\mathcal{P}$ on $\mathbb{R}^d$ is V-uniformly ergodic for a measurable function $V : \mathbb{R}^d \to [1,\infty)$ if*

$$\sup_{\boldsymbol{x}\in\mathbb{R}^d} \sup_{|\phi|\leq V} \frac{|\mathcal{P}^n \phi(\boldsymbol{x}) - \pi(\phi)|}{V(\boldsymbol{x})} \to 0, \qquad \text{as } n \to \infty.$$

By Meyn and Tweedie (2012, Theorem 16.0.1), any V-uniformly ergodic Markov chain is also geometrically ergodic.

Proposition 8 *For any $s > 0$ and f satisfying Assumptions 1 and 2, let $V_s(\boldsymbol{x})$ denote the Lyapunov drift function*

$$V_s(\boldsymbol{x}) = \exp\left(s\|\boldsymbol{x} - \boldsymbol{x}^* + \tfrac{1}{2}h\theta\nabla f(\boldsymbol{x})\|\right), \tag{22}$$

where $\boldsymbol{x}^*$ is a critical point of f. Supposing that (5) holds, the θ-method scheme with transition kernel (6) is V_s-uniformly ergodic provided $\theta \geq 1/2$, or $\theta < 1/2$ and

$$h < \frac{4m}{M^2(1-2\theta)}.$$

Proof It is immediately apparent from the positivity of (6) due to (5) that the iterates of the θ-method scheme are aperiodic and irreducible with respect to Lebesgue measure. Furthermore, it follows from Meyn and Tweedie (2012, Proposition 6.2.8) that all compact sets are small. Therefore, by Meyn and Tweedie (2012, Theorem 15.0.1) and Meyn and Tweedie (2012, Lemma 15.2.8), it suffices to show that

$$\limsup_{\|\boldsymbol{x}\|\to\infty} \frac{\mathcal{P}V_s(\boldsymbol{x})}{V_s(\boldsymbol{x})} = 0,$$

where $\mathcal{P}$ is the Markov transition operator of the θ-method scheme. Indeed, by the definition of $\limsup$, for a given $0 < \lambda < 1$, there exists a $K > 0$ such that

$$\sup_{\|\boldsymbol{x}\|\geq K} \frac{\mathcal{P}V_s(\boldsymbol{x})}{V_s(\boldsymbol{x})} \leq \lambda,$$

and so $\mathcal{P}V_s(\boldsymbol{x}) \leq \lambda V_s(\boldsymbol{x}) + \sup_{\|\boldsymbol{x}\|\leq K} \mathcal{P}V_s(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Letting $\boldsymbol{X}_1$ denote the first step of the θ-method scheme starting from $\boldsymbol{X}_0 = \boldsymbol{x}$, (4) implies

$$\boldsymbol{X}_1 + \tfrac{1}{2}h\theta\nabla f(\boldsymbol{X}_1) \sim \mathcal{N}\left(\boldsymbol{x} - \frac{h(1-\theta)}{2}\nabla f(\boldsymbol{x}),\ h\mathbf{I}\right).$$

Thus, by letting $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we obtain $\mathcal{P}V_s(\boldsymbol{x}) = \mathbb{E}\exp(sg(\boldsymbol{Z}))$, where

$$g(\boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{x}^* - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{x}) + \sqrt{h}\boldsymbol{z}\|.$$

By the reverse triangle inequality, $|g(\boldsymbol{z}_1) - g(\boldsymbol{z}_2)| \leq \sqrt{h}\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$ for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^d$, and hence, g is $\sqrt{h}$-Lipschitz in $\boldsymbol{z}$. Consequently, we can apply the Gaussian concentration inequality (Boucheron et al., 2013, Theorem 5.5) to reveal

$$\mathbb{E}e^{sg(\boldsymbol{Z})} \leq \exp\left(s\mathbb{E}g(\boldsymbol{Z}) + \frac{hs^2}{2}\right).$$

Since by Jensen's inequality,

$$\mathbb{E}g(\boldsymbol{Z}) \leq \sqrt{hd} + \|(\boldsymbol{x} - \boldsymbol{x}^*) - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{x})\|.$$

It follows that

$$\frac{\mathcal{P}V_s(\boldsymbol{x})}{V_s(\boldsymbol{x})} \leq \exp(s\sqrt{hd} + \tfrac{1}{2}hs^2 + s[T_1(\boldsymbol{x}) - T_2(\boldsymbol{x})]),$$

where

$$T_1(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^* - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{x})\| \quad \text{and} \quad T_2(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^* + \tfrac{1}{2}h\theta\nabla f(\boldsymbol{x})\|.$$

Therefore, if we can show that $T_1(\boldsymbol{x}) - T_2(\boldsymbol{x}) \to -\infty$ as $\|\boldsymbol{x}\| \to \infty$, then the result follows. Since $T_1(\boldsymbol{x}) - T_2(\boldsymbol{x}) = (T_1(\boldsymbol{x})^2 - T_2(\boldsymbol{x})^2)/(T_1(\boldsymbol{x}) + T_2(\boldsymbol{x}))$, we may focus on the difference of the squares:

$$\begin{aligned}
T_1(\boldsymbol{x})^2 - T_2(\boldsymbol{x})^2 &= \|(\boldsymbol{x} - \boldsymbol{x}^*) - \tfrac{1}{2}h(1-\theta)\nabla f(\boldsymbol{x})\|^2 - \|(\boldsymbol{x} - \boldsymbol{x}^*) + \tfrac{1}{2}h\theta\nabla f(\boldsymbol{x})\|^2 \\
&= \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - h(1-\theta)\langle\nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*\rangle + \tfrac{1}{4}h^2(1-\theta)^2\|\nabla f(\boldsymbol{x})\|^2 \\
&\qquad - \|\boldsymbol{x} - \boldsymbol{x}^*\|^2 - h\theta\langle\nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*\rangle - \tfrac{1}{4}h^2\theta^2\|\nabla f(\boldsymbol{x})\|^2 \\
&= -h\langle\nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^*\rangle + \tfrac{1}{4}h^2(1-2\theta)\|\nabla f(\boldsymbol{x})\|^2 \\
&\leq -hm\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + c(\boldsymbol{x}^*) + \tfrac{1}{4}h^2\max\{0, 1-2\theta\}M^2\|\boldsymbol{x} - \boldsymbol{x}^*\|^2, \qquad (23)
\end{aligned}$$

where the last inequality follows from (7). Provided that $\frac{1}{2}h(1-2\theta)M^2 < m$ or $\theta \geq 1/2$, (23) will be negative for sufficiently large $\boldsymbol{x}$. Since also

$$T_1(\boldsymbol{x}) + T_2(\boldsymbol{x}) \leq 2\|\boldsymbol{x} - \boldsymbol{x}^*\| + \tfrac{1}{2}h\|\nabla f(\boldsymbol{x})\| \leq (2 + \tfrac{1}{2}hM)\|\boldsymbol{x} - \boldsymbol{x}^*\|,$$

for any $\epsilon > 0$ and sufficiently large $\boldsymbol{x}$,

$$T_1(\boldsymbol{x}) - T_2(\boldsymbol{x}) \leq \frac{-hm + \frac{1}{4}h^2 \max\{0, 1-2\theta\}M^2}{2 + \frac{1}{2}hM}\|\boldsymbol{x} - \boldsymbol{x}^*\| + \epsilon,$$

which implies the difference $T_1(\boldsymbol{x}) - T_2(\boldsymbol{x}) \to -\infty$ as $\|\boldsymbol{x}\| \to \infty$, as required. ■

A.2 Proof of Theorem 2 (W_2 bounds)

Next, using techniques analogous to those of Dalalyan and Karagulyan (2019), we prove Theorem 2. For the sake of brevity, we let $a \wedge b$ denote the minimum of any two quantities a and b. The following estimate is fundamental to the argument.

Lemma 9 *Let $\boldsymbol{L}_t$ be the solution to the (overdamped) Langevin equation*

$$\mathrm{d}\boldsymbol{L}_t = -\tfrac{1}{2}\nabla f\left(\boldsymbol{L}_t\right)\mathrm{d}t + \mathrm{d}\boldsymbol{W}_t$$

for $f \in \mathcal{C}^1(\mathbb{R}^d)$ such that ∇f is M-Lipschitz continuous. Then for any $h > 0$, if $\boldsymbol{L}_0 \sim \pi$,

$$\left\|\int_0^h \nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\mathrm{d}t\right\|_{L^2} \leq \frac{1}{2}h[M\sqrt{hd}(2 + \sqrt{hM}) \wedge 4\sqrt{Md}].$$

Proof Since $\boldsymbol{L}_t$ is stationary, for any $t \geq 0$, $\|\nabla f(\boldsymbol{L}_t)\|_{L^2} \leq \sqrt{Md}$ by Dalalyan (2017a, Lemma 2). Therefore, $\left\|\int_0^h \nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\mathrm{d}t\right\|_{L^2} \leq 2h\sqrt{Md}$. Furthermore, following the same procedure as in Dalalyan and Karagulyan (2019, Lemma 4)

$$\left\|\int_0^h \nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\mathrm{d}t\right\|_{L^2} \leq \frac{1}{4}h^2 M^{3/2} d^{1/2} + \frac{2}{3}h^{3/2} M d^{1/2}.$$

■

With Lemma 9 in hand, we may proceed with the proof of the main result.

Proof [Theorem 2] Letting $\boldsymbol{W}_t$ denote a d-dimensional standard Brownian motion independent of $\boldsymbol{X}_k$ and $\boldsymbol{L}_0 \sim \pi$, define the stochastic process $\boldsymbol{L}$ by

$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \frac{1}{2}\int_0^t \nabla f(\boldsymbol{L}_s)\mathrm{d}s + \boldsymbol{W}_t, \qquad t \geq 0.$$

Evidently, $\boldsymbol{L}_t$ is a realization of (1) and so is a reversible Markov process with $\boldsymbol{L}_t \sim \pi$ for every $t \geq 0$. Now, couple the inexact θ-method scheme $\boldsymbol{X}_k$ satisfying

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \frac{h}{2}[\theta\nabla f(\boldsymbol{X}_{k+1}) + (1-\theta)\nabla f(\boldsymbol{X}_k)] + \sqrt{h}\boldsymbol{Z}_k + \boldsymbol{E}_k,$$

for an appropriate error term $\boldsymbol{E}_k$, to $\boldsymbol{L}_t$, by letting $\boldsymbol{Z}_k = h^{-1/2}[\boldsymbol{W}_{(k+1)h} - \boldsymbol{W}_{kh}]$ for each $k \geq 1$, and choosing $\boldsymbol{L}_0$ such that $W_2(\pi_0, \pi) = \|\boldsymbol{X}_0 - \boldsymbol{L}_0\|_{L^2}$. Observing that, for any $k \geq 1$,

$$\begin{aligned}\boldsymbol{E}_k &= \frac{h}{2}\theta\nabla f(\boldsymbol{X}_{k+1}) + \boldsymbol{X}_{k+1} - \boldsymbol{X}_k + \frac{h}{2}(1-\theta)\nabla f(\boldsymbol{X}_k) - \sqrt{h}\boldsymbol{Z}_k \\ &= \frac{h}{2}\nabla F(\boldsymbol{X}_{k+1}; \boldsymbol{X}_k, \boldsymbol{Z}_k),\end{aligned}$$

by construction, $\|\boldsymbol{E}_k\|_{L^2} \leq \frac{1}{2}h\epsilon$. For each k, let $\boldsymbol{D}_k = \boldsymbol{L}_{kh} - \boldsymbol{X}_k$, observing that $W_2(\pi_0, \pi) = \|\boldsymbol{D}_0\|_2$ and $W_2(\pi_k, \pi) \leq \|\boldsymbol{D}_k\|_2$. Choosing some $k \geq 1$, for the sake of brevity, we denote $\boldsymbol{L}_t^{(k)} = \boldsymbol{L}_{kh+t}$, which now satisfies

$$\begin{aligned}\boldsymbol{L}_h^{(k)} = \boldsymbol{L}_{kh+h} &= \boldsymbol{L}_{kh} - \frac{1}{2}\int_0^h \nabla f(\boldsymbol{L}_{kh+s})\mathrm{d}s + \boldsymbol{W}_{kh+h} - \boldsymbol{W}_{kh} \\ &= \boldsymbol{L}_0^{(k)} - \frac{1}{2}\int_0^h \nabla f(\boldsymbol{L}_s^{(k)})\mathrm{d}s + \sqrt{h}\boldsymbol{Z}_k.\end{aligned}$$

Altogether, we have

$$\boldsymbol{D}_{k+1} = \boldsymbol{D}_k - \tfrac{1}{2}h[(1-\theta)\boldsymbol{U}_k + \theta\tilde{\boldsymbol{U}}_k] - [(1-\theta)\boldsymbol{V}_k + \theta\tilde{\boldsymbol{V}}_k] + \boldsymbol{E}_k,$$

where

$$\begin{aligned}\boldsymbol{U}_k &= \nabla f(\boldsymbol{X}_k + \boldsymbol{D}_k) - \nabla f(\boldsymbol{X}_k) & \boldsymbol{V}_k &= \frac{1}{2}\int_0^h \nabla f(\boldsymbol{L}_s^{(k)}) - \nabla f(\boldsymbol{L}_0^{(k)})\mathrm{d}s \\ \tilde{\boldsymbol{U}}_k &= \nabla f(\boldsymbol{X}_{k+1} + \boldsymbol{D}_{k+1}) - \nabla f(\boldsymbol{X}_{k+1}) & \tilde{\boldsymbol{V}}_k &= \frac{1}{2}\int_0^h \nabla f(\boldsymbol{L}_{h-s}^{(k)}) - \nabla f(\boldsymbol{L}_h^{(k)})\mathrm{d}s.\end{aligned}$$

An application of the fundamental theorem of calculus implies $\boldsymbol{U}_k = \mathbf{F}_k\boldsymbol{D}_k$ and $\tilde{\boldsymbol{U}}_k = \mathbf{F}_{k+1}\boldsymbol{D}_{k+1}$, where

$$\mathbf{F}_k = \int_0^1 \nabla^2 f(\boldsymbol{X}_k + t\boldsymbol{D}_k)\mathrm{d}t.$$

Altogether, $\boldsymbol{D}_{k+1} = \mathbf{S}_k\boldsymbol{D}_k + \boldsymbol{T}_k$ where $\boldsymbol{T}_k = -(\mathbf{I} + \frac{h\theta}{2}\mathbf{F}_{k+1})^{-1}[(1-\theta)\boldsymbol{V}_k + \theta\tilde{\boldsymbol{V}}_k - \boldsymbol{E}_k]$ and

$$\mathbf{S}_k = \left(\mathbf{I} + \frac{h\theta}{2}\mathbf{F}_{k+1}\right)^{-1}\left(\mathbf{I} - \frac{h(1-\theta)}{2}\mathbf{F}_k\right).$$

It can be verified using induction that the solution to this first-order non-homogeneous recurrence relation is given by

$$\boldsymbol{D}_k = \mathbf{S}_{k-1}\cdots\mathbf{S}_0\boldsymbol{D}_0 + \sum_{l=0}^{k-1}\mathbf{S}_{k-1}\cdots\mathbf{S}_{l+1}\boldsymbol{T}_l.$$

Now, observe that by denoting $G(\mathbf{X}) = (\mathbf{I} - \frac{h(1-\theta)}{2}\mathbf{X})(\mathbf{I} + \frac{h\theta}{2}\mathbf{X})^{-1}$, for any $l < k$,

$$\mathbf{S}_{k-1}\cdots\mathbf{S}_l = \left(\mathbf{I} + \frac{h\theta}{2}\mathbf{F}_k\right)^{-1} G(\mathbf{F}_{k-1})\cdots G(\mathbf{F}_{l+1})\left(\mathbf{I} - \frac{h(1-\theta)}{2}\mathbf{F}_l\right). \tag{24}$$

Since the eigenvalues of $\nabla^2 f$ are bounded above by M and below by m, so too are the eigenvalues of $\mathbf{F}_k$ for each k. Therefore,

$$\begin{aligned}\|G(\mathbf{F}_k)\|_2 &= \max_{z\in[m,M]}\left|\frac{1-\frac{1}{2}h(1-\theta)z}{1+\frac{1}{2}h\theta z}\right| \\ &= \max\left\{\frac{1-\frac{1}{2}h(1-\theta)m}{1+\frac{1}{2}h\theta m}, \frac{\frac{1}{2}h(1-\theta)M-1}{\frac{1}{2}h\theta M+1}\right\} =: \rho. \qquad (25)\end{aligned}$$

The transition between these regimes occurs at the point h^* which is the solution to

$$\frac{1-\frac{1}{2}h(1-\theta)m}{1+\frac{1}{2}h\theta m} = \frac{\frac{1}{2}h(1-\theta)M-1}{\frac{1}{2}h\theta M+1}$$

over $h > 0$. Equivalently, it is the solution to

$$\begin{aligned}&\tfrac{1}{2}h\theta M + 1 - \tfrac{1}{4}h^2\theta(1-\theta)mM - \tfrac{1}{2}h(1-\theta)m \\ &\qquad = \tfrac{1}{2}h(1-\theta)M + \tfrac{1}{4}h^2\theta(1-\theta)mM - 1 - \tfrac{1}{2}h\theta m,\end{aligned}$$

and therefore to the quadratic equation

$$\tfrac{1}{2}h(1-2\theta)(m+M) + \tfrac{1}{2}h^2\theta(1-\theta)mM - 2 = 0.$$

It may be readily verified that h^* as defined in (12) is the only positive solution. Furthermore, $\rho < 1$ provided that $\theta \geq 1/2$ or $h < 4/[M(1-2\theta)]$. Also, for any $j,k \geq 1$,

$$\begin{aligned}&\|(\mathbf{I}+\tfrac{1}{2}h\theta\mathbf{F}_j)^{-1}\|_2\|\mathbf{I}-\tfrac{1}{2}h(1-\theta)\mathbf{F}_k\|_2 \\ &\qquad \leq \frac{\max\{1-\frac{1}{2}hm(1-\theta), \frac{1}{2}hM(1-\theta)-1\}}{1+\frac{1}{2}hm\theta} \leq \kappa_h\rho, \qquad (26)\end{aligned}$$

which further implies $\|\mathbf{S}_j\|_2 \leq \kappa_h\rho$ for any j. Now combining (24), (25), and (26), for $k > l$, $\|\prod_{j=l}^{k-1}\mathbf{S}_j\|_2 \leq \kappa_h\rho^{k-l}$, and hence, altogether,

$$\|\boldsymbol{D}_k\|_2 \leq \kappa_h\rho^k\|\boldsymbol{D}_0\|_2 + \sum_{l=0}^{k-1}\kappa_h\rho^{k-l}\|\boldsymbol{T}_l\|_2.$$

Since $\boldsymbol{L}_t$ is reversible and stationary, $\|\boldsymbol{V}_k^*\|_2 = \|\boldsymbol{V}_k\|_2$, and by Lemma 9, $\|\boldsymbol{V}_k\|_2 \leq \frac{1}{2}h[M\sqrt{hd}(2+\sqrt{hM}) \wedge 4\sqrt{Md}]$. Therefore

$$\|\boldsymbol{T}_k\|_2 \leq \frac{\frac{1}{2}h[\epsilon + M\sqrt{hd}(2+\sqrt{hM}) \wedge 4\sqrt{Md}]}{1+\frac{1}{2}hm\theta}.$$

Since

$$\begin{aligned}1-\frac{1-\frac{1}{2}h(1-\theta)m}{1+\frac{1}{2}h\theta m} &= \frac{1+\frac{1}{2}h\theta m - 1 + \frac{1}{2}h(1-\theta)m}{1+\frac{1}{2}h\theta m} = \frac{\frac{1}{2}hm}{1+\frac{1}{2}h\theta m}, \\ 1-\frac{\frac{1}{2}h(1-\theta)M-1}{\frac{1}{2}h\theta M+1} &= \frac{1+\frac{1}{2}h\theta M + 1 - \frac{1}{2}h(1-\theta)M}{\frac{1}{2}h\theta M+1} = \frac{2+\frac{1}{2}h(2\theta-1)M}{\frac{1}{2}h\theta M+1},\end{aligned}$$

applying the closed-form expression for the geometric series,

$$\sum_{l=0}^{k-1} \rho^{k-l} \leq \frac{1}{1-\rho} = \max\left\{\frac{1+\frac{1}{2}h\theta m}{\frac{1}{2}hm}, \frac{\frac{1}{2}h\theta M + 1}{2+\frac{1}{2}h(2\theta-1)M}\right\},$$

and the result follows. ■

A.3 Proof of Theorem 5 (Central Limit Theorem)

Proof The proof makes use of Laplace's method. From (6) and the change of variables theorem, the Markov kernel $\tilde{p}_h(\boldsymbol{y}|\boldsymbol{x})$ for the transition $\boldsymbol{X}_k \mapsto \sqrt{h}(\boldsymbol{X}_{k+1} - \boldsymbol{X}_k^\theta)$ is given by

$$\tilde{p}_h(\boldsymbol{y}|\boldsymbol{x}) = (2\pi)^{-d/2} \det\left(\frac{1}{h}\mathbf{I} + \frac{\theta}{2}\nabla^2 f(\boldsymbol{x}^\theta + h^{-1/2}\boldsymbol{y})\right) \times$$
$$\exp\left(-\frac{1}{2h}\left\|\boldsymbol{x}^\theta - \boldsymbol{x} + h^{-1/2}\boldsymbol{y} + \frac{h\theta}{2}\nabla f(\boldsymbol{x}^\theta + h^{-1/2}\boldsymbol{y}) + \frac{h(1-\theta)}{2}\nabla f(\boldsymbol{x})\right\|^2\right). \quad (27)$$

Letting $q(\boldsymbol{y}|\boldsymbol{x}) = \phi(\boldsymbol{y}; \mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{x}))$ where $\boldsymbol{\Sigma}(\boldsymbol{x}) = (4/\theta^2)\nabla^2 f(\boldsymbol{x}^\theta)^{-2}$, it suffices to show that $\tilde{p}_h(\boldsymbol{y}|\boldsymbol{x}) \to q(\boldsymbol{y}|\boldsymbol{x})$ as $h \to \infty$, pointwise in $\boldsymbol{y}$. Denoting $H_h(\boldsymbol{y}) = h^{-1}\mathbf{I} + \frac{1}{2}\theta\int_0^1 \nabla^2 f(\boldsymbol{x}^\theta + th^{-1/2}\boldsymbol{y})dt$, since $\theta\nabla f(\boldsymbol{x}^\theta) = -(1-\theta)\nabla f(\boldsymbol{x})$, the exponent of (27) becomes

$$-\frac{1}{2h}\|\boldsymbol{x}^\theta - \boldsymbol{x} + \sqrt{h}H_h(\boldsymbol{y})\boldsymbol{y}\|^2 = -\frac{\|\boldsymbol{x}^\theta - \boldsymbol{x}\|^2}{2h} - \frac{(\boldsymbol{x}^\theta - \boldsymbol{x})^\top H_h(\boldsymbol{y})\boldsymbol{y}}{\sqrt{h}} - \frac{1}{2}\boldsymbol{y}^\top H_h(\boldsymbol{y})^2\boldsymbol{y}.$$

Since $H_h(\boldsymbol{y})^2 \to \boldsymbol{\Sigma}(\boldsymbol{x})^{-1}$ and the determinant term converges to $\det(\boldsymbol{\Sigma}^{-1/2})$, the result follows. ■

References

David Anderson and Jonathan Mattingly. A weak trapezoidal method for a class of stochastic differential equations. *Communications in mathematical sciences*, 9:301–318, 03 2011.

Uri M. Ascher. *Numerical methods for evolutionary differential equations.* SIAM, 2008.

Uri M. Ascher and Linda Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1998. ISBN 9781611971392.

Robert B. Bendel and M. Ray Mickey. Population correlation matrices for sampling experiments. *Communications in Statistics-Simulation and Computation*, 7(2):163–182, 1978.

R. Biscay, J. C. Jimenez, J. J. Riera, and P. A. Valdes. Local linearization method for the numerical solution of stochastic differential equations. *Annals of the Institute of Statistical Mathematics*, 48(4):631–644, 1996.

Christopher M. Bishop and Michael E. Tipping. Bayesian regression and classification. *Nato Science Series sub Series III Computer And Systems Sciences*, 190:267–288, 2003.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bruno Casella, Gareth Roberts, and Osnat Stramer. Stability of partially implicit Langevin schemes and their MCMC variants. *Methodology and Computing in Applied Probability*, 13(4):835–854, December 2011.

Chris Chatfield, Jim Zidek, and Jim Lindsey. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2010.

Xiang Cheng and Peter Bartlett. Convergence of Langevin MCMC in KL-divergence. In *Proceedings of Algorithmic Learning Theory*, volume 83, pages 186–211, 2018.

Xiang Cheng, Niladri S. Chatterji, Yasin Abbasi-Yadkori, Peter L. Bartlett, and Michael I. Jordan. Sharp Convergence Rates for Langevin Dynamics in the Nonconvex Setting. *arXiv preprint arXiv:1805.01648*, 2018a.

Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018b.

Patrick L. Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

Richard W. Cottle and Mukund N. Thapa. *Linear and Nonlinear Optimization*. International Series in Operations Research & Management Science. Springer New York, 2017. ISBN 9781493970537.

Rixon Crane and Fred Roosta. DINGO: Distributed Newton-type method for gradient-norm optimization. In *Advances in Neural Information Processing Systems*, pages 9494–9504, 2019.

Arnak S. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017a.

Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017b.

Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Anirban DasGupta. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics.* Springer Texts in Statistics. Springer New York, 2011. ISBN 9781441996336.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

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

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, volume 3. JHU Press, 4 edition, 2012.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar): 723–773, 2012.

Niels Richard Hansen. Geometric ergodicity of discrete-time approximations to multivariate diffusions. *Bernoulli*, 9(4):725–743, 2003.

Wilfred K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. ISSN 00063444.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24. Elsevier, 2014.

David Kahaner, Cleve B. Moler, Stephen Nash, and George E. Forsythe. *Numerical Methods and Software.* Prentice-Hall series in computational mathematics. Prentice Hall, 1989.

Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.

Andrei N. Kolmogorov. Zur umkehrbarkeit der statistischen naturgesetze. *Mathematische Annalen*, 113(1):766–772, 1937.

Marie Kopec. Weak backward error analysis for overdamped Langevin processes. *IMA Journal of Numerical Analysis*, 35(2):583–614, 2014.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *International Conference on Machine Learning*, pages 181–189, 2014.

John Denholm Lambert. *Numerical methods for ordinary differential systems: the initial value problem.* John Wiley & Sons, Inc., 1991.

Florian Maire, Nial Friel, and Pierre Alquier. Informed sub-sampling MCMC: approximate Bayesian inference for large datasets. *Statistics and Computing*, pages 1–34, Jun 2018.

Jonathan C. Mattingly, Andrew M. Stuart, and Desmond J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

Peter McCullagh and John A. Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Jorge Nocedal and Stephen Wright. *Numerical Optimization.* Springer Science & Business Media, 2006.

Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

Marcelo Pereyra. Proximal Markov Chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, Jul 2016. ISSN 1573-1375.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods.* Springer texts in statistics. Springer, 1999. ISBN 9780387987071.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical science*, 16(4):351–367, 2001.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 12 1996.

R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

Fred Roosta, Yang Liu, Peng Xu, and Michael W. Mahoney. Newton-MR: Newton's Method Without Smoothness or Convexity. *arXiv preprint arXiv:1810.00303*, 2018.

Farbod Roosta-Khorasani and Uri M. Ascher. Improved bounds on sample size for implicit matrix trace estimators. *Foundations of Computational Mathematics*, 15(5):1187–1212, 2015.

Farbod Roosta-Khorasani and Michael W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, Nov 2018.

Jun Shao. *Mathematical Statistics.* Springer Texts in Statistics. Springer New York, 2008. ISBN 9780387217185.

Endre Süli and David F. Mayers. *An introduction to numerical analysis.* Cambridge University Press, 2003.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 2093–3027. PMLR, 2018.

Peng Xu, Fred Roosta, and Michael W Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. *Mathematical Programming*, pages 1–36, 2017.