Journal of Machine Learning Research 25 (2024) 1-32 Submitted 2/23; Revised 10/24; Published 10/24

Optimal Scaling for the Proximal Langevin Algorithm in High Dimensions

Natesh S. Pillai PILLAI@FAS.HARVARD.EDU
Department of Statistics
Harvard University
MA 02138, USA

Editor: Matthew Hoffman

Abstract

The Metropolis-adjusted Langevin (MALA) algorithm is a sampling algorithm that incorporates the gradient of the logarithm of the target density in its proposal distribution. In an earlier joint work Pillai et al. (2012), the author had extended the seminal work of Roberts and Rosenthal (1998) and showed that in stationarity, MALA applied to an $N-$dimensional approximation of the target will take $\mathcal{O}(N^{\frac{1}{3}})$ steps to explore its target measure. It was also shown in Roberts and Rosenthal (1998) and Pillai et al. (2012) that, as a consequence of the diffusion limit, the MALA algorithm is optimized at an average acceptance probability of 0.574. In Pereyra (2016), the author introduced the proximal MALA algorithm where the gradient of the log target density is replaced by the proximal function (mainly aimed at implementing MALA for non-differentiable target densities). In this paper, we show that for a wide class of twice differentiable target densities, the proximal MALA enjoys the same optimal scaling as that of MALA in high dimensions and also has an average optimal acceptance probability of 0.574. The results of this paper thus give the following practically useful guideline: for smooth target densities where it is expensive to compute the gradient while implementing MALA, users may replace the gradient with the corresponding proximal function (that can be often computed relatively cheaply via convex optimization) *without* losing any efficiency gains from optimal scaling. We show this for two class of examples. First, for the product of Gaussians, we identify the optimal scale for proximal MALA and show that it is identical to MALA. Next, following the exact framework used in Pillai et al. (2012), we define a version of the proximal MALA algorithm in a Hilbert space. We show that for a certain class of twice differentiable, infinite dimensional *non-product* measures commonly used in applications, the proximal MALA applied to an $N-$dimensional approximation of the target also will take $\mathcal{O}(N^{\frac{1}{3}})$ steps to explore the invariant measure, with an optimal acceptance probability of 0.574. This confirms some of the empirical observations made in Pereyra (2016).

Keywords: Markov Chain Monte Carlo, Metropolis Adjusted Langevin Algorithm, Scaling limit, Diffusion Approximation, Convex Optimization, Proximal Operators, Moreau Envelope.

1. Introduction

The Langevin diffusion in $\mathbb{R}^N$

$$dX_t = \nabla \log \pi^N(X_t)dt + \sqrt{2}\, dW_t \tag{1}$$

under practically realistic regularity assumptions on the measure π^N has π^N as its invariant measure. The Langevin algorithm has been one of the workhorses for sampling probability measures; it is widely used in Bayesian statistics (Robert and Casella, 2004), data assimilation, inverse problems (Stuart, 2010) and machine learning *e.g.,* Welling and Teh (2011); Lamperski (2021), among other areas of data science. The time discretization of X_t with step-size δ gives rise to the Langevin proposal:

$$y = x + \delta \nabla \log \pi^N(x) + \sqrt{2\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N)\ . \tag{2}$$

Consider a π^N-invariant Metropolis Hastings Markov chain $\{x^{k,N}\}_{k\geq 1}$ obtained by proposing y from the current state x according to the kernel $q(x, y)$ given by (2) and then accepted with probability

$$\alpha(x, y) = 1 \wedge \frac{\pi^N(y)q(y, x)}{\pi^N(x)q(x, y)}. \tag{3}$$

The proposal (2) coupled with the accept-reject mechanism above constitutes the Metropolis Adjusted Langevin Algorithm (MALA) (Robert and Casella, 2004). The proposal kernel for the simpler, Random Walk Metropolis (RWM) algorithm is derived from the following random walk:

$$y = x + \sqrt{\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N)\ . \tag{4}$$

An important question regarding the computational complexity of these Markov chains is how should the parameter δ vary as a function of the dimension N. A well-known heuristic for choosing δ is the following: smaller values of δ lead to high acceptance rates but the chain moves very slowly and therefore may not be efficient. Larger values of δ lead to larger moves, but are rejected more often because of smaller acceptance probabilities. The "optimal scale" for the proposal variance thus strikes a balance between making large moves and still having an $\mathcal{O}(1)$ acceptance probability as a function of the dimension N.

To make this heuristic precise, consider the continuous interpolant of the Markov chain $X^{k,N}$:

$$z^N(t) = \Big(\frac{t}{\Delta t} - k\Big)\, x^{k+1,N} + \Big(k + 1 - \frac{t}{\Delta t}\Big)\, x^{k,N}, \qquad \text{for} \qquad k\Delta t \leq t < (k+1)\Delta t. \tag{5}$$

We choose the proposal variance to satisfy $\delta = \ell \Delta t$, with $\Delta t = N^{-\gamma}$ setting the scale in terms of dimension and the parameter ℓ a "tuning" parameter which is independent of the dimension N. We now discuss how to choose γ and ℓ.

Suppose that π^N is the product of N probability densities π,

$$\pi^N(x) \propto \prod_{i=1}^{N} \pi(x_i). \tag{6}$$

For this product measure, the seminal papers Roberts et al. (1997) and Roberts and Rosenthal (1998) respectively showed that, *in stationarity*, the "optimal" choice for γ that maximizes the expected squared jumping distance is $\gamma = 1$ for the RWM algorithm and $\gamma = \frac{1}{3}$

for the MALA. Moreover, the projection of z^N into any single fixed coordinate direction x_i converges weakly in $C([0,T];\mathbb{R})$ to z, the scalar diffusion process of the form:

$$\frac{dz}{dt} = h(\ell)[\log \pi(z)]' + \sqrt{2h(\ell)}\frac{dW}{dt}. \tag{7}$$

Here $h(\ell) > 0$ is a constant determined by the parameter ℓ from the proposal variance. The quantity $h(\ell)$ has the interpretation as the "speed measure" of the limiting diffusion; see Roberts and Rosenthal (2001). Choosing ℓ to maximize $h(\ell)$, thus maximizing the speed of the limiting diffusion, then yields an optimal average acceptance probability of 0.234 for the Random Walk Metropolis Algorithm and 0.574 for MALA. A remarkable feature of these results is that the optimal acceptance probabilities for these two algorithms are "universal" – they hold for a wide range of π.

The above analysis shows that the number of steps required to sample the target measure grows as $\mathcal{O}(N)$ for RWM, but only as $\mathcal{O}(N^{\frac{1}{3}})$ for MALA. This quantifies the efficiency gained by use of MALA over RWM, and in particular from employing local moves informed by the gradient of the logarithm of the target density. These theoretical analyses have inspired much further research as they give useful guidelines for implementation of MALA in high dimensions: in addition to employing an explicit scale in the proposal variance as predicted by the theory, one should "tune" the proposal variance of the RWM and MALA algorithms so as to have acceptance probabilities of 0.234 and 0.574 respectively.

1.1 Proximal MALA algorithm

The proximal MALA algorithm was introduced in Pereyra (2016). For a convex function $f : \mathbb{R}^N \mapsto \mathbb{R}$, $\lambda > 0$ and $\|\cdot\|$ denoting the Euclidean norm, define the proximity operator (also called the λ-Moreau envelope; see Bauschke and Combettes, 2011):

$$\mathrm{Prox}_f^\lambda(x) = \mathrm{argmin}_{y\in\mathbb{R}^N}\Big(f(y) + \frac{1}{2\lambda}\|y-x\|^2\Big).$$

The following two extreme limits are well known for proximal functions (see Bauschke and Combettes, 2011, chap. 12):

$$\lim_{\lambda\to 0}\mathrm{Prox}_f^\lambda(x) = x, \qquad \lim_{\lambda\to\infty} f(\mathrm{Prox}_f^\lambda(x)) = \inf_{y\in\mathbb{R}^N} f(y).$$

Let π^N be a probability density in $\mathbb{R}^N$ and consider its $\lambda-$Moreau approximation (see Equation (3) of Pereyra, 2016):

$$\pi_\lambda^N(x) \propto \sup_{u\in\mathbb{R}^N}\pi(u)\exp\Big(-\frac{1}{2\lambda}\|u-x\|^2\Big).$$

If $\pi^N(x) \propto \exp(-\Psi(x))$ for a convex function Ψ, we have the identity:

$$\pi_\lambda^N(x) \propto \exp\Big\{-\Psi\Big(\mathrm{Prox}_\Psi^\lambda(x)\Big)\Big\}\exp\Big\{-\frac{1}{2\lambda}\|\mathrm{Prox}_\Psi^\lambda(x) - x\|^2\Big\}. \tag{8}$$

In addition, if Ψ is differentiable, we also have the identity (Bauschke and Combettes, 2011, Equation (12.28)):

$$\frac{1}{\lambda}(x - \mathrm{Prox}_\Psi^\lambda(x)) = \nabla\Psi(\mathrm{Prox}_\Psi^\lambda(x)). \tag{9}$$

Equation (9) can be thought of as an implicit gradient. Indeed, the usual explicit Euler discretization for MALA yields:

$$x^{k+1,N} = x^{k,N} - \lambda \nabla\Psi(x^{k,N}) + \sqrt{2\lambda}\, Z^N, \quad Z^N \sim \mathrm{N}(0, \mathrm{I}_N) \tag{10}$$

whereas (9) leads to the implicit update equation

$$x^{k+1,N} = x^{k,N} - \lambda \nabla\Psi(x^{k+1,N}) + \sqrt{2\lambda}\, Z^N \tag{11}$$

or equivalently

$$x^{k+1,N} = \mathrm{Prox}_{\Psi}^{\lambda}(x^{k,N}) + \sqrt{2\lambda}\, Z^N. \tag{12}$$

Motivated by (9) and (12), in Pereyra (2016), the author introduced the following modification of the discrete Langevin proposal [1] (2):

$$y = \left(1 - \frac{\delta}{\lambda}\right) x + \frac{\delta}{\lambda}\mathrm{Prox}_{\Psi}^{\lambda}(x) + \sqrt{2\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N)\ . \tag{13}$$

The proximal MALA Markov chain then proceeds via the accept-reject mechanism (3) using the proposal given in (13).

In Pereyra (2016), the author chose $\delta = \lambda$ on grounds of the stability of the resulting algorithm. We also make this choice. Thus our proximal MALA proposal is given by:

$$y = \mathrm{Prox}_{\Psi}^{\delta}(x) + \sqrt{2\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N)\ . \tag{14}$$

During the revision stages of this paper, the preprint Crucinio et al. (2023) was posted that significantly generalized our results. In Crucinio et al. (2023), the authors show that $\lambda \neq \delta$ leads to sub-optimal results; see Section 8 for more discussion.

One of the main reasons why the proximal MALA was introduced in Pereyra (2016) is that the proposal (14) can be applied to targets even when Ψ is not differentiable: *e.g.*, the Laplace density $\Psi(x) = |x|$. Quoting Pereyra (2016): "finally, similarly to other MH algorithms based on local proposals, proximal MALA may be geometrically ergodic yet perform poorly if the proposal variance δ is either too small or very large. Theoretical and experimental studies of MALA show that for many high-dimensional target densities the value of δ should be set to achieve an acceptance rate of approximately $40\% - 70\%$ (Pillai et al. 2012)."

1.2 Motivation

In this paper, we show that both the MALA algorithm and the proximal-MALA algorithm *enjoy* the same optimal scaling and hence the optimal acceptance probability for a wide range of *differentiable* target measures. Our results thus provide the first theoretical confirmation of the empirical observation above made in Pereyra (2016). It is natural to ask why one should consider differentiable target densities for studying the performance of the proximal MALA algorithm since it was developed mainly for addressing the non-differentiable case. We mention a few reasons that illustrate why such a study is useful.

[1] For notational consistency, we have set $2\delta = \delta'$ where δ' is the analogous parameter in Pereyra's definition; see Equation (9) of Pereyra (2016)

1. The proposal for the MALA algorithm is obtained from the explicit (forward) Euler discretization in (10):

$$x^{k+1,N} = x^{k,N} - \lambda \nabla \Psi(x^{k,N}) + \sqrt{2\lambda}\, Z^N,$$

   whereas the proximal MALA proposal is obtained from the implicit (backward) Euler discretization as described in (11). Thus is interesting to know, and far from obvious *apriori*, that if this small change in the proposal obtained by the implicit discretization (proximal MALA) has better or worse scaling properties than the explicit method (MALA). As mentioned before, one of our main contribution in this paper is to show that for a wide class of differentiable targets, both of these methods have the same optimal scaling. In Section 1.4 we give a heuristic argument showing why this is the case. It is interesting to note that even if the target distribution is non-differentiable only on a set of measure zero (*e.g.*, the Laplace density, $\Psi(x) = |x|$), the proximal MALA does not achieve the $N^{-\frac{1}{3}}$ scaling as it does for smooth targets; see Crucinio et al. (2023).

2. Even if the target density is differentiable, in many practical applications it may be very expensive to compute the gradient, whereas it is often cheap to compute the proximal function via convex optimization. For example, in many applied models encountered in data assimilation and Bayesian inverse problems (Stuart, 2010), the target density is of the form:

$$\pi^N(\Theta|Y) \propto \exp\Big(-\frac{1}{2\sigma^2}\|Y - G(\Theta)\|^2 + h(\Theta)\Big)$$

   where $G : \mathbb{R}^N \mapsto \mathbb{R}$ is an expensive, non-linear function to compute (such as the solution of a climate model obtained via solving a partial differential equation), Θ is a parameter we wish to compute posterior inference for, Y is the observed data and $\exp(h(\Theta))$ denotes the prior distribution for Θ. There have been quite a few papers recently where a sophisticated neural network was used to approximate G when it is a solution of a partial differential equation (Kovachki et al. (2023); Jiang et al. (2023)). In such examples, even for lower dimensional Θ, it can be even more expensive to compute the gradient of a neural network so as to compute the derivative of G with respect to Θ. Thus there is a natural need for developing derivative free sampling algorithms that enjoy the same optimality properties of Langevin algorithms.[2]

3. Optimal scaling is not the only facet of algorithm design; many other factors must be taken into consideration. Even though our results show that the optimal scaling and the optimal acceptance probability for MALA and proximal MALA algorithms are the same, there are many examples in which these two algorithms show vastly different behavior both during the transient phase and at stationarity. It is well known that in many ODEs and PDEs, the implicit discretization is numerically more stable; see Elliott and Stuart (1993) for a construction of an implicit method that is

[2] One such class of algorithms is the recently studied zeroth-order discretization of Langevin algorithms in Roy et al. (2022). It would be of interest to compare the performance of proximal MALA to the algorithms developed in Roy et al. (2022).

much more stable than its explicit counterpart. Let us give another example from Bayesian statistics. Consider a Poisson regression model:

$$Y|X \sim \text{Poisson}(e^X),$$

with the prior distribution $\pi(X) \propto \exp(-\frac{1}{2}X^2)$. The goal is to infer the posterior distribution $\pi(X|Y)$. Suppose that we observed $Y = 1$. Then $\pi(X|Y = 1) \propto e^{-\frac{1}{2}X^2+X-e^X}$. Since $\pi(X|Y = 1)$ has very light tails, the gradient of $\log \pi(X|Y = 1)$ takes very large negative values for $X \gg 1$. Thus if initialized at large values of X, MALA and proximal MALA show very different behavior. As Figure 1 shows, the MALA algorithm gets stuck when initialized at $X = 10$ whereas the proximal MALA mixes after an initial burn in of $\sim$ 50 steps. We conjecture that, for this example, MALA is not even geometrically ergodic whereas proximal MALA is. However, we believe that using the methods in Crucinio et al. (2023), one can show that both MALA and proximal MALA have the same optimal scaling for this example.

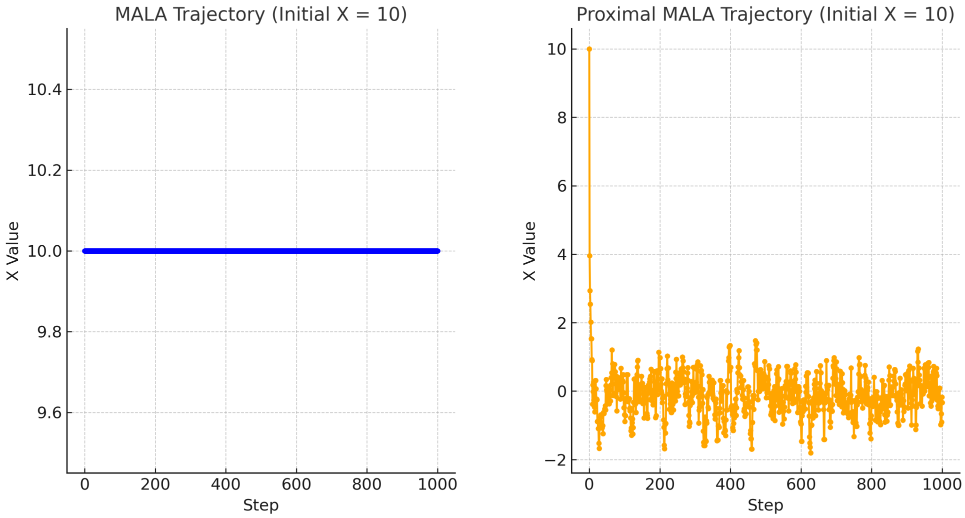

Figure 1: Trajectories of the MALA and Proximal MALA algorithms for the Poisson regression example. Both chains were initialized at $X = 10$.

1.3 Main Results

We study the optimal scaling of proximal MALA in two contexts:

1. When the target measure is a product of standard Gaussians, in Theorem 2 we show that the optimal scale and optimal acceptance probability for the proximal MALA algorithm is identical to that of MALA. The recent work Crucinio et al. (2023) extends our results in this case to a much wider class of densities.

2. For a class of infinite dimensional non-product measures studied in Pillai et al. (2012), we show that the optimal scaling of $N^{-1/3}$ for MALA as worked out in Roberts and

Rosenthal (1998); Pillai et al. (2012) is also optimal for the proximal MALA algorithm when the log density is convex and differentiable; see Theorem 10 for the formal statement of our main result.

The results of our paper thus give the following practically useful guideline: for smooth target densities where the gradient is expensive to compute or numerically unstable while implementing MALA, users may replace the gradient with the corresponding proximal function *without* losing any efficiency gains from optimal scaling; furthermore, users can set the proximal parameter δ to $N^{-1/3}$ and tune the algorithm to have an acceptance probability of 0.574 just as in MALA. Of course, as discussed in Section 1.4, optimal scaling alone does not give a complete picture for preferring one algorithm to the other. Our paper takes the first theoretical steps using convex optimization to study optimal scaling of MCMC algorithms.

1.4 High-level explanation behind the optimal scaling

Let us give a high-level explanation of why the proximal MALA enjoys the same scaling as that of MALA when Ψ is differentiable. When Ψ is smooth, it can be shown under reasonable assumptions on the second derivative of Ψ that:

$$|\mathrm{Prox}_{\Psi}^{\delta}(x) - x| = \mathcal{O}(\delta). \tag{15}$$

Consequently, setting $\lambda = \delta$ in the implicit Euler identity (9) and using (15) yields that

$$\begin{aligned}\mathrm{Prox}_{\Psi}^{\delta}(x) &= x - \delta\nabla\Psi(\mathrm{Prox}_{\Psi}^{\delta}(x))\\ &= x - \delta\nabla\Psi(x) + R(x,\delta), \qquad R(x,\delta) = \mathcal{O}(\delta^2).\end{aligned} \tag{16}$$

The remainder term $R(x,\delta)$ is $\mathcal{O}(\delta^2)$. Comparing this with (14), we see that the proximal MALA proposal can be written as

$$\begin{aligned} y &= x - \delta\nabla\Psi(x) + R(x,\delta) + \sqrt{2\delta}\, Z^N, \qquad Z^N \sim \mathrm{N}(0, \mathrm{I}_N) \\ &= x_{\mathrm{MALA}} + R(x,\delta) \end{aligned} \tag{17}$$

where x_{MALA} is the MALA proposal. In high dimensions, the drift term in the diffusion limit comes from $\mathcal{O}(\delta)$ term; the $\mathcal{O}(\delta^2)$ remainder term $R(x,\delta)$ does not contribute to the diffusion limit and vanishes in the large N limit. Our paper formalizes this observation for a class of infinite dimensional models studied in Pillai et al. (2012); refer to Equation (47), Lemma 19 and the related discussion in Section 4.1.

Remark 1 *Another important theoretical aspect is to study the mixing times of proximal MALA algorithms and obtaining non-asymptotic guarantees. See Durmus et al. (2018) and the references therein. As in the original scaling papers Roberts et al. (1997) and Roberts and Rosenthal (1998), we also do not study the mixing times of the proximal Markov chains in this paper.*

1.5 Infinite Dimensional Diffusions

Motivated by applications in data assimilation, inverse problems and Bayesian nonparametrics (see Stuart (2010) and Hairer et al. (2011)), the papers Mattingly et al. (2012) and

Pillai et al. (2012) extended the results of product measures Roberts and Rosenthal (1998) to certain infinite dimensional *non-product* target measures. In both of these papers, the target measure of interest, π, is on an infinite dimensional real separable Hilbert space $\mathcal{H}$ and is absolutely continuous with respect to a Gaussian measure π_0 on $\mathcal{H}$ with mean zero and covariance operator $\mathcal{C}$. This framework for the analysis of MCMC in high dimensions was first studied in the papers Beskos et al. (2008, 2009); Beskos and Stuart (2009). The Radon-Nikodym derivative defining the target measure is assumed to have the form

$$\frac{d\pi}{d\pi_0}(x) = M_\Psi \exp(-\Psi(x)) \tag{18}$$

for a real-valued functional $\Psi : \mathcal{H}^s \mapsto \mathbb{R}$ defined on a subspace $\mathcal{H}^s \subset \mathcal{H}$ that contains the support of the reference measure π_0; here M_Ψ is a normalizing constant.

It is proved in G. Da Prato and J. Zabczyk (1992); Hairer et al. (2005, 2007) that the measure π is invariant for $\mathcal{H}-$valued SDEs (or stochastic PDEs – SPDEs) with the form

$$\frac{dz}{dt} = -h(\ell)\big(z + \mathcal{C}\nabla\Psi(z)\big) + \sqrt{2\,h(\ell)}\,\frac{dW}{dt}, \quad z(0) = z^0 \tag{19}$$

where W is a Brownian motion (see G. Da Prato and J. Zabczyk (1992)) in $\mathcal{H}$ with covariance operator $\mathcal{C}$ and any constant $h(\ell) > 0$.

In Pillai et al. (2012), the MALA algorithm was studied when applied to a sequence of finite dimensional approximations of π as in (18). The continuous time interpolant of the Markov chain z^N given by (5) is shown to converge weakly to z solving (19) in $C([0,T];\mathcal{H}^s)$. Furthermore, the scaling of the proposal variance which achieves this scaling limit is inversely proportional to $N^{1/3}$ (*i.e.*, corresponds to the exponent $\gamma = 1/3$) and the speed of the limiting diffusion process is maximized at the same universal acceptance probability of 0.574 that was found for product measures Roberts and Rosenthal (1998).

1.6 Notation

Throughout the paper we use the following notation in order to compare sequences and to denote conditional expectations.

- Two sequences $\{\alpha_n\}$ and $\{\beta_n\}$ satisfy $\alpha_n \lesssim \beta_n$ if there exists a constant $K > 0$ satisfying $\alpha_n \le K\beta_n$ for all $n \ge 0$. The notations $\alpha_n \asymp \beta_n$ means that $\alpha_n \lesssim \beta_n$ and $\beta_n \lesssim \alpha_n$.
- Two sequences of real functions $\{f_n\}$ and $\{g_n\}$ defined on the same set D satisfy $f_n \lesssim g_n$ if there exists a constant $K > 0$ satisfying $f_n(x) \le K g_n(x)$ for all $n \ge 0$ and all $x \in D$. The notations $f_n \asymp g_n$ means that $f_n \lesssim g_n$ and $g_n \lesssim f_n$.
- The notation $\mathbb{E}_x\big[f(x,\xi)\big]$ denotes expectation with respect to ξ with the variable x fixed.

2. A Simple Example: Product of Gaussians

We start with a simple case, where the target measure is the product of standard Gaussians:

$$\pi^N(x) \propto \prod_{i=1}^N \exp(-x_i^2/2). \tag{20}$$

The MALA proposal for π^N given in (20) is:

$$y = x(1-\delta) + \sqrt{2\delta}\, Z, \qquad Z \sim \mathrm{N}(0, \mathrm{I}_N).$$

The Metropolis-Hastings acceptance ratio $\alpha(x,y)$ given in (3) with

$$q(x,y) = \prod_{i=1}^{N} \exp\Big(- \frac{1}{4\delta}\big(y_i - x_i(1-\delta)\big)^2\Big).$$

The usual calculation for finding the optimal scale proceeds as follows. Expanding the term $L_n \equiv \log\left(\frac{\pi^N(y)q(y,x)}{\pi^N(x)q(x,y)}\right)$ in δ yields [3]:

$$L_n = -\frac{\delta^{3/2}}{\sqrt{2}} \sum_{i=1}^{N} x_i Z_i + \frac{1}{2}\delta^2 \sum_{i=1}^{N} \left(x_i^2 - Z_i^2\right) + \frac{\delta^{5/2}}{\sqrt{2}} \sum_{i=1}^{N} x_i Z_i - \frac{\delta^3}{4} \sum_{i=1}^{N} x_i^2 + \mathcal{O}\left(\delta^{7/2}\right). \quad (21)$$

Since the chain is at stationarity, the first three summands in (21) have expectations zero:

$$\mathbb{E}^{\pi^N}\mathbb{E}_x(xZ_i) = \mathbb{E}^{\pi^N}\mathbb{E}_x\left(x_i^2 - Z_i^2\right) = \mathbb{E}^{\pi^N}\mathbb{E}_x(x_i Z_i) = 0.$$

Moreover, the variance of the coefficient of the $\mathcal{O}(\delta^{3/2})$ term satisfies:

$$\mathrm{Var}_x(\sum_{i=1}^{N} x_i Z_i) = \sum_{i=1}^{N} x_i^2.$$

Thus if we set $\delta = \ell N^{-1/3}$, using the fact that $\frac{1}{N}\sum_{i=1}^{N} x_i^2 \to 1$ almost surely, we obtain that

$$L_n \Longrightarrow Z_\ell \sim \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2}) \quad (22)$$

and the acceptance probability:

$$\mathbb{E}(1 \wedge e^{L_n}) \to a(\ell) \equiv \mathbb{E}(1 \wedge e^{Z_\ell}).$$

In particular, $L_n = \mathcal{O}(1)$ for $\delta = N^{-1/3}$, and thus the *optimal scale* that makes the size of acceptance probability equal to $\mathcal{O}(1)$ corresponds to $\delta = N^{-1/3}$. The ongoing computation generalizes for quite a large class of product measures π^N far beyond Gaussians, and forms the basis of the diffusion limit obtained in Roberts and Rosenthal (1998). Finally, to have the optimal acceptance probability of 0.574 that maximizes the speed of the limiting diffusion, all one needs to verify is that the limiting Gaussian random variable Z_ℓ satisfies:

$$-2\mathbb{E}(Z_\ell) = \mathrm{Var}(Z_\ell). \quad (23)$$

Indeed, once we have the relation (23), the limiting diffusion has the speed measure:

$$h(\ell) = \ell^2\mathbb{E}(1 \wedge e^{Z_\ell}) = 2\ell^2\Phi(-\frac{K}{2}\ell^3)$$

[3] We used MATHEMATICA for obtaining this expansion; also see Roberts and Rosenthal (1998).

for some constant K that depends on the target measure and Φ is the CDF of the standard Gaussian distribution. As shown in Theorem 2 of Roberts and Rosenthal (1998), the value of ℓ that maximizes $h(\ell)$ is independent of K since making the transformation $u = \frac{K}{2}\ell^3$ yields that

$$\max_{\ell} h(\ell) = 2^{5/3} K^{-2/3} \max_{u} u^{2/3} \Phi(-u)$$

and the maximizer $\hat{u}$ of the latter term is independent of K, see Theorem 2 of Roberts and Rosenthal (1998). Thus the optimal acceptance probability is also independent of K: it is just $\hat{a} = 2\Phi(-\hat{u})$.

Next, we perform the same computation for the proximal MALA algorithm. The proximal MALA proposal for π^N given in (20) is:

$$y = \frac{1}{(1+\delta)} x + \sqrt{2\delta}\, Z, \qquad Z \sim \mathrm{N}(0, \mathrm{I}_N) \tag{24}$$

with the corresponding $q(x,y)$:

$$q(x,y) = \prod_{i=1}^{N} \exp\Big(-\frac{1}{4\delta}\big(y_i - \frac{x_i}{(1+\delta)}\big)^2 \Big).$$

Theorem 2 *For the proximal MALA proposal given in* (24)*, the choice of* $\delta = \ell N^{-1/3}$ *yields an acceptance probability of* $\mathcal{O}(1)$*. The limiting acceptance probability* $a(\ell)$ *can be expressed as* $a(\ell) = \mathbb{E}(1 \wedge e^{\tilde{Z}_\ell})$ *where* $\tilde{Z}_\ell$ *is a Gaussian variable satisfying* (23)*.*

Proof As before, expanding $L_n \equiv \log\left(\frac{\pi^N(y)q(y,x)}{\pi^N(x)q(x,y)}\right)$ in terms of δ yields:

$$\begin{aligned} L_n = &-\frac{3}{\sqrt{2}}\delta^{3/2} \sum_{i=1}^{N} x_i Z_i + \frac{3}{2}\delta^2 \sum_{i=1}^{N} \left(x_i^2 - Z_i^2\right) \\ &+ \delta^{5/2} \frac{7}{\sqrt{2}} \sum_{i=1}^{N} x_i Z_i + \frac{1}{4}\delta^3 \sum_{i=1}^{N} \left(8z_i^2 - 17x_i^2\right) + \mathcal{O}\left(\delta^{7/2}\right). \end{aligned} \tag{25}$$

Again, using the fact that the chain is at stationarity, we see that the summands of $\delta^{3/2}, \delta^2$ and $\delta^{5/2}$ in the expansion (25) all have mean zero. Furthermore, for the choice of $\delta = \ell N^{-1/3}$, we have $L_n \Longrightarrow \tilde{Z}_\ell$ with $\frac{9}{2} = -2\mathbb{E}(\tilde{Z}_\ell) = \mathrm{Var}(\tilde{Z}_\ell)$ satisfying (23), and the proof is finished. ■

While we do not prove a diffusion limit, the arguments laid out in Section 1.4 can be used to prove a diffusion limit for any single component of the piecewise interpolant of the proximal Markov chain described above. Consequently, Theorem 2 yields that the optimal acceptance probability for proximal MALA algorithm is also 0.574 in the case where the target measure is the product of Gaussians.

Remark 3 *While Theorem 2 is only worked out for product of Gaussians, the result and the heuristic arguments given in Section 1.4 strongly suggest that the same optimal scale* *and acceptance probability should hold for a large class of measures obtained as products of smooth, log-concave target densities; this was rightly confirmed in Crucinio et al. (2023). This is because the optimal scale and optimal acceptance probability results are "universal"; the specifics of target distributions should not matter. In particular, the Gaussian distribution (as used in Theorem 2) plays no special role in optimality of MALA and nor should play a role here. We focused on this case for clarity of exposition.*

3. Infinite Dimensional Target Measure

We keep the framework in this paper **identical** to that of Pillai et al. (2012) so that the reader can easily compare our results to that of the MALA algorithm obtained in that paper. The structure of proof of the diffusion limit is also identical to that of Pillai et al. (2012). Recall that our main goal is to show that the proximal MALA proposal has the same performance as that of the infinite dimensional MALA algorithm studied in Pillai et al. (2012). Thus we are not interested in reproving the results of Pillai et al. (2012); instead, we merely wish to highlight only those parts where adding a proximal term (instead of the gradient) in the MALA leads to an alteration of the proof of diffusion limit worked out in Pillai et al. (2012).

Let $\mathcal{H}$ be a separable Hilbert space of real valued functions with scalar product denoted by $\langle \cdot, \cdot \rangle$ and associated norm $\|x\|^2 = \langle x, x \rangle$. Consider a Gaussian probability measure π_0 on $(\mathcal{H}, \| \cdot \|)$ with covariance operator $\mathcal{C}$. The general theory of Gaussian measures G. Da Prato and J. Zabczyk (1992) ensures that the operator $\mathcal{C}$ is positive and trace class. Let $\{\varphi_j, \lambda_j^2\}_{j \geq 1}$ be the eigenfunctions and eigenvalues of the covariance operator $\mathcal{C}$:

$$\mathcal{C}\varphi_j = \lambda_j^2 \varphi_j, \qquad\qquad j \geq 1.$$

We assume a normalization under which the family $\{\varphi_j\}_{j \geq 1}$ forms a complete orthonormal basis in the Hilbert space $\mathcal{H}$, which we refer to us as the Karhunen-Loève basis. Any function $x \in \mathcal{H}$ can be represented in this basis via the expansion

$$x = \sum_{j=1}^{\infty} x_j \, \varphi_j, \qquad x_j \overset{\text{def}}{=} \langle x, \varphi_j \rangle. \tag{26}$$

Throughout this paper we will often identify the function x with its coordinates $\{x_j\}_{j=1}^{\infty} \in \ell^2$ in this eigenbasis, moving freely between the two representations. The Karhunen-Loève expansion (see G. Da Prato and J. Zabczyk (1992), section *White Noise expansions*), refers to the fact that a realization x from the Gaussian measure π_0 can be expressed by allowing the coordinates $\{x_j\}_{j \geq 1}$ in (26) to be independent random variables distributed as $x_j \sim \mathrm{N}(0, \lambda_j^2)$. Thus, in the coordinates $\{x_j\}_{j \geq 1}$, the Gaussian reference measure π_0 has a product structure.

For every $x \in \mathcal{H}$ we have the representation (26). Using this expansion, we define Sobolev-like spaces $\mathcal{H}^r, r \in \mathbb{R}$, with the inner-products and norms defined by

$$\langle x, y \rangle_r \overset{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r} x_j y_j, \qquad\qquad \|x\|_r^2 \overset{\text{def}}{=} \sum_{j=1}^{\infty} j^{2r} \, x_j^2. \tag{27}$$

Notice that $\mathcal{H}^0 = \mathcal{H}$ and $\mathcal{H}^r \subset \mathcal{H} \subset \mathcal{H}^{-r}$ for any $r > 0$. The Hilbert-Schmidt norm $\| \cdot \|_{\mathcal{C}}$ associated to the covariance operator $\mathcal{C}$ is defined as

$$\|x\|_{\mathcal{C}}^2 = \sum_j \lambda_j^{-2} x_j^2.$$

For $x, y \in \mathcal{H}^r$, the outer product operator in $\mathcal{H}^r$ is the operator $x \otimes_{\mathcal{H}^r} y : \mathcal{H}^r \to \mathcal{H}^r$ defined by $(x \otimes_{\mathcal{H}^r} y)z \stackrel{\text{def}}{=} \langle y, z \rangle_r \, x$ for every $z \in \mathcal{H}^r$. For $r \in \mathbb{R}$, let $B_r : \mathcal{H} \mapsto \mathcal{H}$ denote the operator which is diagonal in the basis $\{\varphi_j\}_{j \geq 1}$ with diagonal entries j^{2r}. The operator B_r satisfies $B_r \, \varphi_j = j^{2r} \varphi_j$ so that $B_r^{\frac{1}{2}} \, \varphi_j = j^r \varphi_j$. The operator B_r lets us alternate between the Hilbert space $\mathcal{H}$ and the Sobolev spaces $\mathcal{H}^r$ via the identities $\langle x, y \rangle_r = \langle B_r^{\frac{1}{2}} x, B_r^{\frac{1}{2}} y \rangle$. Since $\|B_r^{-1/2} \varphi_k\|_r = \|\varphi_k\| = 1$, we deduce that $\{B_r^{-1/2} \varphi_k\}_{k \geq 0}$ forms an orthonormal basis for $\mathcal{H}^r$. For a positive, self-adjoint operator $D : \mathcal{H} \mapsto \mathcal{H}$, we define its trace in $\mathcal{H}^r$ by

$$\mathrm{Tr}_{\mathcal{H}^r}(D) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \langle (B_r^{-\frac{1}{2}} \varphi_j), D(B_r^{-\frac{1}{2}} \varphi_j) \rangle_r. \tag{28}$$

Since $\mathrm{Tr}_{\mathcal{H}^r}(D)$ does not depend on the orthonormal basis, the operator D is said to be trace class in $\mathcal{H}^r$ if $\mathrm{Tr}_{\mathcal{H}^r}(D) < \infty$ for some, and hence any, orthonormal basis of $\mathcal{H}^r$. Let us define the operator $\mathcal{C}_r \stackrel{\text{def}}{=} B_r^{1/2} \, \mathcal{C} \, B_r^{1/2}$. Notice that $\mathrm{Tr}_{\mathcal{H}^r}(\mathcal{C}_r) = \sum_{j=1}^{\infty} \lambda_j^2 \, j^{2r}$. In Pillai et al. (2012) it is shown that under the condition

$$\mathrm{Tr}_{\mathcal{H}^r}(\mathcal{C}_r) < \infty, \tag{29}$$

the support of π_0 is included in $\mathcal{H}^r$ in the sense that π_0-almost every function $x \in \mathcal{H}$ belongs to $\mathcal{H}^r$. Furthermore, the induced distribution of π_0 on $\mathcal{H}^r$ is identical to that of a centered Gaussian measure on $\mathcal{H}^r$ with covariance operator $\mathcal{C}_r$. For example, if $\xi \stackrel{\mathcal{D}}{\sim} \pi_0$, then $\mathbb{E}\big[\langle \xi, u \rangle_r \langle \xi, v \rangle_r\big] = \langle u, \mathcal{C}_r v \rangle_r$ for any functions $u, v \in \mathcal{H}^r$. Thus in what follows, we alternate between the Gaussian measures $\mathrm{N}(0, \mathcal{C})$ on $\mathcal{H}$ and $\mathrm{N}(0, \mathcal{C}_r)$ on $\mathcal{H}^r$, for those r for which (29) holds.

3.1 Change of Measure

Our goal is to sample from a measure π defined through the change of probability formula (18). As described above, the condition $\mathrm{Tr}_{\mathcal{H}^r}(\mathcal{C}_r) < \infty$ implies that the measure π_0 has full support on $\mathcal{H}^r$, *i.e.*, $\pi_0(\mathcal{H}^r) = 1$. Consequently, if $\mathrm{Tr}_{\mathcal{H}^r}(\mathcal{C}_r) < \infty$, the functional $\Psi(\cdot)$ needs only to be defined on $\mathcal{H}^r$ in order for the change of probability formula (18) to be valid. In this section we give assumptions on the decay of the eigenvalues of the covariance operator $\mathcal{C}$ of π_0 that ensure the existence of a real number $s > 0$ such that π_0 has full support on $\mathcal{H}^s$. The functional $\Psi(\cdot)$ is assumed to be defined on $\mathcal{H}^s$ and we impose regularity assumptions on $\Psi(\cdot)$ that ensure that the probability distribution π is not too different from π_0, when projected into directions associated with φ_j for j large. For each $x \in \mathcal{H}^s$ the derivative $\nabla \Psi(x)$ is an element of the dual $(\mathcal{H}^s)^*$ of $\mathcal{H}^s$ comprising linear functionals on $\mathcal{H}^s$. However, we may identify $(\mathcal{H}^s)^*$ with $\mathcal{H}^{-s}$ and view $\nabla \Psi(x)$ as an element of $\mathcal{H}^{-s}$ for each $x \in \mathcal{H}^s$. With this identification, the following identity holds

$$\|\nabla \Psi(x)\|_{\mathcal{L}(\mathcal{H}^s, \mathbb{R})} = \|\nabla \Psi(x)\|_{-s}$$

and the second derivative $\partial^2\Psi(x)$ can be identified as an element of $\mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$. To avoid technicalities we assume that $\Psi(\cdot)$ is quadratically bounded, with first derivative linearly bounded and second derivative globally bounded. Weaker assumptions could be dealt with by use of stopping time arguments.

Assumptions 4 *The covariance operator $\mathcal{C}$ and functional Ψ satisfy the following:*

1. ***Decay of Eigenvalues λ_j^2 of $\mathcal{C}$:*** *there is an exponent $\kappa > \frac{1}{2}$ such that*

$$\lambda_j \asymp j^{-\kappa}. \tag{30}$$

2. ***Assumptions on Ψ:*** *The function Ψ is convex. There exist constants $M_i \in \mathbb{R}_+, i \leq 4$ and $s \in [0, \kappa - 1/2)$ such that for all $x \in \mathcal{H}^s$ the functional $\Psi : \mathcal{H}^s \to \mathbb{R}$ satisfies*

$$M_1 \leq \Psi(x) \leq M_2 \Big(1 + \|x\|_s^2\Big) \tag{31}$$

$$\|\nabla\Psi(x)\|_{-s} \leq M_3 \Big(1 + \|x\|_s\Big) \tag{32}$$

$$\|\partial^2\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s,\mathcal{H}^{-s})} \leq M_4. \tag{33}$$

Remark 5 *The convexity of Ψ is not assumed in Pillai et al. (2012). It is not required for the MALA algorithm. In this paper we assume the convexity of Ψ so as to get a unique value for the proximal operator. This assumption is not strictly necessary for our methods to go through. However, since our key aim is to formalize the observation made in* (17)*, we avoid additional complications.*

Remark 6 *The condition $\kappa > \frac{1}{2}$ ensures that the covariance operator $\mathcal{C}$ is trace class in $\mathcal{H}$. In fact, Equation* (29) *shows that $\mathcal{C}_r$ is trace-class in $\mathcal{H}^r$ for any $r < \kappa - \frac{1}{2}$. It follows that π_0 has full measure in $\mathcal{H}^r$ for any $r \in [0, \kappa - 1/2)$. In particular π_0 has full support on $\mathcal{H}^s$.*

Remark 7 *The functional $\Psi(x) = \frac{1}{2}\|x\|_s^2$ satisfies Assumptions 4. It is convex, defined on $\mathcal{H}^s$ and its derivative at $x \in \mathcal{H}^s$ is given by $\nabla\Psi(x) = \sum_{j\geq 0} j^{2s} x_j \varphi_j \in \mathcal{H}^{-s}$ with $\|\nabla\Psi(x)\|_{-s} = \|x\|_s$. The second derivative $\partial^2\Psi(x) \in \mathcal{L}(\mathcal{H}^s, \mathcal{H}^{-s})$ is the linear operator that maps $u \in \mathcal{H}^s$ to $\sum_{j\geq 0} j^{2s}\langle u, \varphi_j\rangle \varphi_j \in \mathcal{H}^s$: its norm satisfies $\|\partial^2\Psi(x)\|_{\mathcal{L}(\mathcal{H}^s,\mathcal{H}^{-s})} = 1$ for any $x \in \mathcal{H}^s$.*

3.2 Finite Dimensional Approximation

We are interested in finite dimensional approximations of the probability distribution π. To this end, we introduce the vector space spanned by the first N eigenfunctions of the covariance operator,

$$X^N \stackrel{\text{def}}{=} \text{span}\big\{\varphi_1, \varphi_2, \ldots, \varphi_N\big\}.$$

Notice that $X^N \subset \mathcal{H}^r$ for any $r \in [0; +\infty)$. In particular, X^N is a subspace of $\mathcal{H}^s$. Next, we define N-dimensional approximations of the functional $\Psi(\cdot)$ and of the reference measure

π_0. To this end, we introduce the orthogonal projection on X^N denoted by $P^N : \mathcal{H}^s \mapsto X^N \subset \mathcal{H}^s$. The functional $\Psi(\cdot)$ is approximated by the functional $\Psi^N : X^N \mapsto \mathbb{R}$ defined by

$$\Psi^N \overset{\text{def}}{=} \Psi \circ P^N. \tag{34}$$

The approximation π_0^N of the reference measure π_0 is the Gaussian measure on X^N given by the law of the random variable

$$\pi_0^N \overset{\mathcal{D}}{\sim} \sum_{j=1}^N \lambda_j \xi_j \varphi_j \;=\; (\mathcal{C}^N)^{\frac{1}{2}}\, \xi^N$$

where ξ_j are i.i.d standard Gaussian random variables, $\xi^N = \sum_{j=1}^N \xi_j \varphi_j$ and $\mathcal{C}^N = P^N \circ \mathcal{C} \circ P^N$. Consequently we have $\pi_0^N = \mathrm{N}(0, \mathcal{C}^N)$. Finally, one can define the approximation π^N of π by the change of probability formula

$$\frac{d\pi^N}{d\pi_0^N}(x) = M_{\Psi^N} \exp\big(-\Psi^N(x)\big) \tag{35}$$

where M_{Ψ^N} is a normalization constant. Notice that the probability distribution π^N is supported on X^N and has Lebesgue density[4] on X^N equal to

$$\pi^N(x) \;\propto\; \exp\Big(-\frac{1}{2}\|x\|_{\mathcal{C}^N}^2 - \Psi^N(x)\Big). \tag{36}$$

In formula (36), the Hilbert-Schmidt norm $\|\cdot\|_{\mathcal{C}^N}$ on X^N is given by the scalar product $\langle u, v\rangle_{\mathcal{C}^N} = \langle u, (\mathcal{C}^N)^{-1} v\rangle$ for all $u, v \in X^N$. The operator $\mathcal{C}^N$ is invertible on X^N because the eigenvalues of $\mathcal{C}$ are assumed to be strictly positive. The quantity $\mathcal{C}^N \nabla \log \pi^N(x)$ is repeatedly used in the text and in particular appears in the function $\mu^N(x)$ given by

$$\mu^N(x) = -\Big(P^N x + \mathcal{C}^N \nabla \Psi^N(x)\Big) \tag{37}$$

which is $\mathcal{C}^N \nabla \log \pi^N(x)$. This function is the drift of an ergodic Langevin diffusion that leaves π^N invariants. Similarly, one defines the function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ given by

$$\mu(x) = -\Big(x + \mathcal{C} \nabla \Psi(x)\Big) \tag{38}$$

which is $\mathcal{C} \nabla \log \pi(x)$. In Lemmas 4.1 and 4.3 of Pillai et al. (2012), it is shown that for π_0-almost every function $x \in \mathcal{H}$, we have $\lim_{N\to\infty} \mu^N(x) = \mu(x)$; see Section 7.1 below. This quantifies the manner in which $\mu^N(\cdot)$ is an approximation of $\mu(\cdot)$.

The next lemma gathers various regularity estimates on the functional $\Psi(\cdot)$ and $\Psi^N(\cdot)$ that are repeatedly used in the sequel. These are simple consequences of Assumptions 4 and proofs can be found in Mattingly et al. (2012) and Pillai et al. (2012).

Lemma 8 (Properties of Ψ) *Let the functional $\Psi(\cdot)$ satisfy Assumptions 4 and consider the functional $\Psi^N(\cdot)$ defined by Equation* (34). *The following estimates hold.*

[4] For ease of notation we do not distinguish between a measure and its density, nor do we distinguish between the representation of the measure in X^N or in coordinates in $\mathbb{R}^N$

1. *The functionals $\Psi^N : \mathcal{H}^s \to \mathbb{R}$ satisfy the same conditions imposed on Ψ given by Equations* (31), (32) *and* (33) *with constants that can be chosen independent of N.*

2. *The function $\mathcal{C}\nabla\Psi : \mathcal{H}^s \to \mathcal{H}^s$ is globally Lipschitz on $\mathcal{H}^s$: there exists a constant $M_5 > 0$ such that*

   $$\|\mathcal{C}\nabla\Psi(x) - \mathcal{C}\nabla\Psi(y)\|_s \le M_5 \|x-y\|_s \qquad \forall x, y \in \mathcal{H}^s.$$

   Moreover, the functions $\mathcal{C}^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ also satisfy this estimate with a constant that can be chosen independent of N.

3. *The functional $\Psi(\cdot) : \mathcal{H}^s \to \mathbb{R}$ satisfies a "one-sided" Taylor formula*[5]*. There exists a constant $M_6 > 0$ such that*

   $$\Psi(y) - \Big(\Psi(x) + \langle \nabla\Psi(x), y-x\rangle\Big) \le M_6 \|x-y\|_s^2 \qquad \forall x, y \in \mathcal{H}^s. \tag{39}$$

   Moreover, the functionals $\Psi^N(\cdot)$ also satisfy the above estimates with a constant that can be chosen independent of N.

Remark 9 *The regularity Lemma 8 shows in particular that the function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ defined by* (38) *is globally Lipschitz on $\mathcal{H}^s$. Similarly, it follows that $\mathcal{C}^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ and $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ given by* (37) *are globally Lipschitz with Lipschitz constants that can be chosen uniformly in N.*

4. The proximal MALA in Hilbert space

In this section, we construct a version of the proximal MALA algorithm of Pereyra (2016) in the Hilbert space $\mathcal{H}^s$. The proximal operators are well defined in an infinite dimensional Hilbert space. The reader is referred to Bauschke and Combettes (2011) for a book length treatment. For a function $g : \mathcal{H}^s \mapsto (-\infty, \infty]$ and $\lambda > 0$, define the proximal function

$$\mathrm{Prox}_g^\lambda(x) = \mathrm{argmin}_{y\in\mathcal{H}^s}\Big(g(y) + \frac{1}{2\lambda}\|x-y\|_s^2\Big). \tag{40}$$

If g is convex, Proposition 12.15 of Bauschke and Combettes (2011) yields that $\mathrm{Prox}_g^\lambda(x)$ is convex and differentiable. Moreover the minimizer in (40) is unique due to the convexity of g. Similar to (8), define the function E_g^λ:

$$\mathrm{E}_g^\lambda(x) \propto \exp\Big\{-g\Big(\mathrm{Prox}_g^\lambda(x)\Big)\Big\}\exp\Big\{-\frac{1}{2\lambda}\|\mathrm{Prox}_g^\lambda(x) - x\|^2\Big\}. \tag{41}$$

The function $-\log \mathrm{E}_g^\lambda(x)$ is the λ-Moreau-Yoshida envelope of g. We also have the identity (Bauschke and Combettes (2011), Proposition 12.29 and Corollary 17.6):

$$-\nabla \log \mathrm{E}_g^\lambda(x) = \frac{1}{\lambda}(x - \mathrm{Prox}_g^\lambda(x)) = \nabla g(\mathrm{Prox}_g^\lambda(x)). \tag{42}$$

[5] We extend $\langle\cdot,\cdot\rangle$ from an inner-product on $\mathcal{H}$ to the dual pairing between $\mathcal{H}^{-s}$ and $\mathcal{H}^s$.

4.1 The Proximal-MALA Algorithm

Recall from (36) that our target measure is

$$\pi^N(x) \;\propto\; \exp\Big(-\frac{1}{2}\|x\|^2_{\mathcal{C}^N} - \Psi^N(x)\Big).$$

Our algorithm is motivated by the fact that the probability measure π^N defined by Equation (35) is invariant with respect to the Langevin diffusion process

$$\begin{aligned}\frac{dz}{dt} &= \mathcal{C}^N \nabla \log \pi^N(z) + \sqrt{2}\,\frac{dW^N}{dt} \\ &= \mathcal{C}^N \mu^N(z) + \sqrt{2}\,\frac{dW^N}{dt}\end{aligned} \tag{43}$$

where W^N is a Brownian motion in $\mathcal{H}^s$ with covariance operator $\mathcal{C}^N$ and μ^N is as defined in (37).

To obtain a proximal algorithm that is analogous to Pereyra's algorithm given in (14), we replace $\Psi^N(x)$ by its δ-Moreau-Yoshida envelope to obtain:

$$\pi^N_\lambda(x) \propto \exp\Big(-\frac{1}{2}\|x\|^2_{\mathcal{C}^N}\Big)\mathrm{E}^\lambda_{\Psi^N}(x)$$

where $\mathrm{E}^\delta_{\Psi^N}(x)$ is defined as in Equation (41) with $g = \Psi^N$. Now the usual MALA proposal for $\pi^N_\lambda(x)$ gives the analogue to Pereyra's proximal algorithm. Indeed, the MALA proposal for $\pi^N_\lambda(x)$ (with the choice of $\lambda = \delta$) gives

$$\begin{aligned} y &= x + \delta\,\mathcal{C}^N \nabla \log \pi^N_\lambda(x) + \sqrt{2\delta}\,(\mathcal{C}^N)^{\frac{1}{2}}\xi^N \\ &= x + \delta\,\mathcal{C}^N\Big(-(\mathcal{C}^N)^{-1}x + \nabla \log \mathrm{E}^\delta_{\Psi^N}(x)\Big) + \sqrt{2\delta}\,(\mathcal{C}^N)^{\frac{1}{2}}\xi^N \\ &= (1 - \delta - \mathcal{C}^N)x + \mathcal{C}^N \mathrm{Prox}^\delta_{\Psi^N}(x) + \sqrt{2\delta}\,(\mathcal{C}^N)^{\frac{1}{2}}\xi^N \qquad \delta = \ell N^{-\frac{1}{3}} \\ &\equiv x_{\mathrm{Prox-MALA}} \end{aligned} \tag{44}$$

where the third equality follows from Equations (41) and (42).

Applying (42) with $\Psi = \Psi^N$ and $\lambda = \delta$, we obtain that

$$\begin{aligned}\mathrm{Prox}^\delta_{\Psi^N}(x) &= x - \delta \nabla \Psi^N(\mathrm{Prox}^\delta_{\Psi^N}(x)) \\ &\approx x - \delta \nabla \Psi^N(x) + \mathcal{O}(\delta^2).\end{aligned}$$

Consequently, on X^N,

$$\begin{aligned}(1 - \delta - \mathcal{C}^N)x + \mathcal{C}^N \mathrm{Prox}^\delta_{\Psi^N}(x) &\approx x - \delta(P^N x + \mathcal{C}^N \nabla \Psi^N(x)) \\ &= x + \delta \mu^N(x).\end{aligned} \tag{45}$$

Let

$$x_{\mathrm{MALA}} = x + \delta \mu^N(x) + \sqrt{2\delta}\,(\mathcal{C}^N)^{\frac{1}{2}}\xi^N \qquad \text{where} \qquad \delta = \ell N^{-\frac{1}{3}} \tag{46}$$

denote the usual MALA proposal obtained from the Euler discretization of the infinite dimensional diffusion (43). Notice that $(\mathcal{C}^N)^{\frac{1}{2}}\xi^N \overset{\mathcal{D}}{\sim} \mathrm{N}(0, \mathcal{C}^N)$. The calculation done in (45) shows that our proximal MALA proposal (44) closely tags the MALA proposal:

$$x_{\text{Prox}-\text{MALA}} = x_{\text{MALA}} + R^N(x, \delta) \tag{47}$$

where the term

$$R^N(x, \delta) \equiv \delta\, \mathcal{C}^N \Big(\mathrm{Prox}^{\delta}_{\Psi^N}(x) - x \Big) \tag{48}$$

can be thought of as the added "error" induced by the proximal MALA proposal as compared to MALA. As shown in Lemma 13, we have $\|R^N(x, \delta)\|_{\mathcal{C}^N} \lesssim \delta^2(1 + \|x\|_s) = \mathcal{O}(\delta^2)$. As in the product measure case, for optimal scaling only terms of $\mathcal{O}(\delta^{3/2})$ and lower order contribute; thus the contribution from this remainder term to the scaling drops out in the large N limit. Consequently, the optimal scaling and the diffusion limits for the proximal MALA algorithm follow from the corresponding results for the MALA algorithm.

For streamlining further calculations, we will write the $x_{\text{Prox}-\text{MALA}}$ proposal from (44) as

$$y = x + \delta\mu^N(x) + R^N(x, \delta) + \sqrt{2\delta}\, (\mathcal{C}^N)^{\frac{1}{2}}\xi^N \qquad \text{where} \qquad \delta = \ell N^{-\frac{1}{3}}. \tag{49}$$

4.2 Time evolution of the proximal MALA chain

We introduce a related parameter

$$\Delta t := \ell^{-1}\delta = N^{-\frac{1}{3}}$$

which will be the natural time-step for the limiting diffusion process derived from the proposal above, after inclusion of an accept-reject mechanism. The scaling of Δt, and hence δ, with N will ensure that the average acceptance probability is $\mathcal{O}(1)$ as N grows.

Following Pillai et al. (2012), we will study the Markov chain $x^N = \{x^{k,N}\}_{k\geq 0}$ resulting from Metropolizing the proximal proposal (49) when it is started at stationarity: the initial position $x^{0,N}$ is distributed as π^N and thus lies in X^N. Therefore, the Markov chain evolves in X^N; as a consequence, only the first N components of an expansion in the eigenbasis of $\mathcal{C}$ are nonzero and the algorithm can be implemented in $\mathbb{R}^N$. However the analysis is cleaner when written in $X^N \subset \mathcal{H}^s$. The acceptance probability only depends on the first N coordinates of x and y and has the form

$$\alpha^N(x, \xi^N) = 1 \wedge \frac{\pi^N(y) T^N(y, x)}{\pi^N(x) T^N(x, y)} = 1 \wedge e^{Q^N(x, \xi^N)} \tag{50}$$

where the proposal y is given by Equation (49). The function $T^N(\cdot, \cdot)$ is the density of the Langevin proposals (49) and is given by

$$T^N(x, y) \propto \exp\Big\{ -\frac{1}{4\delta} \|y - x - \delta\mu^N(x) - R^N(x, \delta)\|^2_{\mathcal{C}^N} \Big\}.$$

The local mean acceptance probability $\alpha^N(x)$ is defined by

$$\alpha^N(x) = \mathbb{E}_x\big[\alpha^N(x, \xi^N)\big]. \tag{51}$$

It is the expected acceptance probability when the algorithm stands at $x \in \mathcal{H}$. The Markov chain $x^N = \{x^{k,N}\}_{k \geq 0}$ can also be expressed as

$$\begin{cases} y^{k,N} & = x^{k,N} + \delta\mu^N(x^{k,N}) + R^N(x^{k,N}, \delta) + \sqrt{2\delta}\,(\mathcal{C}^N)^{\frac{1}{2}}\, \xi^{k,N} \\ x^{k+1,N} & = \gamma^{k,N} y^{k,N} + (1 - \gamma^{k,N})\, x^{k,N} \end{cases} \tag{52}$$

where $\xi^{k,N}$ are i.i.d samples distributed as ξ^N and $\gamma^{k,N} = \gamma^N(x^{k,N}, \xi^{k,N})$ creates a Bernoulli random sequence with k^{th} success probability $\alpha^N(x^{k,N}, \xi^{k,N})$. We may view the Bernoulli random variable as $\gamma^{k,N} = 1_{\{U^k < \alpha^N(x^{k,N}, \xi^{k,N})\}}$ where $U^k \overset{\mathcal{D}}{\sim}$ Uniform$(0,1)$ is independent from $x^{k,N}$ and $\xi^{k,N}$.

In summary, the Markov chain that we have described in $\mathcal{H}^s$ is, when projected onto X^N, equivalent to a proximal MALA algorithm on $\mathbb{R}^N$ for the Lebesgue density (36). Recall that the target measure π in (18) is the invariant measure of the SPDE (19). Our goal is to obtain an invariance principle for the continuous interpolant (5) of the Markov chain $x^N = \{x^{k,N}\}_{k \geq 0}$ started in stationarity, *i.e*, to show weak convergence in $C([0,T]; \mathcal{H}^s)$ of $z^N(t)$ to the solution $z(t)$ of the SPDE (19), as the dimension $N \to \infty$.

5. Main Result

In this section, we present the main result of this paper. Consider the constant $\alpha(\ell) = \mathbb{E}\big[1 \wedge e^{Z_\ell}\big]$ where $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$ and define the speed function

$$h(\ell) = \ell\alpha(\ell). \tag{53}$$

The quantity $\alpha(\ell)$ represents the limiting expected acceptance probability of the MALA algorithm while $h(\ell)$ is the asymptotic speed function of the limiting diffusion.

Theorem 10 *Let the initial condition $x^{0,N}$ of the proximal MALA algorithm be such that $x^{0,N} \sim \pi^N$ and let $z^N(t)$ be a piecewise linear, continuous interpolant of the proximal MALA algorithm* (52) *with $\Delta t = N^{-1/3}$. Then, for any $T > 0$, $z^N(t)$ converges weakly in $C([0,T], \mathcal{H}^s)$ to the diffusion process $z(t)$ given by*

$$\frac{dz}{dt} = -h(\ell)\big(z + \mathcal{C}\nabla\Psi(z)\big) + \sqrt{2\,h(\ell)}\,\frac{dW}{dt}, \quad z(0) = z^0 \sim \pi \tag{54}$$

with the constant $h(\ell)$ as given in (53). *Choosing ℓ so as to maximize the speed function $h(\ell)$ leads to the acceptance probability of 0.574 for the proximal MALA algorithm.*

Remark 11 *The fact that choosing ℓ so as to maximize the speed function $h(\ell)$ leads to the optimal universal acceptance probability of 0.574 is known since Roberts and Rosenthal (1998), and is also shown in Pillai et al. (2012). Thus to prove Theorem 10, we need only establish the diffusion limit.*

5.1 Proof Strategy

The acceptance probability of the proposal (49) is equal to $\alpha^N(x, \xi^N) = 1 \wedge e^{Q^N(x, \xi^N)}$ and the quantity $\alpha^N(x) = \mathbb{E}_x[\alpha^N(x, \xi^N)]$ given by (51) represents the mean acceptance probability when the Markov chain x^N stands at x. Recall the quantity Q^N in Equation (50). This quantity may be expressed as

$$
\begin{aligned}
Q^N(x,\xi^N) = &-\frac{1}{2}\Big(\|y\|^2_{\mathcal{C}^N} - \|x\|^2_{\mathcal{C}^N}\Big) - \Big(\Psi^N(y) - \Psi^N(x)\Big) \\
&- \frac{1}{4\delta}\Big\{\|x-y-\delta\mu^N(y) - R^N(y,\delta)\|^2_{\mathcal{C}^N} - \|y-x-\delta\mu^N(x) - R^N(x,\delta)\|^2_{\mathcal{C}^N}\Big\}.
\end{aligned}
\tag{55}
$$

The main observation (also used in Pillai et al. (2012)) is that $Q^N(x,\xi^N)$ can be approximated by a Gaussian random variable

$$
Q^N(x,\xi^N) \;\approx\; Z_\ell \tag{56}
$$

where $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4},\frac{\ell^3}{2})$. These approximations are made rigorous in Lemma 16 and Lemma 17. Therefore, the Bernoulli random variable $\gamma^N(x,\xi^N)$ with success probability $1\wedge e^{Q^N(x,\xi^N)}$ can be approximated by a Bernoulli random variable, independent of x, with success probability equal to

$$
\alpha(\ell) = \mathbb{E}\big[1\wedge e^{Z_\ell}\big]. \tag{57}
$$

Thus, the limiting acceptance probability of the MALA algorithm is as given in Equation (57).

Recall that $\Delta t = N^{-\frac{1}{3}}$. With this notation we introduce the drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ given by

$$
d^N(x) = \big(h(\ell)\Delta t\big)^{-1}\mathbb{E}\big[x^{1,N} - x^{0,N}\,|x^{0,N} = x\big] \tag{58}
$$

and the martingale difference array $\{\Gamma^{k,N} : k \geq 0\}$ defined by $\Gamma^{k,N} = \Gamma^N(x^{k,N},\xi^{k,N})$ with

$$
\Gamma^{k,N} = \big(2h(\ell)\Delta t\big)^{-\frac{1}{2}}\Big(x^{k+1,N} - x^{k,N} - h(\ell)\Delta t\, d^N(x^{k,N})\Big). \tag{59}
$$

The normalization constant $h(\ell)$ defined in Equation (53) ensures that the drift function d^N and the martingale difference array $\{\Gamma^{k,N}\}$ are asymptotically independent from the parameter ℓ. The drift-martingale decomposition of the Markov chain $\{x^{k,N}\}_k$ then reads

$$
x^{k+1,N} - x^{k,N} = h(\ell)\Delta t d^N(x^{k,N}) + \sqrt{2h(\ell)\Delta t}\,\Gamma^{k,N}. \tag{60}
$$

Lemma 19 and Lemma 20 exploit the Gaussian behaviour of $Q^N(x,\xi^N)$ described in Equation (56) in order to give quantitative versions of the following approximations,

$$
d^N(x) \;\approx\; \mu(x) \qquad\text{and}\qquad \Gamma^{k,N} \;\approx\; \mathrm{N}(0,C) \tag{61}
$$

where $\mu(x) = -\Big(x + C\nabla\Psi(x)\Big)$. From Equation (60) it follows that for large N the evolution of the Markov chain ressembles the Euler discretization of the limiting diffusion (19). The next step consists of proving an invariance principle for a rescaled version of the martingale difference array $\{\Gamma^{k,N}\}$. The continuous process $W^N \in \mathcal{C}([0;T],\mathcal{H}^s)$ is defined as

$$
W^N(t) = \sqrt{\Delta t}\,\sum_{j=0}^{k}\Gamma^{j,N} \;+\; \frac{t-k\Delta t}{\sqrt{\Delta t}}\,\Gamma^{k+1,N} \qquad\text{for}\qquad k\Delta t \leq t < (k+1)\Delta t. \tag{62}
$$

The sequence of processes $\{W^N\}$ converges weakly in $\mathcal{C}([0;T],\mathcal{H}^s)$ to a Brownian motion W in $\mathcal{H}^s$ with covariance operator equal to C_s. Indeed, Proposition 21 proves the stronger result

$$(x^{0,N}, W^N) \Longrightarrow (z^0, W)$$

where $\Longrightarrow$ denotes weak convergence in $\mathcal{H}^s \times \mathcal{C}([0;T],\mathcal{H}^s)$ and $z^0 \overset{\mathcal{D}}{\sim} \pi$ is independent of the limiting Brownian motion W. Once we have the invariance principle and the converge of the drift and diffusion terms, the "Master Theorem" in Pillai et al. (2012) (see Proposition 3.1 of Pillai et al. (2012)) gives the required diffusion limit.

6. Proof of the Main Result

In this section, we give the proof of the Theorem 10. To this end, we use Proposition 3.1 of Pillai et al. (2012). According to Proposition 3.1 of Pillai et al. (2012), to show the diffusion limit, we must show the following three conditions.

1. **Convergence of initial conditions:** π^N converges in distribution to the probability measure π where π has a finite first moment, that is $\mathbb{E}^{\pi}[\|x\|_s] < \infty$.
2. **Invariance principle:** the sequence $(x^{0,N}, W^N)$ defined by Equation (62) converges weakly in $\mathcal{H}^s \times \mathcal{C}([0,T],\mathcal{H}^s)$ to (z^0, W) where $z^0 \overset{\mathcal{D}}{\sim} \pi$ and W is a Brownian motion in $\mathcal{H}^s$, independent from z^0, with covariance operator C_s.
3. **Convergence of the drift:** there exists a globally Lipschitz function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ that satisfies

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\left[\|d^N(x) - \mu(x)\|_s\right] = 0.$$

Item (1.) above follows from Lemma 4.3 of Pillai et al. (2012)); also see Section 7.1 below. Item (2.) is proved in Proposition 21. Item (3.) is proved in Lemma 19. Thus we have established all three conditions required by Proposition 3.1 of Pillai et al. (2012) and thus the proof of our main result is finished.

7. Key Estimates

In this section, we prove some key estimates for the proximal operator, and and also collect some key approximation properties of μ^N and π^N from Pillai et al. (2012). These properties will be repeatedly used throughout.

7.1 Approximation properties of μ^N and π^N

- For π_0-almost every function $x \in \mathcal{H}^s$, the approximation $\mu^N(x) \approx \mu(x)$ holds as N goes to infinity. Indeed, under Assumption 4, the sequences of functions $\mu^N : \mathcal{H}^s \to \mathcal{H}^s$ satisfies (see Lemma 4.1 of Pillai et al. (2012)),

$$\pi_0\Big(\Big\{x \in \mathcal{H}^s : \lim_{N\to\infty} \|\mu^N(x) - \mu(x)\|_s = 0 \Big\}\Big) = 1. \tag{63}$$

- Under the Assumptions 4 the normalization constants M_{Ψ^N} are uniformly bounded so that for any measurable functional $f : \mathcal{H} \mapsto \mathbb{R}$, we have from Lemma 4.3 of Pillai et al. (2012) that
$$\mathbb{E}^{\pi^N}\big[|f(x)|\big] \lesssim \mathbb{E}^{\pi_0}\big[|f(x)|\big].$$
Moreover, the sequence of probability measure π^N satisfies $\pi^N \implies \pi$ where $\implies$ denotes weak convergence in $\mathcal{H}^s$.

- By Fernique's theorem G. Da Prato and J. Zabczyk (1992), for any exponent $p \geq 0$ we have
$$\mathbb{E}^{\pi^0}\big[\|x\|_s^p\big] < \infty.$$
We also have that for any $p \geq 0$
$$\sup_{N\in\mathbb{N}} \mathbb{E}^{\pi^N}\big[\|x\|_s^p\big] \;<\; \infty.$$

7.2 Estimates involving proximal functions and the remainder term

Recall the constant M_6 from (39).

Lemma 12 *For any $x \in \mathcal{H}^s$ and $N \in \mathbb{N}$ and for all $\delta < \frac{1}{2M_6}$,*
$$\|\mathrm{Prox}^{\delta}_{\Psi^N}(x) - x\|_s \lesssim \delta(1 + \|x\|_s).$$

Proof Set $x^* = \mathrm{Prox}^{\delta}_{\Psi^N}(x)$. Since x^* minimizes the map:
$$y \mapsto \Big(\Psi^N(y) + \frac{1}{2\delta}\|y - x\|_s^2\Big),$$
from our assumptions in (39) and (32), it follows that
$$\begin{aligned}\frac{1}{2\delta}\|x^* - x\|_s^2 &\leq \Psi^N(x) - \Psi^N(x^*) = |\Psi^N(x^*) - \Psi^N(x)|\\ &\leq |\langle \nabla\Psi^N(x), x^* - x\rangle| + M_6\|x^* - x\|_s^2\\ &\leq M_3(1 + \|x\|_s)\|x^* - x\|_s + M_6\|x^* - x\|_s^2.\end{aligned}$$
Dividing by the term $\|x^* - x\|_s$ throughout and simplifying yields
$$\|x^* - x\|_s \leq \delta\frac{M_3}{(1 - 2\delta M_6)}(1 + \|x\|_s) \lesssim \delta(1 + \|x\|_s)$$
and the proof is done. ■

Lemma 13 *Recall the remainder term $R^N(x, \delta)$ from (48). For any $x \in \mathcal{H}^s$, $N \in \mathbb{N}$ and for all $\delta < \frac{1}{2M_6}$,*
$$\|R^N(x, \delta)\|_{\mathcal{C}^N} \lesssim \delta^2(1 + \|x\|_s), \qquad \|R^N(x, \delta)\|_s \lesssim \delta^2(1 + \|x\|_s).$$

Proof Set $x^* = \mathrm{Prox}^{\delta}_{\Psi^N}(x)$. Then $R^N(x,\delta) = \delta\, \mathcal{C}^N(x^* - x)$. Thus

$$\begin{aligned}\|R^N(x,\delta)\|^2_{\mathcal{C}^N} &= \langle R^N(x,\delta), (\mathcal{C}^N)^{-1} R^N(x,\delta)\rangle \\ &= \delta^2 \langle \mathcal{C}^N(x^* - x), (x^* - x)\rangle \\ &\lesssim \delta^2 \|x^* - x\|_s^2 \lesssim \delta^4 (1 + \|x\|_s^2)\end{aligned}$$

where the last inequality follows from Lemma 12 showing the first inequality. The second inequality follows similarly:

$$\|R^N(x,\delta)\|_s^2 = \delta^2 \|\mathcal{C}^N(x^* - x)\|_s^2 \lesssim \delta^2 \|x^* - x\|_s^2 \lesssim \delta^4 (1 + \|x\|_s^2)$$

and the proof is done. ■

Next lemma shows that the size of the jump $y - x$ is of order $\sqrt{\Delta t}$.

Lemma 14 *Consider y given by* (49)*. Under Assumptions 4, for any $p \geq 1$ we have*

$$\mathbb{E}^{\pi^N}_x \Big[\|y - x\|_s^p \Big] \lesssim (\Delta t)^{\frac{p}{2}} \cdot (1 + \|x\|_s^p).$$

Proof Under Assumption 4 the function μ^N is globally Lipschitz on $\mathcal{H}^s$, with Lipschitz constant that can be chosen independent from N. Thus using Lemma 13 we obtain that

$$\begin{aligned}\|y - x\|_s &\lesssim \Delta t (1 + \|x\|_s) + \|R^N(x,\delta)\|_s + \sqrt{\Delta t}\, \|\mathcal{C}^{\frac{1}{2}} \xi^N\|_s \\ &\lesssim \Delta t (1 + \|x\|_s) + (\Delta t)^2 (1 + \|x\|_s) + \sqrt{\Delta t}\, \|\mathcal{C}^{\frac{1}{2}} \xi^N\|_s \\ &\lesssim \Delta t (1 + \|x\|_s) + \sqrt{\Delta t}\, \|\mathcal{C}^{\frac{1}{2}} \xi^N\|_s.\end{aligned}$$

We have $\mathbb{E}^{\pi^0}\Big[\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s^p\Big] \leq \mathbb{E}^{\pi^0}\Big[\|\zeta\|_s^p\Big] < \infty$, where $\zeta \overset{\mathcal{D}}{\sim} \mathrm{N}(0,\mathcal{C})$. Consequently, $\mathbb{E}^{\pi^0}\Big[\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s^p\Big]$ is uniformly bounded as a function of N, proving the lemma. ■

Consider y given by (49) and recall from (47) that

$$y = x_{\mathrm{MALA}} + R^N(x,\delta).$$

Lemma 15 *We have*

$$\begin{aligned} a^N(x,\delta) &\equiv \|y\|^2_{\mathcal{C}^N} - \|x_{\mathrm{MALA}}\|^2_{\mathcal{C}^N} \\ \mathbb{E}^{\pi^N} a^N(x,\delta) &\lesssim \delta^2\end{aligned}$$

Proof From (47) we have

$$\begin{aligned}\|y\|^2_{\mathcal{C}^N} - \|x_{\mathrm{MALA}}\|^2_{\mathcal{C}^N} &= a^N(x,\delta) \\ a^N(x,\delta) &\equiv 2\langle x_{\mathrm{MALA}}, R^N(x)\rangle_{\mathcal{C}^N} + \|R^N(x,\delta)\|^2_{\mathcal{C}^N}.\end{aligned} \tag{64}$$

From (48), we obtain

$$\begin{aligned}|\langle x_{\mathrm{MALA}}, R^N(x,\delta)\rangle_{\mathcal{C}^N}| &= |\langle x_{\mathrm{MALA}}, (\mathcal{C}^N)^{-1} R^N(x,\delta)\rangle| \\ &\leq \|x_{\mathrm{MALA}}\|_s \|R^N(x,\delta)\|_{\mathcal{C}^N}.\end{aligned}$$

From Lemma 14 we deduce that

$$\|x_{\text{MALA}}\|_s \lesssim (1+\delta)(1+\|x\|_s) + \sqrt{\delta}\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s\,.$$

Combining this with Lemma 13 yields that

$$|\langle x_{\text{MALA}}, R^N(x,\delta)\rangle_{\mathcal{C}^N}| \lesssim \delta^2(1+\|x\|_s^2)(1+\sqrt{\delta}\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s)\,.$$

Thus

$$\mathbb{E}^{\pi^N}(|\langle x_{\text{MALA}}, R^N(x,\delta)\rangle_{\mathcal{C}^N}|) \lesssim \delta^2\,\mathbb{E}^{\pi^N}(1+\|x\|_s^2)(1+\sqrt{\delta}\|\mathcal{C}^{\frac{1}{2}}\xi^N\|_s) \lesssim \delta^2. \tag{65}$$

Thus from (64), (65) and Lemma 13 we deduce that

$$\mathbb{E}^{\pi^N}(a^N(x,\delta) \lesssim \delta^2$$

and the proof is finished. ■

7.3 Gaussian approximation of Q^N

Recall the quantity Q^N defined in Equation (55). This section proves that Q^N has a Gaussian behavior in the sense that

$$Q^N(x,\xi^N) = Z^N(x,\xi^N) \;+\; i^N(x,\xi^N) \;+\; \mathbf{e}^N(x,\xi^N) \tag{66}$$

where the quantities Z^N and i^N are equal to

$$Z^N(x,\xi^N) = -\frac{\ell^3}{4} - \frac{\ell^{\frac{3}{2}}}{\sqrt{2}}N^{-\frac{1}{2}}\sum_{j=1}^{N}\lambda_j^{-1}\xi_j x_j \tag{67}$$

$$i^N(x,\xi^N) = \frac{1}{2}(\ell\Delta t)^2\Big(\|x\|_{\mathcal{C}^N}^2 - \|(\mathcal{C}^N)^{\frac{1}{2}}\xi^N\|_{\mathcal{C}^N}^2\Big) \tag{68}$$

with i^N and e^N small. Thus the principal contributions to Q^N comes from the random variable $Z^N(x,\xi^N)$. Notice that, for each fixed $x \in \mathcal{H}^s$, the random variable $Z^N(x,\xi^N)$ is Gaussian. Furthermore, the Karhunen-Loève expansion of π_0 shows that for π_0-almost every choice of function $x \in \mathcal{H}$ the sequence $\big\{Z^N(x,\xi^N)\big\}_{N\geq 1}$ converges in law to the distribution of $Z_\ell \overset{\mathcal{D}}{\sim} \mathrm{N}(-\frac{\ell^3}{4}, \frac{\ell^3}{2})$. The next lemma rigorously bounds the error terms $\mathbf{e}^N(x,\xi^N)$ and $i^N(x,\xi^N)$: we show that i^N is an error term of order $\mathcal{O}(N^{-\frac{1}{6}})$ and $\mathbf{e}^N(x,\xi)$ is an error term of order $\mathcal{O}(N^{-\frac{1}{3}})$. In Lemma 17 we then quantify the convergence of $Z^N(x,\xi^N)$ to Z_ℓ.

Lemma 16 (Gaussian Approximation) *Let $p \geq 1$ be an integer. Under Assumptions 4, $Q^N(x,\xi^N)$ has the expansion given in* (66) *and the error terms i^N and $\mathbf{e}^N$ in the Gaussian approximation* (66) *satisfy*

$$\Big(\mathbb{E}^{\pi^N}\big[|i^N(x,\xi^N)|^p\big]\Big)^{\frac{1}{p}} = \mathcal{O}(N^{-\frac{1}{6}}) \qquad \textit{and} \qquad \Big(\mathbb{E}^{\pi^N}\big[|\mathbf{e}^N(x,\xi^N)|^p\big]\Big)^{\frac{1}{p}} = \mathcal{O}(N^{-\frac{1}{3}}). \tag{69}$$

Proof As in Lemma 4.4 of Pillai et al. (2012), without loss of generality, we suppose $p = 2q$. The quantity Q^N is defined in Equation (55) and expanding terms leads to

$$Q^N(x, \xi^N) \;=\; I_1 \;+\; I_2 \;+\; I_3 + \; I_4$$

where the quantities I_1, I_2, I_3 and I_4 are given by

$$\begin{aligned}
I_1 &= -\frac{1}{2}\big(\|y\|^2_{\mathcal{C}^N} - \|x\|^2_{\mathcal{C}^N}\big) - \frac{1}{4\ell\Delta t}\big(\|x - y(1-\ell\Delta t)\|^2_{\mathcal{C}^N} - \|y - x(1-\ell\Delta t)\|^2_{\mathcal{C}^N}\big)\\
I_2 &= -\Big(\Psi^N(y) - \Psi^N(x)\Big) - \frac{1}{2}\Big(\langle x - y(1-\ell\Delta t), \mathcal{C}^N\nabla\Psi^N(y)\rangle_{\mathcal{C}^N} - \langle y - x(1-\ell\Delta t), \mathcal{C}^N\nabla\Psi^N(x)\rangle_{\mathcal{C}^N}\Big)\\
I_3 &= -\frac{1}{4\ell\Delta t}\Big\{\|\ell\Delta t\,\mathcal{C}^N\nabla\Psi^N(y) + R^N(y,\delta)\|^2_{\mathcal{C}^N} - \|\ell\Delta t\,\mathcal{C}^N\nabla\Psi^N(x) + R^N(x,\delta)\|^2_{\mathcal{C}^N}\Big\}\\
I_4 &= -\frac{1}{2\ell\Delta t}\Big\{\langle x - y(1-\ell\Delta t), R^N(y,\delta)\rangle_{\mathcal{C}^N} - \langle y - x(1-\ell\Delta t), R^N(x,\delta)\rangle_{\mathcal{C}^N}\Big\}.
\end{aligned}$$

The term I_1 arises purely from the Gaussian part of the target measure π^N and from the Gaussian part of the proposal. The other terms come from the change of probability involving the functional Ψ^N. By the calculation identical to page 2343 of Pillai et al. (2012), we can simplify the the term I_1 to be:

$$I_1 = -\frac{\ell\Delta t}{4}\Big(\|y\|^2_{\mathcal{C}^N} - \|x\|^2_{\mathcal{C}^N}\Big). \tag{70}$$

The term I_1 is shown to be $\mathcal{O}(1)$ and constitutes the main contribution to Q^N. Before analyzing I_1 in more detail, we show that I_2, I_3 and I_4 are $\mathcal{O}(N^{-\frac{1}{3}})$:

$$\Big(\mathbb{E}^{\pi^N}[I_2^{2q}]\Big)^{\frac{1}{2q}} + \Big(\mathbb{E}^{\pi^N}[I_3^{2q}]\Big)^{\frac{1}{2q}} + \Big(\mathbb{E}^{\pi^N}[I_4^{2q}]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{3}}). \tag{71}$$

- By a calculation nearly identical to the one in Lemma 4.4 of Pillai et al. (2012) (the only change being the use of our Lemma 14 instead of their Lemma 4.2) we obtain that

$$\Big(\mathbb{E}^{\pi^N}[I_2^{2q}]\Big)^{\frac{1}{2q}} \;=\; \mathcal{O}(N^{-\frac{1}{3}}). \tag{72}$$

- Using the definition of $R^N(x,\delta)$ from (48), we obtain that

$$\begin{aligned}
\mathbb{E}^{\pi^N}\big[I_3^{2q}\big] \lesssim \Delta t^{2q}\,\mathbb{E}^{\pi^N}\Big[|\langle\nabla\Psi^N(x), \mathcal{C}^N\nabla\Psi^N(x)\rangle|^q + |\langle\nabla\Psi^N(y), \mathcal{C}^N\nabla\Psi^N(y)\rangle|^q\Big]\\
+ \Delta t^{-2q}\,\mathbb{E}^{\pi^N}\Big[\|R^N(x,\delta)\|^{2q}_{\mathcal{C}^N} + \|R^N(y,\delta)\|^{2q}_{\mathcal{C}^N}\Big].
\end{aligned}$$

Lemma 8 states $\mathcal{C}^N\nabla\Psi^N : \mathcal{H}^s \to \mathcal{H}^s$ is globally Lipschitz, with a Lipschitz constant that can be chosen uniformly in N. Therefore,

$$\|\mathcal{C}^N\nabla\Psi^N(z)\|_s \lesssim 1 + \|z\|_s. \tag{73}$$

Since $\|\mathcal{C}^N\nabla\Psi^N(z)\|^2_{\mathcal{C}^N} = \langle\nabla\Psi^N(z), \mathcal{C}^N\nabla\Psi^N(z)\rangle$, the bound (32) gives

$$\mathbb{E}^{\pi^N}\big[I_3^{2q}\big] \lesssim \Delta t^{2q}\,\mathbb{E}\Big[\langle\nabla\Psi^N(x), \mathcal{C}^N\nabla\Psi^N(x)\rangle^q + \langle\nabla\Psi^N(y), \mathcal{C}^N\nabla\Psi^N(y)\rangle^q\Big]$$

$$\lesssim \Delta t^{2q}\, \mathbb{E}^{\pi^N}\Big[(1+\|x\|_s)^{2q} + (1+\|y\|_s)^{2q}\Big]$$

$$\lesssim \Delta t^{2q}\, \mathbb{E}^{\pi^N}\Big[1+\|x\|_s^{2q} + \|y\|_s^{2q}\Big] \;\lesssim\; \Delta t^{2q} \;=\; \Big(N^{-\frac{1}{3}}\Big)^{2q}. \tag{74}$$

Similarly, from Lemma 13 and 14,

$$\Delta t^{-2q}\, \mathbb{E}^{\pi^N}\Big[\|R^N(x,\delta)\|_{\mathcal{C}^N}^{2q} + \|R^N(y,\delta)\|_{\mathcal{C}^N}^{2q}\Big] \lesssim \Delta t^{6q}\, \mathbb{E}^{\pi^N}\Big[1+\|x\|_s^{2q} + \|y\|_s^{2q}\Big] \;\lesssim\; \Delta t^{6q}$$

$$\lesssim \Big(N^{-\frac{1}{3}}\Big)^{6q} \lesssim \Big(N^{-\frac{1}{3}}\Big)^{2q}. \tag{75}$$

Thus from (74) and (75), we conclude that

$$\Big(\mathbb{E}^{\pi^N}[I_3^{2q}]\Big)^{\frac{1}{2q}} \lesssim \Big(N^{-\frac{1}{3}}\Big)^{2q}. \tag{76}$$

- Finally, we tackle the term I_4:

$$\mathbb{E}^{\pi^N}\big[I_4^{2q}\big] \lesssim \Delta t^{-2q}\, \mathbb{E}\Big[\|x - y(1-\ell\Delta t)\|_s^{2q}\, \|(\mathcal{C}^N)^{-1} R^N(y,\delta)\|_s^{2q}$$
$$+ \|y - x(1-\ell\Delta t)\|_s^{2q}\, \|(\mathcal{C}^N)^{-1} R^N(x,\delta)\|_s^{2q}\Big].$$

From Lemma 14, we obtain that $\mathbb{E}^{\pi^N}(\|y - x(1-\ell\Delta t)\|_s^{4q}) \lesssim (\Delta t)^{2q}\cdot(1+\|x\|_s^{4q})$ and $\mathbb{E}^{\pi^N}(\|x - y(1-\ell\Delta t)\|_s^{2q}) \lesssim (\Delta t)^{2q}\cdot(1+\|x\|_s^{4q})$. Similarly, from Lemma 13 we gather that $\mathbb{E}^{\pi^N}\|R^N(x,\delta)\|_{\mathcal{C}^N}^{4q} \lesssim \delta^{8q}(1+\|x\|_s^{4q})$. Putting these two together and using the Cauchy-Schwartz inequality gives,

$$\mathbb{E}^{\pi^N}\big[I_4^{2q}\big] \lesssim \Delta t^{3q}\, \mathbb{E}^{\pi^N}\Big[1+\|x\|_s^{2q} + \|y\|_s^{2q}\Big] \lesssim \Big(N^{-\frac{1}{3}}\Big)^{2q}. \tag{77}$$

Equations (72), (76) and (77) imply the requisite estimate in (71).
Next, we tackle the term I_1. Recall from from (70) that

$$I_1 = -\frac{\ell\Delta t}{4}\Big(\|y\|_{\mathcal{C}^N}^2 - \|x\|_{\mathcal{C}^N}^2\Big).$$

From Lemma 15 we obtain that

$$\|y\|_{\mathcal{C}^N}^2 = \|x_{\mathrm{MALA}}\|_{\mathcal{C}^N}^2 + a^N(x,\ell\Delta t), \qquad \mathbb{E}^{\pi^N} a^N(x,\ell\Delta t) \lesssim (\Delta t)^2. \tag{78}$$

Consequently,

$$I_1 = -\frac{\ell\Delta t}{4}\Big(\|x_{\mathrm{MALA}}\|_{\mathcal{C}^N}^2 - \|x\|_{\mathcal{C}^N}^2\Big) - \frac{\ell\Delta t}{4} a^N(x,\ell\Delta t).$$

From Lemma 4.4 of Pillai et al. (2012), we deduce that

$$-\frac{\ell\Delta t}{4}\Big(\|x_{\mathrm{MALA}}\|_{\mathcal{C}^N}^2 - \|x\|_{\mathcal{C}^N}^2\Big) = Z^N(x,\xi^N) \;+\; i^N(x,\xi^N) \;+\; b^N(x,\xi^N) \tag{79}$$

with $Z^N(x,\xi^N)$ and $i^N(x,\xi^N)$ given by Equation (67) and (68) and

$$\Big(\mathbb{E}^{\pi^N}\big[b^N(x,\xi^N)^{2q}\big]\Big)^{\frac{1}{2q}} \;=\; \mathcal{O}(N^{-\frac{1}{3}}).$$

Lemma 4.4 of Pillai et al. (2012) also shows that

$$\Big(\mathbb{E}^{\pi^N}\big[i^N(x,\xi^N)^{2q}\big]\Big)^{\frac{1}{2q}} = \mathcal{O}(N^{-\frac{1}{6}}). \tag{80}$$

The proof of the lemma now follows from (71), (78) and (79). ■

We recall Lemma 4.5 of Pillai et al. (2012):

Lemma 17 *Pillai et al. (2012)(Lemma 4.5)* **(Asymptotic independence)** *Let $p \geq 1$ be a positive integer and $f : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function. Consider error terms $\boldsymbol{e}^N_\star(x,\xi)$ satisfying*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}[\boldsymbol{e}^N_\star(x,\xi^N)^p] = 0.$$

Define the functions $\bar{f}^N : \mathbb{R} \to \mathbb{R}$ and the constant $\bar{f} \in \mathbb{R}$ by

$$\bar{f}^N(x) = \mathbb{E}_x\Big[f\big(Z^N(x,\xi^N) + \boldsymbol{e}^N_\star(x,\xi^N)\big)\Big] \qquad \textit{and} \qquad \bar{f} = \mathbb{E}[f(Z_\ell)].$$

Then the function f^N is highly concentrated around its mean in the sense that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[|\bar{f}^N(x) - \bar{f}|^p\Big] = 0.$$

Corollary 18 *Let $p \geq 1$ be a positive. The local mean acceptance probability $\alpha^N(x)$ defined in Equation* (51) *satisfies*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\big[|\alpha^N(x) - \alpha(\ell)|^p\big] = 0.$$

Proof The function $f(z) = 1 \wedge e^z$ is 1-Lipschitz and $\alpha(\ell) = \mathbb{E}[f(Z_\ell)]$. Also,

$$\alpha^N(x) = \mathbb{E}_x\Big[f(Q^N(x,\xi^N))\Big] = \mathbb{E}_x\big[f(Z^N(x,\xi^N) + \mathbf{e}^N_\star(x,\xi^N)\big]$$

with $\mathbf{e}^N_\star(x,\xi^N) = i^N(x,\xi^N) + \mathbf{e}^N(x,\xi^N)$. Lemma 16 shows that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}[\mathbf{e}^N_\star(x,\xi)^p] = 0$$

and therefore Lemma 17 gives the conclusion. ■

7.4 Drift approximation

This section proves that the approximate drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ defined in Equation (58) converges to the drift function $\mu : \mathcal{H}^s \to \mathcal{H}^s$ of the limiting diffusion (54).

Lemma 19 (Drift Approximation) *Let Assumptions 4 hold. The drift function $d^N : \mathcal{H}^s \to \mathcal{H}^s$ converges to μ in the sense that*

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[\|d^N(x) - \mu(x)\|_s^2\Big] = 0.$$

Proof Now that we have established the relevant estimates, the proof of this lemma is nearly identical to that of Lemma 4.7 of Pillai et al. (2012), but also needs to account for the extra error term induced by the proximal operator. The approximate drift d^N is given by Equation (58). The definition of the local mean acceptance probability $\alpha^N(x)$ given by Equation (51) shows that d^N can also be expressed as

$$d^N(x) = \Big(\alpha^N(x)\alpha(\ell)^{-1}\Big)\mu^N(x) + \mathcal{R}^N_{\text{Prox}}(x, \Delta t) + \sqrt{2\ell}h(\ell)^{-1}(\Delta t)^{-\frac{1}{2}}\varepsilon^N(x) \qquad (81)$$

where $\mu^N(x) = -\Big(P^N x + \mathcal{C}^N \nabla \Psi^N(x)\Big)$; the term $\varepsilon^N(x)$ is defined by

$$\varepsilon^N(x) \;=\; \mathbb{E}_x\big[\gamma^N(x,\xi^N)\,\mathcal{C}^{\frac{1}{2}}\xi^N\big] \;=\; \mathbb{E}_x\big[\big(1 \wedge e^{Q^N(x,\xi^N)}\big)\,\mathcal{C}^{\frac{1}{2}}\xi^N\big]$$

and the term $\mathcal{R}^N_{\text{Prox}}(x, \Delta t)$ is the error term induced by the proximal approximation:

$$\mathcal{R}^N_{\text{Prox}}(x, \Delta t) = \frac{\alpha^N(x)}{h(\ell)}\frac{R^n(x, \ell\Delta)}{\Delta t}.$$

To prove Lemma 19 it suffices to verify that

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\Big[\big\|\big(\alpha^N(x)\alpha(\ell)^{-1}\big)\mu^N(x) - \mu(x)\big\|_s^2\Big] = 0 \qquad (82)$$

$$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\|\mathcal{R}^N_{\text{Prox}}(x, \Delta t)\|_s^2 = 0 \qquad (83)$$

$$\lim_{N\to\infty} (\Delta t)^{-1}\,\mathbb{E}^{\pi^N}\Big[\|\varepsilon^N(x)\|_s^2\Big] = 0. \qquad (84)$$

- Equation (82) follows directly from Lemma 4.7 of Pillai et al. (2012).
- Next, using the fact that $|\alpha^N(x)| \le 1$ and Lemma 13,

  $$\begin{aligned}\|\mathcal{R}^N_{\text{Prox}}(x, \Delta t)\|_s^2 &\lesssim \Big(\alpha^N(x)\Big)^2\Big\|\frac{R^n(x, \ell\Delta t)}{\Delta t}\Big\|_s^2 \\ &\lesssim (\Delta t)^2(1 + \|x\|_s^2)\end{aligned}$$

  and thus we have

  $$\lim_{N\to\infty} \mathbb{E}^{\pi^N}\|\mathcal{R}^N_{\text{Prox}}(x, \Delta t)\|_s^2 = \lim_{N\to\infty} N^{-2/3}\mathbb{E}^{\pi^N}(1 + \|x\|_s^2) = 0$$

  establishing (83).

- Let us prove Equation (84). If the Bernoulli random variable $\gamma^N(x,\xi^N)$ were independent from the noise term $(\mathcal{C}^N)^{\frac{1}{2}}\xi^N$, it would follow that $\varepsilon^N(x)=0$. In general $\gamma^N(x,\xi^N)$ is not independent from $(\mathcal{C}^N)^{\frac{1}{2}}\xi^N$ so that $\varepsilon^N(x)$ is not equal to zero. Nevertheless, as quantified by Lemma 17, the Bernoulli random variable $\gamma^N(x,\xi^N)$ is asymptotically independent from the current position x and from the noise term $(\mathcal{C}^N)^{\frac{1}{2}}\xi^N$. Consequently, we can prove in Equation (86) that the quantity $\varepsilon^N(x)$ is small. To this end, we establish that each component $\langle\varepsilon(x),\hat{\varphi}_j\rangle_s^2$ satisfies

$$\mathbb{E}^{\pi^N}\big[\langle\varepsilon^N(x),\hat{\varphi}_j\rangle_s^2\big] \quad\lesssim\quad N^{-1}\mathbb{E}^{\pi^N}[\langle x,\hat{\varphi}_j\rangle_s^2]+N^{-\frac{2}{3}}(j^s\lambda_j)^2. \tag{85}$$

  Summation of Equation (85) over $j=1,\ldots,N$ leads to

$$\mathbb{E}^{\pi^N}\Big[\|\varepsilon^N(x)\|_s^2\Big] \quad\lesssim\quad N^{-1}\mathbb{E}^{\pi^N}\big[\|x\|_s^2\big]+N^{-\frac{2}{3}}\operatorname{Tr}_{\mathcal{H}^s}(\mathcal{C}_s) \quad\lesssim\quad N^{-\frac{2}{3}}, \tag{86}$$

  which gives the proof of Equation (84). To prove Equation (85) for a fixed index $j\in\mathbb{N}$, the quantity $Q^N(x,\xi)$ is decomposed as a sum of a term independent from ξ_j and another remaining term of small magnitude. To this end we introduce

$$\begin{cases} Q^N(x,\xi^N) & = Q_j^N(x,\xi^N)+Q_{j,\perp}^N(x,\xi^N) \\ Q_j^N(x,\xi^N) & = -\frac{1}{\sqrt{2}}\ell^{\frac{3}{2}}N^{-\frac{1}{2}}\lambda_j^{-1}x_j\xi_j-\frac{1}{2}\ell^2N^{-\frac{2}{3}}\lambda_j^2\xi_j^2+\mathbf{e}^N(x,\xi^N). \end{cases} \tag{87}$$

  The definitions of $Z^N(x,\xi^N)$ and $i^N(x,\xi^N)$ in Equation (67) and (68) readily show that $Q_{j,\perp}^N(x,\xi^N)$ is independent from ξ_j. The noise term satisfies $\mathcal{C}^{\frac{1}{2}}\xi^N=\sum_{j=1}^N(j^s\lambda_j)\xi_j\hat{\varphi}_j$. Since $Q_{j,\perp}^N(x,\xi^N)$ and ξ_j are independent and $z\mapsto 1\wedge e^z$ is 1-Lipschitz, it follows that

$$\begin{aligned}\langle\varepsilon^N(x),\hat{\varphi}_j\rangle_s^2 &= (j^s\lambda_j)^2\,\Big(\mathbb{E}_x\big[\big(1\wedge e^{Q^N(x,\xi^N)}\big)\;\xi_j\big]\Big)^2 \\ &= (j^s\lambda_j)^2\,\Big(\mathbb{E}_x\big[[\big(1\wedge e^{Q^N(x,\xi^N)}\big)-\big(1\wedge e^{Q_{j,\perp}^N(x,\xi^N)}\big)]\;\xi_j\big]\Big)^2 \\ &\lesssim (j^s\lambda_j)^2\mathbb{E}_x\big[|Q^N(x,\xi^N))-Q_{j,\perp}^N(x,\xi^N)|^2\big] \\ &= (j^s\lambda_j)^2\mathbb{E}_x\big[Q_j^N(x,\xi^N)^2\big].\end{aligned}$$

  By Lemma 16 $\mathbb{E}^{\pi^N}\big[\mathbf{e}^N(x,\xi^N)^2\big]\lesssim N^{-\frac{2}{3}}$. Therefore,

$$\begin{aligned}(j^s\lambda_j)^2\mathbb{E}^{\pi^N}\big[Q_j^N(x,\xi^N)^2\big] &\lesssim (j^s\lambda_j)^2\Big\{N^{-1}\lambda_j^{-2}\mathbb{E}^{\pi^N}\big[x_j^2\xi_j^2\big]+N^{-\frac{4}{3}}\mathbb{E}^{\pi^N}\big[\lambda_j^4\xi_j^4\big]+\mathbb{E}^{\pi^N}\big[\mathbf{e}^N(x,\xi)^2\big]\Big\} \\ &\lesssim N^{-1}\,\mathbb{E}^{\pi^N}\big[(j^sx_j)^2\xi_j^2\big]+(j^s\lambda_j)^2(N^{-\frac{4}{3}}+N^{-\frac{2}{3}}) \\ &\lesssim N^{-1}\,\mathbb{E}^{\pi^N}\big[\langle x,\hat{\varphi}_j\rangle_s^2\big]+(j^s\lambda_j)^2N^{-\frac{2}{3}} \\ &\lesssim N^{-1}\,\mathbb{E}^{\pi^N}\big[\langle x,\hat{\varphi}_j\rangle_s^2\big]+(j^s\lambda_j)^2N^{-\frac{2}{3}},\end{aligned}$$

  which finishes the proof of Equation (85).

Thus we have established (82), (83) and (84) and the proof is finished. ■

7.5 Noise approximation

Recall the definition (59) of the martingale difference $\Gamma^{k,N}$. In this section we estimate the error in the approximation $\Gamma^{k,N} \approx \mathrm{N}(0, \mathcal{C}_s)$. To this end we introduce the covariance operator

$$D^N(x) = \mathbb{E}_x\Big[\Gamma^{k,N} \otimes_{\mathcal{H}^s} \Gamma^{k,N} \,|x^{k,N} = x\Big].$$

For any $x, u, v \in \mathcal{H}^s$ the operator $D^N(x)$ satisfies

$$\mathbb{E}\Big[\langle \Gamma^{k,N}, u\rangle_s \langle \Gamma^{k,N}, v\rangle_s \,|x^{k,N} = x\Big] \;=\; \langle u, D^N(x) v\rangle_s.$$

The next lemma gives a quantitative version of the approximation $D^N(x) \approx \mathcal{C}_s$.

Lemma 20 *Let Assumptions 4 hold. For any pair of indices $i, j \geq 0$ the operator $D^N(x) : \mathcal{H}^s \to \mathcal{H}^s$ satisfies*

$$\lim_{N\to\infty} \quad \mathbb{E}^{\pi^N} \big|\langle \hat{\varphi}_i, D^N(x)\hat{\varphi}_j\rangle_s - \langle \hat{\varphi}_i, \mathcal{C}_s \hat{\varphi}_j\rangle_s\big| \;=\; 0 \tag{88}$$

and, furthermore,

$$\lim_{N\to\infty} \quad \mathbb{E}^{\pi^N} \big|\mathrm{Tr}_{\mathcal{H}^s}(D^N(x)) - \mathrm{Tr}_{\mathcal{H}^s}(\mathcal{C}_s)\big| \;=\; 0. \tag{89}$$

Proof This lemma follows directly from Lemma 4.8 of Pillai et al. (2012), since the only estimate needed for the proof of Lemma 4.8 of Pillai et al. (2012) is the Gaussian approximation and the estimate for $\mathbf{e}^N(x, \xi^N)$ established in Lemma 16. Thus the proof is finished. ■

7.6 Martingale Invariance Principle

This section proves that the process W^N defined in Equation (62) converges to a Brownian motion.

Proposition 21 *Let Assumptions 4 hold. Let $z^0 \sim \pi$ and $W^N(t)$ the process defined in equation (62) and $x^{0,N} \overset{\mathcal{D}}{\sim} \pi^N$ the starting position of the Markov chain x^N. Then*

$$(x^{0,N}, W^N) \Longrightarrow (z^0, W), \tag{90}$$

where $\Longrightarrow$ denotes weak convergence in $\mathcal{H}^s \times C([0,T]; \mathcal{H}^s)$, and W is a $\mathcal{H}^s$-valued Brownian motion with covariance operator $\mathcal{C}_s$. Furthermore the limiting Brownian motion W is independent of the initial condition z^0.

Proof

This proof involves verifying three conditions of Proposition 5.1 of Berger (1986) and is identical to that of Proposition 4.10 of Pillai et al. (2012). The only change required is to use our Lemma 20 instead of their Lemma 4.8. Therefore we omit the details of the rest of the proof. ■

8. Closing Comments

There are a number of related issues that are of great practical interest:

- In Theorem 1 of Crucinio et al. (2023), the authors extended Theorem 2 to a general class of product measures.
- As mentioned in the introduction, we choose $\lambda = \delta$. In Crucinio et al. (2023), it is shown that for differentiable targets, this choice is optimal. When $\delta \to 0$ quicker than λ, proximal MALA is less efficient than MALA.
- Of course, the most interesting case is when the log-target is not differentiable. In this scenario, Crucinio et al. (2023) show that the applicability of proximal MALA comes at a cost – the algorithm scales smaller than $O(N^{-\frac{1}{3}})$ and is less efficient than its smooth counterpart.
- A similar result should be of interest when proximal functions are used for implementing the Hybrid Monte Carlo algorithm; see Chaari et al. (2016).

Acknowledgements

The author thanks Gareth Roberts and Andrew Stuart for introducing him to optimal scaling of Markov Chain Monte Carlo Methods and several inspiring conversations, Dean Foster and Robert Stine for introducing him to proximal algorithms and related ADMM methods, Alex Belloni, David Dunson, Lucas Janson, Eric Laber, Jonathan Mattingly, Sayan Mukherjee, Debdeep Pati, Christian Robert, Aaron Smith, Pragya Sur and Alex Thiery for very helpful comments and their collaboration over the years. Part of this work was done when the author was an Amazon Scholar. He thanks Robert Stine and Andrew Ruud for their encouragement to work on this problem.

References

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

E. Berger. Asymptotic Behaviour of a Class of Stochastic Approximation Procedures. *Probability Theory and Related Fields*, 71(4):517–552, 1986. ISSN 0178-8051.

A. Beskos and A. Stuart. Mcmc Methods for Sampling Function Space. In *Invited Lectures, Sixth International Congress on Industrial and Applied Mathematics, ICIAM07, Editors Rolf Jeltsch and Gerhard Wanner*, pages 337–364. European Mathematical Society, 2009.

A. Beskos, G. Roberts, A. Stuart, and J. Voss. An MCMC Method for Diffusion Bridges. *Stochastics and Dynamics*, 8(3):319–350, 2008.

A. Beskos, G. Roberts, and A. Stuart. Optimal Scalings of Metropolis-Hastings Algorithms for Non-Product Targets in High Dimensions. *Annals of Applied Probability*, 19:863–898, 2009.

L. Chaari, J.-Y. Tourneret, C. Chaux, and H. Batatia. A Hamiltonian Monte Carlo Method for Non-Smooth Energy Sampling. *IEEE Transactions on Signal Processing*, 64(21):5585–5594, 2016.

F. R. Crucinio, A. Durmus, P. Jiménez, and G. O. Roberts. Optimal Scaling Results for a Wide Class of Proximal MALA Algorithms, 2023. arXiv preprint arXiv:2301.02446.

A. Durmus, E. Moulines, and M. Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

C. M. Elliott and A. Stuart. The Global Dynamics of Discrete Semilinear Parabolic Equations. *SIAM Journal on Numerical Analysis*, 30(6):1622–1663, 1993.

G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

M. Hairer, A. M. Stuart, J. Voss, and P. Wiberg. Analysis of SPDEs Arising in Path Sampling. Part I: The Gaussian Case. *Communications in Mathematical Sciences*, 3: 587–603, 2005.

M. Hairer, A. M. Stuart, and J. Voss. Analysis of SPDEs Arising in Path Sampling. Part II: The Nonlinear Case. *Annals of Applied Probability*, 17(5-6):1657–1706, 2007. ISSN 1050-5164.

M. Hairer, A. M. Stuart, and J. Voss. Signal Processing Problems on Function Space: Bayesian Formulation, Stochastic PDEs and Effective MCMC Methods. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

Z. Jiang, J. Jiang, Q. Yao, and G. Yang. A Neural Network-Based PDE Solving Algorithm with High Precision. *Scientific Reports*, 13(1):4479, 2023.

N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar. Neural Operator: Learning Maps Between Function Spaces with Applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

A. Lamperski. Projected Stochastic Gradient Langevin Algorithms for Constrained Sampling and Non-Convex Learning. In *Conference on Learning Theory*, pages 2891–2937. PMLR, 2021.

J. C. Mattingly, N. S. Pillai, and A. M. Stuart. Diffusion Limits of the Random Walk Metropolis Algorithm in High Dimensions. *The Annals of Applied Probability*, 22(3): 881–930, 2012.

M. Pereyra. Proximal Markov Chain Monte Carlo Algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

N. S. Pillai, A. M. Stuart, and A. H. Thiéry. Optimal Scaling and Diffusion Limits for the Langevin Algorithm in High Dimensions. *The Annals of Applied Probability*, 22(6): 2320–2356, 2012.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004. ISBN 0-387-21239-6.

G. O. Roberts and J. S. Rosenthal. Optimal Scaling of Discrete Approximations to Langevin Diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998. ISSN 1369-7412.

G. O. Roberts and J. S. Rosenthal. Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, 16(4):351–367, 2001. ISSN 0883-4237.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *Annals of Applied Probability*, 7(1):110–120, 1997.

A. Roy, L. Shen, K. Balasubramanian, and S. Ghadimi. Stochastic Zeroth-Order Discretizations of Langevin Diffusions for Bayesian Inference. *Bernoulli*, 28(3):1810–1834, 2022.

A. Stuart. Inverse Problems: a Bayesian Perspective. *Acta Numerica*, 19, 2010.

M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688. Citeseer, 2011.