

SAMPLING FROM ROUGH ENERGY LANDSCAPES*

PETR PLECHÁČ† AND GIDEON SIMPSON‡

Abstract. We examine challenges to sampling from Boltzmann distributions associated with multiscale energy landscapes. The multiscale features, or “roughness”, corresponds to highly oscillatory, but bounded, perturbations of a smooth landscape. Through a combination of numerical experiments and analysis we demonstrate that the performance of Metropolis adjusted Langevin algorithm can be severely attenuated as the roughness increases. In contrast, we prove that random walk Metropolis is insensitive to such roughness. We also formulate two alternative sampling strategies that incorporate large scale features of the energy landscape, while resisting the impact of fine scale roughness; these also outperform random walk Metropolis. Numerical experiments on these landscapes are presented that confirm our predictions. Open questions and numerical challenges are also highlighted.

Keywords. Markov chain Monte Carlo; random walk Metropolis; Metropolis adjusted Langevin; rough energy landscapes; multi-scale energy landscapes; mean squared displacement.

AMS subject classifications. 65C05; 65C40; 60J22.

1. Introduction

In this work, we consider the task of sampling from a Boltzmann distribution,

$$\mu(dx) = Z^{-1}e^{-\beta V(x)}dx, \quad Z = \int e^{-\beta V(x)}dx, \quad V: \mathbb{R}^n \rightarrow \mathbb{R}, \quad (1.1)$$

when V is, in some sense “rough”, or “rugged”. We are particularly interested in multiscale landscapes of the form

$$V_\epsilon(x) = V_0(x) + V_1(x, x/\epsilon), \quad \epsilon > 0. \quad (1.2)$$

Here, V_0 is a smooth, long range, trapping potential ($V_0(x) \rightarrow \infty$ as $|x| \rightarrow \infty$) that is bounded from below, and V_1 is bounded with local short wavelength features. $V_1(x, y)$ will also be assumed to be smooth. An example of such a rough landscape, and its impact on the associated distribution, is shown in Figure 1.1. Model potentials like (1.2) serve as prototypes for rough and multiscale landscapes found in disordered media and soft matter, [1–5]. The goal of the present work is to assess how such roughness impacts the performance of well known Markov chain Monte Carlo (MCMC) sampling strategies like random walk Metropolis (RWM) and Metropolis adjusted Langevin (MALA).

We recall that RWM and MALA generate samples for $e^{-\beta V(x)}$ with proposals¹

$$\text{RWM: } X_{k+1}^p = X_k + \sqrt{\beta^{-1}}\sigma\xi_{k+1}, \quad \xi_{k+1} \sim N(0, I), \quad (1.3)$$

$$\text{MALA: } X_{k+1}^p = X_k - \frac{\sigma^2}{2}\nabla V(X_k) + \sqrt{\beta^{-1}}\sigma\xi_{k+1}, \quad \xi_{k+1} \sim N(0, I). \quad (1.4)$$

These proposals are then accepted or rejected with the appropriate rule to ensure detailed balance with respect to $\mu(dx) \propto e^{-\beta V(x)}dx$.

As an example, sample MALA paths with $\sigma = 1$ for the landscapes in Figure 1.1 are shown in Figure 1.2. A path for the smooth landscape exhibits better mixing than

*Received: June 14, 2019; Accepted (in revised form): July 07, 2020. Communicated by Grigoris A. Pavliotis.

†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA (plechac@udel.edu).

‡Department of Mathematics, Drexel University, Philadelphia, PA 19139, USA (grs53@drexel.edu).

¹Proposals will be denoted with a superscript p.

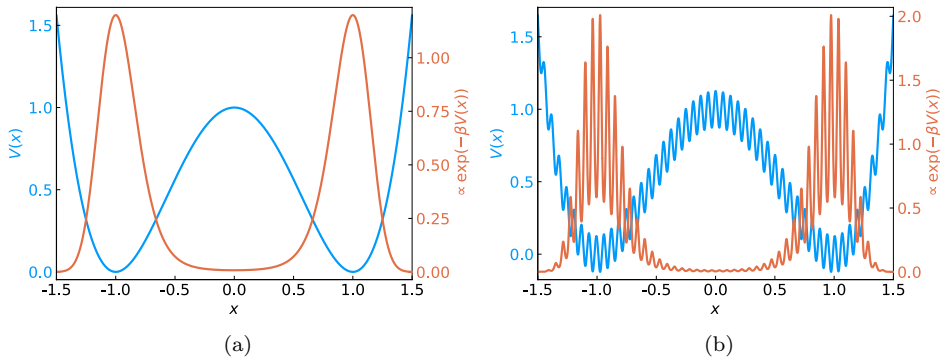


FIG. 1.1. In (a), we see a smooth multimodal energy landscape. In contrast, the landscape in (b), is rough, with many internal energy barriers. The underlying potentials are $V_0(x) = (x^2 - 1)^2$ and $V_\epsilon(x) = V_0(x) + \frac{1}{8} \cos(x/\epsilon)$ with $\epsilon = 0.01$ and $\beta = 5$ (color online).

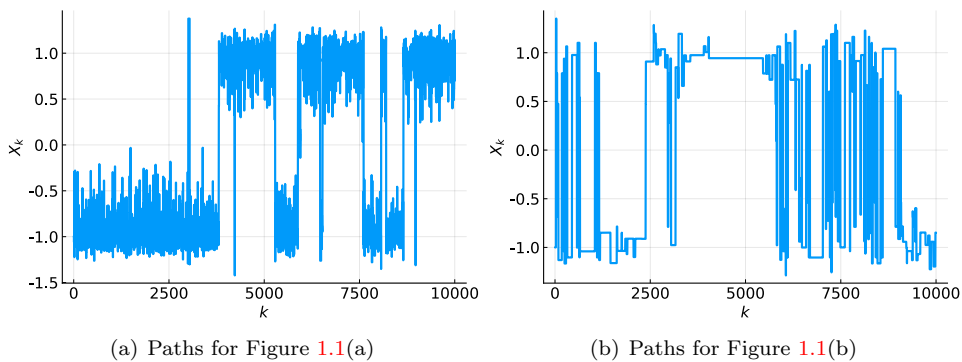


FIG. 1.2. Sample paths corresponding to the energy landscapes in Figure 1.1. The samples were generated using MALA with $\sigma = 1$ (color online).

the one for the rough landscape. On the rough landscape the trajectory stagnates.² This begs the question of whether or not $\sigma = 1$ was merely a poorly chosen value – perhaps with a different value the rough landscape would also be efficiently sampled. Large values of σ result in proposals with greater magnitude, but few will be accepted, and the trajectory will move slowly. Conversely, small values of σ produce more readily accepted proposals, but their size limits exploration of the state space. Consequently, an optimal choice of σ is anticipated for each distribution.

Assuming we tune our sampler to the optimal σ for each V_ϵ we seek to assess how ϵ impacts sampling performance. In this work, optimality, at a fixed value of ϵ and at a fixed dimension, will refer to maximization of some measurement of mixing, discussed below over the set of numerical parameters, i.e. σ in the case of MALA. Ultimately, our work indicates that even at the optimal value of proposal variance, MALA will cease to be effective as $\epsilon \rightarrow 0$. In contrast, even a poorly tuned RWM sampler remains robust in the $\epsilon \rightarrow 0$ limit.

In a “global” sense, robustness refers to the stability of the mixing and asymptotic variance properties of the chain as ϵ vanishes. This can be quantified through the

²Stagnation corresponds to persistent rejection of proposals.

spectral gap of the transition operator \mathcal{T} of a given MCMC method³

$$\text{Gap}(\mathcal{T}) \equiv 1 - \sup_{\lambda \in \sigma(\mathcal{T}) \setminus \{1\}} \lambda \equiv 1 - \Lambda. \quad (1.5)$$

The gap controls both the mixing of the process and the time-averaged variance constant, [6]. In particular, if the gap vanishes ($\Lambda \rightarrow 1$), then the mixing collapses and the variance bound explodes. The relationship between the gap and these quantities is given below.

Experimentally accessible measurements of mixing can be found by looking at observables. In particular, we consider the mean square displacement of the chain in stationarity. If this remains positive in the $\epsilon \rightarrow 0$ limit, it provides a “local” (in the sense of a single observable) notion of robustness. In addition to being straightforward to estimate through simulations, the mean squared displacement provides an upper bound on the spectral gap.

To obtain better performance than RWM, we also formulate two related sampling strategies that incorporate information about the large scale (i.e. long wavelength) features of the energy landscape through V_0 in (1.2). Indeed, our results, particularly Theorem 2.1 and Corollary 2.1 show that for potentials that can be decomposed as in (1.2) with V_0 smooth and trapping and V_1 rough but bounded, if the proposal of the sampling strategy is ϵ -independent, then the performance of the method will also be ϵ -independent.

1.1. Review of prior work

The question of optimizing σ to maximize performance was initially examined in [7, 8] and has been subsequently studied in other works, including [9–17]. Many of these works consider an energy landscape of the type

$$V_n(x) = \sum_{i=1}^n v(x_i), \quad x = (x_1, \dots, x_n) \in \mathbb{R}^n. \quad (1.6)$$

This choice of the potentials induces Boltzmann distributions that are products (for brevity, we take $\beta = 1$):

$$\mu(dx) \propto \prod_{i=1}^n e^{-v(x_i)} dx_i. \quad (1.7)$$

Thus, coordinates only interact through the accept/reject step of the method. Some results for potentials other than (1.6) have also been obtained. In [9], the authors treat distributions which have a density $e^{-\Phi(x)}$ with respect to a product measure (1.7). In earlier works, [7–9], it was often assumed that whichever sampler is studied, the process is in stationarity. This has been relaxed in the more recent results, [10–13, 16, 17].

Some of our results will also specialize to the rough analog of (1.6),

$$V_{\epsilon,n}(x) = \sum_{i=1}^n \underbrace{v_0(x_i) + v_1(x_i, x_i/\epsilon)}_{\equiv v_{\epsilon}(x_i)}. \quad (1.8)$$

As in the case of (1.2), we assume v_0 is trapping while v_1 is uniformly bounded. Both $v_0(x_i)$ and $v_1(x_i, y_i)$ are assumed to be smooth functions.

³Some authors refer to (1.5) as the interval, $\text{Int}(\mathcal{T})$. In [6], $\text{Gap}(\mathcal{T})$ is instead defined as $1 - \sup |\lambda|$.

In the case of (1.6), in stationarity, the performance can be measured by considering the mean square displacement (MSD)

$$\text{MSD}_n = \mathbb{E}_\mu[|X_{k+1} - X_k|^2]. \quad (1.9)$$

The strategy of [9] is to use the MSD as a proxy for mixing, and σ is selected to maximize it in the limit of $n \rightarrow \infty$. Indeed, this leads to the results that, as $n \rightarrow \infty$,

$$\text{MSD}_n/n = \ell^2 n^{-I} a(\ell; v) + o(n^{-I}), \quad I > 0, \quad (1.10)$$

where I and a both depend on the method, but *not* on the dimension n . The choice of σ is then related to ℓ by

$$\sigma^2 = \ell^2 n^{-I} \quad (1.11)$$

The function a is the mean acceptance rate in the $n \rightarrow \infty$ limit, [18].

Consequently, we can maximize this measure of performance as $n \rightarrow \infty$ by solving

$$\ell_\star = \operatorname{argmax}_\ell \ell^2 a(\ell; v). \quad (1.12)$$

The optimal σ_\star is inferred from (1.11), and there is an associated optimal acceptance rate, $a(\ell_\star)$. This optimization provides a strategy for tuning the value of σ to achieve the optimal acceptance rate approximately 23% for RWM. Analogously, one tunes MALA to have a 57% acceptance rate, [7, 8], and Hamiltonian Monte Carlo (HMC) to have a 65% acceptance rate, [15].

For RWM $I = 1$ while it is $I = 1/3$ for MALA. The function a has the explicit form

$$a(\ell; v) = 2\Phi\left(-\frac{\ell^{1/I}}{2}\sqrt{\mathcal{K}[v]}\right), \quad (1.13)$$

with \mathcal{K} a functional involving derivatives of v ; Φ is the standard normal ($N(0, 1)$) cumulative distribution function. A similar result holds for HMC, [15].

There are caveats to applying these results in practical computations, [19]. In particular, the results are obtained as $n \rightarrow \infty$ for distributions of form (1.7), and often assume the process to be in stationarity. Many distributions of interest will not be of this form, so target acceptance rates, like 23% for RWM, may be inappropriate. However, in the recent work [20], it was demonstrated that for a more general class of distributions than (1.7), 23% remains the optimal acceptance rate for RWM.

We mention these results because they motivate certain aspects of this work, such as the examination of product measures and the examination of MSD_n as a proxy for mixing. However, we emphasize that this work is focused on problems at fixed n , letting $\epsilon \rightarrow 0$. The preceding results can provide guidance in the $\epsilon \rightarrow 0$ limit, but they would necessitate first taking $n \rightarrow \infty$.

We also highlight the recent work in [10] which studies “ridged” densities associated with potentials of the form

$$V_\epsilon(x) = V_0(x_1) + V_1(x_1, x_2/\epsilon), \quad x = (x_1, x_2) \in \mathbb{R}^{n=n_1+n_2}. \quad (1.14)$$

Here, roughness is only present in a subset of the degrees of freedom (x_2). Examining RWM for such a problem, the authors are able to derive a limiting diffusion from which they can find an optimal step size. This limiting diffusion has a state dependent diffusion

coefficient. Both the drift and diffusion coefficients are nontrivial requiring averaging against the rough degrees of freedom.

Related to the work on ridged densities, and the present work, is [21]. In this work, the authors consider the case that one of the coordinates is scaled differently than the others. This would correspond to a potential like

$$V_\epsilon(x) = V(x_1/\epsilon, x_2, \dots, x_n), \quad x \in \mathbb{R}^n. \quad (1.15)$$

In [21] the authors also look for algorithms that are less sensitive to length scale variations in the gradients. They obtain results on MALA and HMC showing poor behavior in the $\epsilon \rightarrow 0$ limit. An important tool that they use in their analysis is the Dirichlet form and its relationship to the spectral gap; we, too, make use of that approach.

Another relevant work is [22]. There, the authors sought to perform gradient based sampling on non-differentiable energy landscapes and proposed using a Moreau-Yosida regularization. This approach is related to one of the mechanisms that we propose in order to overcome roughness, though our potentials are smooth, but highly oscillatory.

1.2. Measures of performance and notions of robustness. As mentioned, the key metrics that we use to assess performance are the spectral gap, (1.5), along with the MSD, (1.9). We recall the relationships amongst these quantities, [6, 21]. First, the relaxation to the equilibrium in the total variation (TV) and the time-averaged variance constant (TAVC) are controlled by Λ and $\text{Gap}(\mathcal{T})$:

$$\lim_{k \rightarrow \infty} \frac{1}{k} \log \|\rho_0 \mathcal{T}^k - \mu\|_{\text{TV}} \leq \log \Lambda = \log(1 - \text{Gap}(\mathcal{T})), \quad (1.16)$$

$$\lim_{k \rightarrow \infty} \frac{1}{k} \text{Var}_\mu \left(\sum_{j=0}^{k-1} f(X_j) \right) \leq \frac{1 + \Lambda}{1 - \Lambda} \text{Var}(f(X_k)) = \frac{2 - \text{Gap}(\mathcal{T})}{\text{Gap}(\mathcal{T})} \text{Var}(f(X_k)), \quad (1.17)$$

for an initial distribution with $d\rho_0/d\mu \in L^2(\mu)$ and any function $f \in L^2(\mu)$. Consequently, if $\text{Gap}(\mathcal{T}) \rightarrow 0$, mixing ceases and there is no upper bound on the TAVC. On the other hand, if the gap remains positive then there is a priori bound on the TAVC. Due to the inequality in (1.16) a positive spectral gap does *not* imply a lower bound on mixing.

A key relationship is between spectral gap and the Dirichlet form

$$\inf_{f \in L^2_{0,1}(\mu)} \frac{1}{2} \mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] = 1 - \sup_{\lambda \in \sigma(\mathcal{T}) \setminus \{1\}} \lambda = \text{Gap}(\mathcal{T}), \quad (1.18)$$

where $L^2_{0,1}(\mu)$ is the subset of mean zero, unit variance functions in $L^2(\mu)$. An elementary computation reveals that this is equivalent to

$$\text{Gap}(\mathcal{T}) = \inf_{f \in L^2(\mu)} \frac{\mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2]}{2 \text{Var}_\mu(f(X))}. \quad (1.19)$$

Consequently, for any square integrable function, if we can estimate the term on the right-hand side of (1.19), we have obtained an upper bound on the spectral gap. This allows us to use (1.9) as a proxy for the gap; if we find, numerically, that $\text{MSD}_n \rightarrow 0$ as $\epsilon \rightarrow 0$, that is strong empirical evidence that $\text{Gap}(\mathcal{T}_\epsilon) \rightarrow 0$ too. Again, due to the infimum in the Dirichlet form, positivity of the one step jumping distance of any observable, including MSD_n , does not imply positivity of the gap.

We thus formalize two notions of robustness. Our robustness criteria for a method with a transition operator \mathcal{T}_ϵ for sampling the Boltzmann distribution of potential V_ϵ are

$$\text{Global Robustness Criterion: } \liminf_{\epsilon \rightarrow 0} \text{Gap}(\mathcal{T}_\epsilon) > 0, \quad (1.20)$$

$$\text{Local Robustness Criterion: } \liminf_{\epsilon \rightarrow 0} \frac{\text{MSD}_n}{2 \text{Var}_\mu(X)} > 0. \quad (1.21)$$

Analogous robustness conditions can be constructed for other observables. Failure to be locally robust immediately implies the method cannot be globally robust. Likewise, a globally robust method automatically implies local robustness. However, there may be distributions and methods for which, on a particular observable, the method is locally robust, but for which it fails to be globally robust.

In (1.20) and (1.21), no mention is made of the choice of the numerical parameter σ . Some methods may fail to be robust if σ is not chosen carefully. Thus, we introduce two conditional notions of robustness that depend on σ

$$\text{Global Robustness Criterion with Optimal } \sigma: \liminf_{\epsilon \rightarrow 0} \sup_{\sigma} \text{Gap}(\mathcal{T}_\epsilon(\sigma)) > 0, \quad (1.22)$$

$$\text{Local Robustness Criterion with Optimal } \sigma: \liminf_{\epsilon \rightarrow 0} \sup_{\sigma} \frac{\text{MSD}_n(\sigma)}{2 \text{Var}_\mu(X)} > 0. \quad (1.23)$$

These notions generalize to methods with additional numerical parameters. Obviously, if a method fails to satisfy (1.22), it will fail to satisfy (1.20). Conversely, if the method satisfies (1.20), then it will also satisfy (1.22). Analogous relationships can be formulated for local robustness.

1.3. Results on performance in the presence of roughness. Our key observations and results are:

- (i) At a fixed dimension n , the performance of RWM for potentials of the form (1.2), is globally robust in the sense of (1.20). This is a consequence of Corollary 2.1. Indeed, any method that uses ϵ -independent proposals will similarly satisfy (1.20). Proposals with a sufficiently mild ϵ -dependence will also be robust; see Corollary 2.3. The methods need not be optimally tuned for (1.20) to hold.
- (ii) A rigorous result is established that, subject to certain assumptions, MALA fails to be globally robust when σ is inadequately scaled for potentials like (1.8). This is the content of Theorem 2.3. Specifically, when σ is too large relative to ϵ , the spectral gap will close.
- (iii) Numerical experiments and an explicit example indicate that for $n = 1$, the optimal scaling of MALA is $\sigma \propto \sqrt{\epsilon}$, so that $\text{MSD}_1 \propto \epsilon$. In contrast, for n sufficiently large, the empirical optimal scaling is $\sigma \propto \epsilon$, so that $\text{MSD}_n/n \propto \epsilon^2$. The experiments also indicate that MALA is not locally robust even at an empirically determined optimal σ , (1.23). Thus, the method suffers generically in the $\epsilon \rightarrow 0$.
- (iv) We formulate two alternative methods for potentials of type (1.2) that use large scale information contained in V_0 . The first method, which we call *Modified MALA*, uses the proposals (at $\beta = 1$)

$$X_{k+1}^p = X_k - \frac{\sigma^2}{2} \nabla V_0(X_k) + \sigma \xi_{k+1}. \quad (1.24)$$

This method fits in the class of sampling methods studied in [15] where the authors rigorously established that for the path-wise accuracy of Metropolized integrators it is sufficient to accurately simulate the diffusion term, which (1.24) does.

The second method, we call an *Independence Sampler*, uses proposals (at $\beta = 1$)

$$X_{k+1}^{\text{P}} = Y_{k+1} \sim e^{-V_0(x)} dx. \quad (1.25)$$

The Y_k samples are assumed to be independent, generated by an auxiliary process. Both of these methods are also insensitive to the roughness and outperform RWM. As both of these methods use ϵ -independent proposals Corollary 2.1 allows us to conclude they will also be globally robust in the $\epsilon \rightarrow 0$ limit.

- (v) For potentials that do not admit an obvious decomposition like (1.2) we propose using local entropy approximation, [23, 24] to extract the large scale information needed for either the Modified MALA method or the Independence Sampler.

In Section 2 we identify a bound on the performance of RWM and other algorithms, and we consider the asymptotic behavior of MALA as $\epsilon \rightarrow 0$. In Section 3 we present alternative methods that are also robust to $\epsilon \rightarrow 0$. Numerical experiments are presented in Section 4, and we conclude with a discussion in Section 5.

2. Bounds on performance

In this section we present bounds on performance with respect to the robustness criteria. For any MCMC method, let $q(x \rightarrow y)$ denote the associated proposal kernel, and define

$$R(x, y) = V(x) - V(y) + \log \frac{q(y \rightarrow x)}{q(x \rightarrow y)}. \quad (2.1)$$

Consequently, the proposal X_n^{P} is accepted with probability $F(R(X_n, X_n^{\text{P}}))$. The two forms of F that we consider here are

$$\text{Metropolis: } F(r) = 1 \wedge e^r, \quad (2.2a)$$

$$\text{Barker: } F(r) = (1 + e^{-r})^{-1}. \quad (2.2b)$$

2.1. Roughness independent bounds. For potentials of type (1.2) one can obtain ϵ -independent upper and lower bounds on a variety of quantities. Indeed, by the boundedness assumption in the introduction, we are assured that

$$\text{osc } V_1 = \sup_x V_1(x, x/\epsilon) - \inf_x V_1(x, x/\epsilon) < \infty. \quad (2.3)$$

Our main results in this subsection are the following:

THEOREM 2.1. *Let V_ϵ be a potential of type (1.2) and assume V_1 has uniform in ϵ bounded oscillation in the sense of (2.3). If the sampling strategy uses proposals $q(x \rightarrow y)$ that are ϵ -independent, then for any f , the one step jumping distance is bounded by ϵ independent constants:*

$$\begin{aligned} e^{-2 \text{osc } V_1} \mathbb{E}_{\mu_0} [|f(X_{k+1}) - f(X_k)|^2] &\leq \mathbb{E}_\mu [|f(X_{k+1}) - f(X_k)|^2] \\ &\leq e^{2 \text{osc } V_1} \mathbb{E}_{\mu_0} [|f(X_{k+1}) - f(X_k)|^2]. \end{aligned} \quad (2.4)$$

In (2.4), $\mu_0(dx) \propto e^{-V_0(x)} dx$.

A corollary to this result provides ϵ independent bounds on the spectral gap:

COROLLARY 2.1. *Under the same assumptions as Theorem 2.1*

$$e^{-3 \operatorname{osc} V_1} \operatorname{Gap}(\mathcal{T}_0) \leq \operatorname{Gap}(\mathcal{T}_\epsilon) \leq e^{3 \operatorname{osc} V_1} \operatorname{Gap}(\mathcal{T}_0) \quad (2.5)$$

where \mathcal{T}_0 is the transition operator of the method with proposals generated by q sampling μ_0 .

REMARK 2.1. We emphasize that these results, and the results in the rest of Section 2.1, are all at a fixed dimension n . Constants, like $e^{\pm 2 \operatorname{osc} V_1}$ in (2.4), may depend in an unfavorable way on n . Additionally, Theorem 2.1 and Corollary 2.1 hold independently of the choice of any numerical parameters, like σ in RWM.

To prove Theorem 2.1 and its corollary, we first prove the following bounds on the distribution.

LEMMA 2.1. *Let V_ϵ be a potential of type (1.2), and assume V_1 has uniform in ϵ bounded oscillation in the sense of (2.3). Then*

$$e^{-\operatorname{osc} V_1} \mu_0(dx) \leq \mu(dx) \leq e^{\operatorname{osc} V_1} \mu_0(dx) \quad (2.6)$$

Proof. The proof of this follows from direct estimates on the densities. First,

$$Z^{-1} e^{-V(x)} \geq Z^{-1} e^{-V_0(x)} e^{-\sup V_1(x, x/\epsilon)}$$

while

$$Z = \int e^{-V(x)} \leq \underbrace{\int e^{-V_0(x)} dx}_{\equiv Z_0} e^{-\inf V_1(x, x/\epsilon)}$$

□

As an immediate consequence we have bounds for the mean.

COROLLARY 2.2. *Let V_ϵ satisfy the same assumptions as in Lemma 2.1. Then for any non-negative observable, f , which may depend on a small parameter ϵ ,*

$$e^{-\operatorname{osc} V_1} \mathbb{E}_{\mu_0}[f(X)] \leq \mathbb{E}_\mu[f(X)] \leq e^{\operatorname{osc} V_1} \mathbb{E}_{\mu_0}[f(X)]. \quad (2.7)$$

Similarly we obtain bounds on the variance.

LEMMA 2.2. *Let V_ϵ satisfy the same assumptions as in Lemma 2.1. Then $L^2(\mu)$ and $L^2(\mu_0)$ are equivalent as sets, and*

$$e^{-\operatorname{osc} V_1} \operatorname{Var}_{\mu_0}(f(X)) \leq \operatorname{Var}_\mu(f(X)) \leq e^{\operatorname{osc} V_1} \operatorname{Var}_{\mu_0}(f(X)). \quad (2.8)$$

Proof. For $f \in L^2(\mu)$, Lemma 2.1 ensures

$$e^{-\operatorname{osc} V_1} \int f(x)^2 \mu_0(dx) \leq \int f(x)^2 \mu(dx) < \infty.$$

Hence, $f \in L^2(\mu_0)$ too and $L^2(\mu) \subset L^2(\mu_0)$. A similar computation shows the reverse inclusion. Next, for $f \in L^2(\mu)$, let $f_c(x) = f(x) - \mathbb{E}_{\mu_0}[f(X)]$. Then, using Lemma 2.1 again,

$$\begin{aligned} \operatorname{Var}_\mu(f(X)) &= \operatorname{Var}_\mu(f_c(X)) \\ &= \mathbb{E}_\mu[f_c(X)^2] - \mathbb{E}_\mu[f_c(X)]^2 \leq \mathbb{E}_\mu[f_c(X)^2] \\ &\leq e^{\operatorname{osc} V_1} \mathbb{E}_{\mu_0}[f_c(X)^2] = e^{\operatorname{osc} V_1} \operatorname{Var}_{\mu_0}[f(X)]. \end{aligned}$$

Reversing the roles of μ and μ_0 establishes the analogous lower bound in (2.8). \square

REMARK 2.2. For potentials satisfying the assumptions of the preceding results, it will often be sufficient to examine $\mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2]$ since a prior bound on the variance is provided by Lemma 2.2.

Proof. (Proof of Theorem 2.1.) Given potential $V = V_0 + V_1$ of the form (1.2) satisfying the assumptions:

$$R(x, y) = \underbrace{(V(x) - V_0(x))}_{V_1(x, x/\epsilon)} - \underbrace{(V(y) - V_0(y))}_{V_1(y, y/\epsilon)} + \underbrace{\log \left(\frac{e^{-V_0(y)} q(y \rightarrow x)}{e^{-V_0(x)} q(x \rightarrow y)} \right)}_{R_0(x, y)}. \quad (2.9)$$

Since $\text{osc } V_1$ is bounded,

$$R_0(x, y) + \text{osc } V_1 \geq R(x, y) \geq R_0(x, y) - \text{osc } V_1. \quad (2.10)$$

Consequently, for either choice of (2.2),

$$e^{\text{osc } V_1} F(R_0(x, y)) \geq F(R(x, y)) \geq e^{-\text{osc } V_1} F(R_0(x, y)). \quad (2.11)$$

Since the proposal, $q(x \rightarrow y)$, is also assumed ϵ -independent, then, for any X_0 ,

$$\begin{aligned} \mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] &= \mathbb{E}_\mu[|f(X_{k+1}^p) - f(X_k)|^2 F(R(X_k, X_{k+1}^p))] \\ &\geq e^{-\text{osc } V_1} \mathbb{E}_\mu[|f(X_{k+1}^p) - f(X_k)|^2 F(R_0(X_k, X_{k+1}^p))] \end{aligned} \quad (2.12a)$$

$$\mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] \leq e^{\text{osc } V_1} \mathbb{E}_\mu[|f(X_{k+1}^p) - f(X_k)|^2 F(R_0(X_k, X_{k+1}^p))]. \quad (2.12b)$$

In the preceding upper and lower bounds on $\mathbb{E}[|X_1 - X_0|^2]$, no ϵ is present. If we now apply Lemma 2.1 to the upper and lower bounds in (2.12), we obtain (2.4). \square

Corollary 2.1 follows from the theorem and Lemma 2.2.

Proof. (Proof of Corollary 2.1.) For any non-constant f ,

$$\frac{\mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2]}{2 \text{Var}_\mu(f(X))} \leq e^{3 \text{osc } V_1} \frac{\mathbb{E}_{\mu_0}[|f(X_{k+1}) - f(X_k)|^2]}{2 \text{Var}_{\mu_0}(f(X))}.$$

Taking the infimum over $L^2(\mu)$,

$$\text{Gap}(\mathcal{T}_\epsilon) \leq e^{3 \text{osc } V_1} \inf_{f \in L^2(\mu)} \frac{\mathbb{E}_{\mu_0}[|f(X_{k+1}) - f(X_k)|^2]}{2 \text{Var}_{\mu_0}(f(X))} = e^{3 \text{osc } V_1} \text{Gap}(\mathcal{T}_0).$$

This last equality is due to $L^2(\mu)$ and $L^2(\mu_0)$ being equivalent as sets. An analogous computation establishes the lower bound. \square

Theorem 2.1 and Corollary 2.1 immediately apply to RWM, as it has a proposal independent of ϵ . Consequently, for RWM,

$$\text{MSD}_n = \mathbb{E}_\mu[|X_{k+1} - X_k|^2] \sim \sigma^2.$$

In contrast, as MALA proposals include gradients of the potential, for potentials of the form (1.2), these results will not apply.

The results can be modified to allow for proposals that have some ϵ dependence:

THEOREM 2.2. *Let V_ϵ be a potential of the type (1.2), and assume V_1 has bounded oscillation, uniformly in ϵ , in the sense of (2.3). Assume the sampling strategy's proposal kernel, $q(x \rightarrow y)$, satisfies the inequality*

$$Cq_0(x \rightarrow y) \leq q(x \rightarrow y) \leq Dq_0(x \rightarrow y), \quad (2.13)$$

where $q_0(x \rightarrow y)$ is an ϵ -independent proposal kernel. Then the performance, as measured by MSD_n , is ϵ -independent:

$$\begin{aligned} e^{-2 \text{osc } V_1} \frac{C}{D} \mathbb{E}_{\mu_0}[|f(X_{k+1}) - f(X_k)|^2] &\leq \mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] \\ &\leq e^{2 \text{osc } V_1} \frac{D}{C} \mathbb{E}_{\mu_0}[|f(X_{k+1}) - f(X_k)|^2]. \end{aligned} \quad (2.14)$$

Proof. As in the proof of Theorem 2.1, we begin by writing

$$\begin{aligned} R(x, y) &= (V(x) - V_0(x)) - (V(y) - V_0(y)) + \log \left(\frac{e^{-V_0(y)} q(y \rightarrow x)}{e^{-V_0(x)} q(x \rightarrow y)} \right) \\ &\leq V_1(x) - V_1(y) + \log \left(\frac{D e^{-V_0(y)} q_0(y \rightarrow x)}{C e^{-V_0(x)} q_0(x \rightarrow y)} \right) \\ &\leq \text{osc } V_1 + \log \frac{D}{C} + R_0(x, y). \end{aligned}$$

Analogously,

$$R(x, y) \geq -\text{osc } V_1 + \log \frac{C}{D} + R_0(x, y).$$

Consequently, for a function f ,

$$\begin{aligned} \mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] &\leq e^{\text{osc } V_1} \frac{D}{C} \mathbb{E}_\mu[|f(X_{k+1}^p) - f(X_k)|^2 F(R_0(X_k, X_{k+1}^p))] \\ \mathbb{E}_\mu[|f(X_{k+1}) - f(X_k)|^2] &\geq e^{-\text{osc } V_1} \frac{C}{D} \mathbb{E}_\mu[|f(X_{k+1}^p) - f(X_k)|^2 F(R_0(X_k, X_{k+1}^p))]. \end{aligned}$$

Using Lemma 2.1 yields the result. \square

An analog of Corollary 2.1 can then be established, which we present without proof:

COROLLARY 2.3. *Under the same assumptions as Theorem 2.2,*

$$e^{-3 \text{osc } V_1} \frac{C}{D} \text{Gap}(\mathcal{T}_0) \leq \text{Gap}(\mathcal{T}_\epsilon) \leq e^{3 \text{osc } V_1} \frac{D}{C} \text{Gap}(\mathcal{T}_0). \quad (2.15)$$

A method where Theorem 2.2 would apply is the tamed (or truncated) MALA, [25–27]. Given a parameter $\delta > 0$, one form of the tamed MALA (at $\beta = 1$) is

$$X_{k+1}^p = X_k - \frac{\sigma^2}{2} \frac{\nabla V(X_k)}{1 \vee (\delta |\nabla V(X_k)|)} + \sigma \xi_{k+1}, \quad \xi_{k+1} \sim N(0, I). \quad (2.16)$$

This has the effect of ensuring that the gradient term is never larger than $1/\delta$, mitigating stiff regimes of the state space.

COROLLARY 2.4. *The performance, as measured by the spectral gap, of the tamed MALA on potentials of the type (1.2), is insensitive to ϵ .*

Proof. For the tamed MALA as implemented in (2.16), let

$$f_\delta(X_k) = -\frac{\nabla V(X_k)}{1 \vee (\delta|\nabla V(X_k)|)}.$$

Then its proposal density is

$$q(x \rightarrow y) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left| y - x - \frac{\sigma^2}{2} f_\delta(x) \right|^2 \right\}. \quad (2.17)$$

Since $|f_\delta(x)| \leq \delta^{-1}$,

$$2|y - x|^2 - \frac{\sigma^4}{2\delta^2} \leq |y - x - \frac{\sigma^2}{2} f_\delta(x)|^2 \leq 2|y - x|^2 + \frac{\sigma^4}{2\delta^2}. \quad (2.18)$$

Setting

$$q_0(x \rightarrow y) \propto \exp \left\{ -\frac{1}{2\sigma^2} |y - x|^2 \right\}, \quad (2.19)$$

we have that the tamed MALA proposal kernel satisfies

$$e^{-\frac{\sigma^4}{2\delta^2}} q_0(x \rightarrow y) \leq q(x \rightarrow y) \leq e^{\frac{\sigma^4}{2\delta^2}} q_0(x \rightarrow y). \quad (2.20)$$

Thus, Corollary 2.3 applies. \square

2.2. Optimal proposal variance for MALA. Next, we perform a formal calculation on the performance of MALA with respect to proposal variance σ and roughness ϵ . Though this does not lead to any immediate conclusions, it may guide future analysis.

We define $M(\sigma)$ to be the MSD_n obtained with σ , assuming stationarity, $x \sim \mu$, with the Barker rule, (2.2b). Thus, the optimal value of σ can be obtained by solving $M'(\sigma) = 0$. Proceeding with this strategy we have

$$M(\sigma) = \int |x - y|^2 F(R(x, y; \sigma)) g(y; x, \sigma) dy \mu(dx), \quad (2.21)$$

where g is the Gaussian density of the normal distribution $N(x - \frac{\sigma^2}{2} \nabla V(x), \sigma^2 I)$ and $\mu(dx)$ is the Boltzmann distribution.

PROPOSITION 2.1. *At a critical point of (2.21), where $M'(\sigma) = 0$,*

$$0 = \frac{\sigma^4}{2} \underbrace{\mathbb{E} [|x - y|^2 F^2 |\nabla V(x)|^2]}_{\equiv A} + \sigma^2 n \underbrace{\mathbb{E} [|x - y|^2 F]}_{=M} - \underbrace{\mathbb{E} [|x - y|^4 F]}_{\equiv C}, \quad (2.22)$$

or

$$\sigma^2 = 2 \left\{ \frac{nM}{C} + \sqrt{\left(\frac{nM}{C} \right)^2 + \frac{2A}{C}} \right\}^{-1}. \quad (2.23)$$

While this is a formal calculation, it does not require any particular assumptions on the potential beyond the associated distribution being normalizable and having the necessary moments.

Proof. See Appendix A for the derivation of (2.22). The expression (2.23) follows by solving for σ^2 . \square

2.3. Insights from the high dimensional limit. We recall from the introduction that for product distributions such as (1.7), as $n \rightarrow \infty$, the MSD_n per degree of freedom is given by (1.10). Here, we consider the $n \rightarrow \infty$ limit at fixed ϵ , and then examine the impact of the roughness through ϵ . This corresponds to a consideration of the order $\lim_{\epsilon \rightarrow 0} \lim_{n \rightarrow \infty}$. Elsewhere in this manuscript, we focus on fixed n while letting $\epsilon \rightarrow 0$.

The optimal ℓ for (1.10) can be found using (1.12) and (1.13). From (1.11) we can then find the optimal choice of σ and obtain the scaling of the MSD_n/n in (1.10). After changing coordinates to $\lambda = -\ell^{1/I} \sqrt{\mathcal{K}[v]}/2$, and optimizing in this transformed coordinate, one can deduce that the optimal $\ell^2 \sim \mathcal{K}^{-I}$. Consequently, for potentials like (1.8), the optimal scaling, as $n \rightarrow \infty$, is $\sigma^2 \sim (n\mathcal{K})^{-I}$ and

- (i) For RWM, $I = 1$ and $\mathcal{K}[v_\epsilon] = \mathbb{E}_\mu[(v'_\epsilon)^2]$, so that $\sigma^2 \sim \epsilon^2/n$;
- (ii) For MALA, $I = 1/3$ and $\mathcal{K}[v_\epsilon] = \mathbb{E}_\mu[5(v''_\epsilon)^2 + 3(v''_\epsilon)^3]/48 \sim \epsilon^{-6}$ so that $\sigma^2 \sim \epsilon^2/n^{1/3}$.

In this limit, both RWM and MALA will fail to be locally robust at optimal σ (in the sense of (1.23)) as $\epsilon \rightarrow 0$. Thus, they will also suffer when a suboptimal value is selected for any observable. This does not contradict the analysis of Section 2.1. There, in the product case, our lower bound would have the prefactor $e^{-n \text{osc } v_1}$, which vanishes more rapidly than the ϵ^2/n we have here.

2.4. Scaling in dimension one. We consider the question of sampling from the distribution associated with $V(x) = \frac{1}{2}\epsilon^{-1}x^2$ in $n = 1$, using MALA. While this may seem odd, as this corresponds to a Gaussian distribution, it reveals a distinct scaling in low dimensions. It can be interpreted as the necessary scaling to efficiently sample individual modes of the highly multi-modal landscape in Figure 1.1(b), or alternatively, as the analog of the stiff differential equation problem $\dot{x} = -\epsilon^{-1}x$, $0 < \epsilon \ll 1$.

In this case,

$$X_{k+1}^p = X_k - \frac{\sigma^2}{2}\epsilon^{-1}X_k + \sigma\xi_{k+1}, \quad (2.24)$$

and the term in the accept/reject rule is

$$R(x, y) = \frac{\sigma^2}{8\epsilon^2}(x^2 - y^2). \quad (2.25)$$

With the aid of a computer algebra system (see, also, Appendix B) we conclude

$$\begin{aligned} \text{MSD}_1 &= \epsilon \frac{2\delta}{\pi(4 + \delta(-2 + \delta))} \left\{ (8 + \delta^3) \arctan \left(\sqrt{\frac{8}{\delta^3}} \right) - 2\sqrt{2}\delta^{3/2} \right\} \\ &= \epsilon m(\delta). \end{aligned} \quad (2.26)$$

A consequence of this computation is that it provides an explicit example for which MALA will not satisfy the local robustness criterion, even with optimal σ , (1.23).

The function $m(\delta)$ has a single maximum (see Figure B.1), and the maximum value is achieved at $\delta_\star = 1.27797$, and $m(\delta_\star) = 1.8494$. Thus, the optimal scaling for the time step is

$$\sigma = \sqrt{2\delta_\star\epsilon} \approx 1.59873\sqrt{\epsilon}.$$

This scaling is different from the scaling found in Section 2.2. We note that this optimal choice is smaller than the Euler-Maruyama stability threshold which requires $\sigma \leq 2\sqrt{\epsilon}$. Additionally, the acceptance rate at this optimal value is 0.70, which is quite different than the $n \rightarrow \infty$ acceptance rate of 0.57.

2.5. Poorly scaled proposals. We show rigorously that, under particular assumptions, if σ is improperly scaled, MALA will fail to be robust in the sense of (1.20). Our approach is based on the method of [21] (see, also, [6, 28]) using the relationship between the Dirichlet form and the spectral gap, (1.5).

As in [21], the interval Λ is bounded in terms of the *conductance* of a set, K , with $\mu(K) \in (0, 1/2)$. Using the observable

$$f(x) = \frac{1}{\sqrt{(1 - \mu(K))\mu(K)}}(1_K(x) - \mu(K)), \quad (2.27)$$

we immediately have for a method with the transition operator \mathcal{T}

$$\text{Gap}(\mathcal{T}) \leq 2\mathbb{P}_\mu(X_{k+1} \in K^c \mid X_k \in K). \quad (2.28)$$

PROPOSITION 2.2. *Assume that*

- (1) *The potential is $V(x) = V_0(x) + V_1(x)$, where V_0 is trapping and $V_1(x) = V_1(x, x/\epsilon)$ is bounded.*
- (2) *A concentration inequality holds for μ*

$$\mathbb{P}_\mu(|X| \geq t) \leq C_0 e^{-\kappa t^p} \quad (2.29)$$

where C_0 , κ , and p are independent of ϵ .

- (3) *There exists $m > 0$ such that for all ϵ sufficiently small, the set $K = \{x \mid |x| > m\}$ satisfies*

$$0 < \underline{\mu}_K < \mu(K) < \bar{\mu}_K < \frac{1}{2}, \quad (2.30)$$

where $\underline{\mu}_K$ and $\bar{\mu}_K$ are independent of ϵ .

- (4) *The gradient ∇V_0 satisfies a growth bound with an exponent $\delta > 0$, such that for all x*

$$|\nabla V_0(x)| \leq D_0 + D_1|x|^\delta. \quad (2.31)$$

- (5) *Let $\alpha \geq 0$, $\theta \in \mathbb{R}$, and $\gamma > 0$ satisfy*

$$2\alpha + \delta\gamma + \theta < 1. \quad (2.32)$$

Let $\sigma = \epsilon^\alpha$, and define the sets

$$K_{\epsilon, \gamma} = \{m < |x| \leq \epsilon^{-\gamma}\} \subset K, \quad (2.33)$$

$$L_{\epsilon, \theta} = \{x \mid |\nabla V_1(x)| > \epsilon^{\theta-1}\}. \quad (2.34)$$

Then it holds

$$\begin{aligned} \text{Gap}(\mathcal{T}_{\text{MALA}}) &\lesssim e^{-\frac{1}{32}\epsilon^{-2(1-\alpha-\theta)}} + e^{-\kappa\epsilon^{-p\gamma}} \\ &\quad + \mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon, \theta}^c, X_k \in K \cap L_{\epsilon, \theta}^c). \end{aligned} \quad (2.35)$$

In (2.35), the relation \lesssim is introduced. Generically, if $a \lesssim b$, then there is a constant, C , independent of b , such that $a \leq Cb$.

Before proving the result, a few remarks are in order.

- (1) The concentration inequality for μ will hold provided μ_0 , the distribution associated with V_0 , has a concentration inequality. This is a consequence of Corollary 2.2.
- (2) The bound (2.30) will similarly hold provided such a condition holds for V_0 . Again, this is a consequence of Corollary 2.2 by selecting a set K for which $0 < \mu_0(K) < \frac{1}{2}e^{-\text{osc } V_1}$.
- (3) The set $L_{\epsilon,\theta}$ captures the points in space where ∇V_1 will be large. To obtain a rate, it is necessary to know the measure of $L_{\epsilon,\theta}^c$, requiring additional details on the structure of the potential. In particular, it is essential to estimate

$$\mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon,\theta}^c, X_k \in K \cap L_{\epsilon,\theta}^c), \quad (2.36)$$

which is clearly related to the conductance of set K with the additional constraints that X_k and X_{k+1} both reside in $L_{\epsilon,\theta}^c$, the set where the gradient is “small”. See Figure 2.1 for an illustration of these sets.

- (4) To see how the set (2.36) can become small, consider the case that $V_0 = x^2/2$ and $V_1 = \cos(x/\epsilon)$. Then $\nabla V_1(x) = -\sin(x/\epsilon)/\epsilon$. For any $\theta > 0$ and all $\epsilon > 0$ sufficiently small

$$L_{\epsilon,\theta}^c = \{x \mid |\sin(x/\epsilon)| \leq \epsilon^\theta\} \subset \bigcup_{k \in \mathbb{Z}} [-\epsilon^{1+\theta}, \epsilon^{1+\theta}] + \epsilon\pi k.$$

Then, recalling that $g(y; x, \sigma)$ is the Gaussian density of the proposal

$$\mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon,\theta}^c, X_k \in K \cap L_{\epsilon,\theta}^c) \leq \int_{K \cap L_{\epsilon,\theta}^c} \mu(dx) \int_{K^c \cap L_{\epsilon,\theta}^c} g(y; x, \sigma) dy.$$

For any $x \in K \cap L_{\epsilon,\theta}^c$

$$\begin{aligned} \int_{K^c \cap L_{\epsilon,\theta}^c} g(y; x, \sigma) dy &\lesssim \sum_{|\epsilon k \pi| \leq m} \int_{[-\epsilon^{1+\theta}, \epsilon^{1+\theta}] + \epsilon\pi k} \frac{1}{\sqrt{2\pi\sigma^2}} dy \\ &\lesssim \epsilon^{1+\theta-\alpha} \sum_{|\epsilon k \pi| \leq m} 1 \lesssim \epsilon^{\theta-\alpha}. \end{aligned} \quad (2.37)$$

Additionally,

$$\begin{aligned} \mu(K \cap L_{\epsilon,\theta}^c) &\lesssim \mu_0(K \cap L_{\epsilon,\theta}^c) \lesssim \sum_{|\epsilon k \pi| > m} \int_{[-\epsilon^{1+\theta}, \epsilon^{1+\theta}] + \epsilon\pi k} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &\lesssim \sum_{|\epsilon k \pi| > m} \epsilon^{1+\theta} e^{-(\epsilon k \pi)^2/2} \\ &\lesssim \epsilon^{1+\theta} \int_{\frac{m}{\epsilon\pi}}^{\infty} e^{-(\epsilon k \pi)^2/2} dk \lesssim \epsilon^\theta. \end{aligned} \quad (2.38)$$

Combining (2.37) and (2.38), (2.36) is thus bounded as

$$\mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon,\theta}^c, X_k \in K \cap L_{\epsilon,\theta}^c) \lesssim \epsilon^{2\theta-\alpha}$$

For $\theta > 1/2$ and γ and α sufficiently small the above expression is $\ll \epsilon$.

- (5) As the result requires $\alpha < 1/2$, $\sigma \gg \sqrt{\epsilon}$ in this regime.

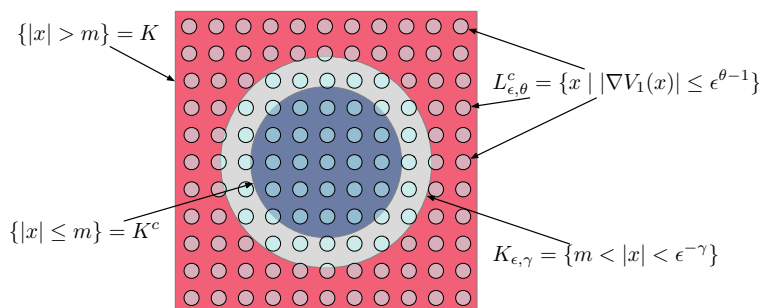


FIG. 2.1. A diagram of the sets discussed in Proposition 2.2. For the sake of clarity, we have plotted $L_{\epsilon, \theta}^c$, and not $L_{\epsilon, \theta}$, the set where the gradient of V_1 is “small”.

Proof. (Proof of Proposition 2.2.) Since $\mu(K)$ is, by assumptions, uniformly bounded away from 0 and $1/2$ as $\epsilon \rightarrow 0$, it is sufficient to study $\mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K)$. For all sufficiently small ϵ we can write

$$\begin{aligned} \mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K) &= \mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K_{\epsilon, \gamma}) \\ &\quad + \mathbb{P}_\mu(X_{k+1} \in K^c, |X_k| > \epsilon^{-\gamma}) \\ &\leq \mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K_{\epsilon, \gamma}) + C_0 e^{-\kappa \epsilon^{-\gamma p}}, \end{aligned}$$

where we have used the assumed concentration inequality at the end.

Next, we divide the set depending on whether X_k and X_{k+1} lie in the high or low gradient sets

$$\begin{aligned} \mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K_{\epsilon, \gamma}) &= \mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \\ &\quad + \mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon, \theta}, X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}^c) \\ &\quad + \mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon, \theta}^c, X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}^c). \end{aligned} \quad (2.39)$$

We consider the first of the three terms on the right-hand side of (2.39):

$$\begin{aligned} &\mathbb{P}_\mu(X_{k+1} \in K^c, X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \\ &= \mathbb{P}_\mu(X_{k+1} \in K^c \mid X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \mathbb{P}(X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \\ &\leq \mathbb{P}_\mu(X_{k+1} \in K^c \mid X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \\ &\leq \mathbb{P}_\mu(X_{k+1}^p \in K^c \mid X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}). \end{aligned}$$

Conditioned on $X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}$,

$$\begin{aligned} |X_{k+1}^p| &= |X_k - \frac{\sigma^2}{2} \nabla V_0(X_k) - \frac{\sigma^2}{2} \nabla V_1(X_k) + \sigma \xi_{k+1}| \\ &\geq \frac{\sigma^2}{2} |\nabla V_1(X_k)| - |X_k| - \frac{\sigma^2}{2} |\nabla V_0(X_k)| - \sigma |\xi_{k+1}| \\ &\geq \frac{1}{2} \epsilon^{2\alpha + \theta - 1} - \epsilon^{-\gamma} - \frac{D_0}{2} \epsilon^{2\alpha} - \frac{D_1}{2} \epsilon^{2\alpha - \delta \gamma} - \epsilon^\alpha |\xi_{k+1}|, \end{aligned}$$

therefore

$$\begin{aligned} &\mathbb{P}_\mu(m \geq |X_{k+1}^p| \mid X_k \in K_{\epsilon, \gamma} \cap L_{\epsilon, \theta}) \\ &\leq \mathbb{P}(m \geq \frac{1}{2} \epsilon^{2\alpha + \theta - 1} - \epsilon^{-\gamma} - \frac{D_0}{2} \epsilon^{2\alpha} - \frac{D_1}{2} \epsilon^{2\alpha - \delta \gamma} - \epsilon^\alpha |\xi_{k+1}|) \\ &= \mathbb{P}(|\xi_{k+1}| \geq \frac{1}{2} \epsilon^{\alpha + \theta - 1} - \epsilon^{-\gamma - \alpha} - \frac{D_0}{2} \epsilon^\alpha - \frac{D_1}{2} \epsilon^{\alpha - \delta \gamma} - \epsilon^{-\alpha} m) \\ &= \mathbb{P}(|\xi_{k+1}| \geq \epsilon^{-(1 - \alpha - \theta)} (\frac{1}{2} - \epsilon^{1 - \theta - 2\alpha - \delta \gamma} - \frac{D_0}{2} \epsilon^{1 - \theta} - \frac{D_1}{2} \epsilon^{1 - \theta - \delta \gamma} - \epsilon^{1 - \theta - 2\alpha} m)). \end{aligned}$$

By our assumptions on the exponents, for all ϵ small enough,

$$\mathbb{P}_\mu(m \geq |X_{k+1}^p| \mid X_k \in K_{\epsilon,\gamma} \cap L_{\epsilon,\theta}) \leq \mathbb{P}(|\xi_{k+1}| \geq \tfrac{1}{4}\epsilon^{-(1-\alpha-\theta)}) \lesssim e^{-\frac{1}{32}\epsilon^{-2(1-\alpha-\theta)}},$$

where we have a tail inequality on $N(0, I_n)$.

For the second term in (2.39) we use reversibility to get

$$\mathbb{P}_\mu(X_{k+1} \in K^c \cap L_{\epsilon,\theta}, X_k \in K_{\epsilon,\gamma} \cap L_{\epsilon,\theta}^c) = \mathbb{P}_\mu(X_{k+1} \in K_{\epsilon,\gamma} \cap L_{\epsilon,\theta}^c, X_k \in K^c \cap L_{\epsilon,\theta}).$$

Using the same approach as above we have

$$\begin{aligned} & \mathbb{P}_\mu(X_{k+1} \in K_{\epsilon,\gamma} \cap L_{\epsilon,\theta}^c \mid X_k \in K^c \cap L_{\epsilon,\theta}) \\ & \leq \mathbb{P}_\mu(\epsilon^{-\gamma} \geq |X_{k+1}^p| \mid X_k \in K^c \cap L_{\epsilon,\theta}) \\ & \leq \mathbb{P}(\epsilon^{-\gamma} \geq \tfrac{1}{2}\epsilon^{2\alpha+\theta-1} - \epsilon^{-\alpha}m - \tfrac{D_0}{2}\epsilon^{2\alpha} - \tfrac{D_1}{2}\epsilon^{2\alpha}m^\delta - \epsilon^\alpha|\xi_{k+1}|) \\ & = \mathbb{P}(|\xi_{k+1}| \geq \tfrac{1}{2}\epsilon^{\alpha+\theta-1} - \epsilon^{-\alpha}m - \tfrac{D_0}{2}\epsilon^\alpha - \tfrac{D_1}{2}\epsilon^\alpha m^\delta - \epsilon^{-\alpha-\gamma}) \\ & = \mathbb{P}(|\xi_{k+1}| \geq \epsilon^{-(1-\alpha-\theta)}(\tfrac{1}{2} - \epsilon^{1-2\alpha-\theta}m - \tfrac{D_0}{2}\epsilon^{1-\theta} - \tfrac{D_1}{2}\epsilon^{1-\theta}m^\delta - \epsilon^{1-\theta-2\alpha-\gamma})). \end{aligned}$$

Again, by our assumptions on the exponents and a tail inequality on the Gaussian

$$\mathbb{P}_\mu(X_{k+1} \in K_{\epsilon,\gamma} \cap L_{\epsilon,\theta}^c \mid X_k \in K^c \cap L_{\epsilon,\theta}) \leq \mathbb{P}(|\xi_{k+1}| \geq \tfrac{1}{4}\epsilon^{-(1-\alpha-\theta)}) \lesssim e^{-\frac{1}{32}\epsilon^{-2(1-\alpha-\theta)}}.$$

Combining these estimates we obtain (2.35). \square

The next proposition provides an estimate of the last term in (2.35) for a particular class of potentials.

THEOREM 2.3. *Assume that*

(1) *There exists $p \geq 1$ and constants V_{\min} and $C_p > 0$*

$$V_0(x) \geq V_{\min} + C_p \sum_{i=1}^n |x_i|^p, \quad (2.40)$$

(2) *The growth bound of Proposition 2.2 holds for ∇V_0 .*

(3) *The potential V_1 takes the form*

$$V_1(x) = \sum_{i=1}^n A_i \cos\left(\frac{x}{\ell_i \epsilon}\right) \quad (2.41a)$$

$$0 < \underline{A} \leq |A_i| \leq \overline{A} < \infty, \quad 0 < \underline{\ell} \leq \ell_i \leq \overline{\ell} < \infty. \quad (2.41b)$$

(4) α, γ, δ , and θ satisfy (2.32).

Then there exists a constant κ such that

$$\text{Gap}(\mathcal{T}_{\text{MALA}}) \lesssim e^{-\frac{1}{2}\kappa\epsilon^{-\gamma p}} + e^{-\frac{1}{32}\epsilon^{-2(1-\alpha-\theta)}} + \epsilon^{n(2\theta-\alpha)}. \quad (2.42)$$

Before proving this result we make a few remarks:

- (1) If V_0 is harmonic it will satisfy these assumptions with $p = 2$ and $\delta = 1$.
- (2) As long as $\alpha < 2\theta$, the performance will deteriorate as $\epsilon \rightarrow 0$. By (2.32), this necessitates $\alpha < 2/5$.

- (3) In the numerical experiments discussed below, MSD_1 was found to be maximized when $\sigma \sim \sqrt{\epsilon}$, while in higher dimensions, MSD_n was maximized when $\sigma \sim \epsilon$. Thus, there is a “gap” between values of α for which Theorem 2.3 predicts poor performance, and values for which we numerically observe peak performance. We interpret this as a gap in the analysis because, even at the optimal choice of σ , our results show $\text{MSD}_n \rightarrow 0$ as $\epsilon \rightarrow 0$, implying the method is not locally (or globally) robust in the sense of (1.23).
- (4) As n increases, the gap vanishes more and more rapidly as $\epsilon \rightarrow 0$.

Proof. (Proof of Theorem 2.3.) A concentration inequality will obviously hold for μ , as such an inequality holds for μ_0 , with $\mathbb{P}_{\mu_0}(|X| > t) \lesssim e^{-\frac{1}{2}\kappa t^p}$; this is a consequence of (2.40). The constant κ is related to C_p , p , and the dimension.

Since the concentration inequality holds, and V_0 is trapping while V_1 is bounded, Corollary 2.2 ensures that a set $K = \{|x| > m\}$ exists and satisfies (2.30). Thus, the assumptions of Proposition 2.2 hold.

Since

$$\mathbb{P}_{\mu}(X_{k+1} \in K^c \cap L_{\epsilon, \theta}^c, X_k \in K \cap L_{\epsilon, \theta}^c) \leq \int_{K \cap L_{\epsilon, \theta}^c} \mu(dx) \int_{K^c \cap L_{\epsilon, \theta}^c} g(y; x, \sigma) dy,$$

it will be sufficient to estimate $\mu(K \cap L_{\epsilon, \theta}^c)$ and the integral $\int_{K^c \cap L_{\epsilon, \theta}^c} g(y; x, \sigma) dy$ for $x \in K \cap L_{\epsilon, \theta}^c$.

We first bound the set $L_{\epsilon, \theta}^c$. For all ϵ small enough

$$L_{\epsilon, \theta}^c \subset \prod_{i=1}^n \left\{ \left| \frac{A_i}{\ell_i} \sin \left(\frac{x_i}{\ell_i \epsilon} \right) \right| \leq \epsilon^\theta \right\} \subset \prod_{i=1}^n \bigcup_{k_i \in \mathbb{Z}} \underbrace{\left[-\frac{\ell_i^2}{|A_i|} \epsilon^{1+\theta}, \frac{\ell_i^2}{|A_i|} \epsilon^{1+\theta} \right] + \epsilon \ell_i \pi k_i}_{\equiv B_{i, k_i}}.$$

Next, we bound the set K by

$$K \subset \bigcup_{j=1}^n \{|x_j| > m/\sqrt{n}\}$$

Consequently, for all ϵ small enough,

$$L_{\epsilon, \theta}^c \cap K \subset \bigcup_{j=1}^n \left(\bigcup_{|k_j| > \frac{m}{\epsilon \ell_j \pi \sqrt{n}}} B_{j, k_j} \times \prod_{i \neq j} \bigcup_{k_i \in \mathbb{Z}} B_{i, k_i} \right). \quad (2.43)$$

Similarly

$$K^c \subset \prod_{j=1}^n \{|x_j| \leq m\},$$

and

$$L_{\epsilon, \theta}^c \cap K^c \subset \prod_{i=1}^n \bigcup_{|k_i| < \frac{m}{\epsilon \ell_i \pi}} B_{i, k_i}. \quad (2.44)$$

Now, using (2.40) and (2.43),

$$\begin{aligned} \mu(L_{\epsilon,\theta}^c \cap K) &\leq \sum_{j=1}^n \mu \left(\bigcup_{|k_j| > \frac{m}{\epsilon \ell_j \pi \sqrt{n}}} B_{j,k_j} \times \prod_{i \neq j} \bigcup_{k_i \in \mathbb{Z}} B_{i,k_i} \right) \\ &\lesssim \sum_{j=1}^n \left(\sum_{|k_j| > \frac{m}{\epsilon \ell_j \pi \sqrt{n}}} \int_{B_{j,k_j}} e^{-C_p |x_j|^p} dx_j \right) \prod_{i \neq j} \left(\sum_{k_i} \int_{B_{i,k_i}} e^{-C_p |x_i|^p} dx_i \right). \end{aligned}$$

The first term in the above sum is as in the example (2.38)

$$\begin{aligned} \sum_{|k_j| > \frac{m}{\epsilon \ell_j \pi \sqrt{n}}} \int_{B_{j,k_j}} e^{-C_p |x_j|^2} dx_j &\lesssim \epsilon^{1+\theta} \sum_{|k_j| > \frac{m}{\epsilon \ell_j \pi \sqrt{n}}} e^{-C_p (\epsilon \ell_j \pi k_j)^p} \\ &\lesssim \epsilon^{1+\theta} \int_{\frac{m}{\epsilon \ell_j \pi \sqrt{n}}}^{\infty} e^{-C_p (\epsilon \ell_j \pi k_j)^p} dk_j \lesssim \epsilon^{\theta}. \end{aligned}$$

The other terms are treated the same way

$$\sum_{k_i} \int_{B_{i,k_i}} e^{-C_p |x_i|^p} dx_i \lesssim \epsilon^{1+\theta} \int_0^{\infty} e^{-C_p (\epsilon \ell_i \pi k_i)^p} dk_i \lesssim \epsilon^{\theta}.$$

Consequently

$$\mu(L_{\epsilon,\theta}^c \cap K) \lesssim \epsilon^{n\theta}. \quad (2.45)$$

For $x \in L_{\epsilon,\theta}^c \cap K$, using (2.44),

$$\begin{aligned} \int_{L_{\epsilon,\theta}^c \cap K^c} g(y; x, \sigma) dy &\lesssim \prod_{i=1}^n \sigma^{-1} \text{Leb.} \left(\bigcup_{|k_i| < \frac{m}{\epsilon \ell_i \pi \sqrt{n}}} \left[-2 \frac{\ell_i}{A_i} \epsilon^{1+\theta}, 2 \frac{\ell_i}{A_i} \epsilon^{1+\theta} \right] + \epsilon \ell_i \pi k_i \right) \\ &\lesssim \prod_{i=1}^n \left(\epsilon^{1+\theta-\alpha} \sum_{|k_i| < \frac{m}{\epsilon \ell_i \pi \sqrt{n}}} 1 \right) \lesssim \epsilon^{n(\theta-\alpha)}. \end{aligned} \quad (2.46)$$

Combining this with (2.45) yields the result. \square

3. Alternative algorithms for rough landscapes

Having concluded that there are impediments to MALA for sampling on rough landscapes, but desiring a method that may perform better than RWM, we propose two methods here. Both these methods require a smoothed version of the potential which may not be accessible. Indeed, as the potential of interest is unlikely to admit a simple decomposition like (1.2), we also propose a method for approximating a smoothed landscape.

3.1. Modified MALA. The first method, which we call Modified MALA, involves generating proposals from an auxiliary potential, U . The algorithm generates the proposals

$$X_{k+1}^p = X_k - \frac{\sigma^2}{2} \nabla U(X_k) + \sigma \xi_{k+1}, \quad (3.1)$$

and $R(x, y)$ is modified to be

$$R(x, y) = V(x) - V(y) - \frac{1}{2\sigma^2}|x - y + \nabla U(y) \frac{\sigma^2}{2}|^2 + \frac{1}{2\sigma^2}|y - x + \nabla U(x) \frac{\sigma^2}{2}|^2. \quad (3.2)$$

If the scale separation in (1.2) holds, we might take $U = V_0$ to get (1.24). Since the proposal is ϵ -independent, Corollary 2.1 applies to this method.

3.2. Independence sampler. The second method we propose is to generate auxiliary samples, Y_k , from another distribution $\propto e^{-U(x)}dx$, and perform independence sampling against $\mu(dx) \propto e^{-V(x)}dx$. More explicitly,

$$\Delta(x) = V(x) - U(x), \quad R(x, y) = \Delta(x) - \Delta(y) \quad (3.3a)$$

$$X_{k+1} = \begin{cases} Y_{k+1}, & \text{with probability } F(R(X_k, Y_{k+1})), \\ X_k, & \text{otherwise.} \end{cases} \quad (3.3b)$$

This approach is agnostic as to how samples from the auxiliary distributions are generated – any strategy which produces (approximately) independent samples will be satisfactory. When the assumptions about (1.2) hold, then we would take $U = V_0$ giving (1.25). Corollary 2.1 again applies to this method. Given the choice of U and the mechanism for sampling from its distribution, this algorithm has no free parameters – there is no σ that can be tuned.

In our numerical examples below, where the potentials take the form (1.8), we numerically approximate an inverse function sampler on each degree of freedom, generating i.i.d. samples $\propto e^{-v_0(x_i)}$ for $i = 1, \dots, n$. This approach, of course, will not be practical for general landscapes, and we return to this issue in the discussion.

3.3. Finding smoothed landscapes. A question that remains is what to do when (1.2) fails to hold. One option is to use physical intuition about the problem to identify a potential $U(x)$ that has suitable properties. More systematically, we can use the local entropy approach formulated in [23, 24], or, equivalently the Moreau-Yosida approximation to estimate a smoothed version of $V(x)$.

Given $\gamma > 0$, we define V_γ as

$$V_\gamma(x) = -\beta^{-1} \log \left\{ \int \left(\frac{\beta}{2\pi\gamma} \right)^{\frac{n}{2}} e^{-\frac{\beta|x-y|^2}{2\gamma}} e^{-\beta V(y)} dy \right\}, \quad (3.4)$$

which corresponds to a Gaussian filter of the Boltzmann distribution. The associated gradient is then

$$\nabla V_\gamma(x) = \gamma^{-1} \int_{\mathbb{R}^n} (x - y) \rho_\infty(y | x) dy = \gamma^{-1} (x - \mathbb{E}^{\rho_\infty(\cdot | x)}[Y]). \quad (3.5)$$

The density, $\rho_\infty(y | x)$, can be defined by first introducing

$$U_\gamma(y | x) = V(y) + \frac{1}{2\gamma}|x - y|^2, \quad (3.6)$$

so that

$$\rho_\infty(y | x) \propto e^{-\beta U_\gamma(y | x)}.$$

The potential V_γ could be estimated with the simple Monte Carlo scheme

$$V_\gamma(x) \approx -\beta^{-1} \log \left(\frac{1}{N_s} \sum_{j=1}^{N_s} e^{-\beta V(x + Y^{(j)})} \right), \quad Y^{(j)} \sim N(0, \gamma\beta^{-1}I). \quad (3.7)$$

Likewise, (3.5) could be estimated by integrating the auxiliary diffusion

$$dY_t = -\nabla U_\gamma(Y_t \mid x)dt + \sqrt{2\beta^{-1}}dW_t^y. \quad (3.8)$$

If V is assumed to grow at infinity, then for sufficiently small γ , $U_\gamma(y \mid x)$ will be convex about $y = x$, and (3.8) will converge to equilibrium exponentially fast. The gradient ∇V_γ can then be estimated by Monte Carlo

$$\nabla V_\gamma(x) \approx \frac{1}{N_s} \sum_{j=1}^{N_s} \gamma^{-1}(x - Y^{(j)}), \quad (3.9)$$

where each $Y^{(j)} \sim \rho_\infty(\cdot \mid x)$. Of course, $Y^{(j)}$ must now be appropriately sampled. Running short i.i.d. samples of (3.8) using, for instance, MALA, the algorithm introduces three additional numerical parameters: N_s , the number of samples; δt , the fine scale time step; and $N_{\delta t}$, the number of fine scale time steps. We must also specify initial conditions.

With estimates of V_γ and ∇V_γ , we then use them as the smoothed landscapes in our coupled independence sampler scheme or Modified MALA scheme.

4. Numerical experiments

In this section we demonstrate, numerically, how roughness can impede MALA, and how RWM, along with the methods discussed in Section 3, can resist the roughness. In these computational examples, we use the Metropolis rule, (2.2a). By Remark 2.2, it is sufficient to study MSD_n without including the variance factor in (1.21) or (1.23).

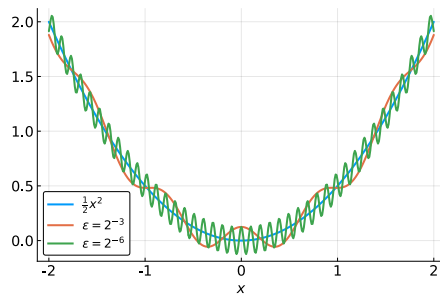


FIG. 4.1. Harmonic potential (4.1) with an additive rough term (color online).

4.1. Rough harmonic potential. As a first example, we consider a potential of the type (1.8), with

$$v_\epsilon(x_i) = \frac{1}{2}x_i^2 + \frac{1}{8}\cos(x_i/\epsilon). \quad (4.1)$$

The potential is depicted in Figure 4.1 at various values of ϵ .

We explore this potential by varying both ϵ and the dimension n . A priori, we do not know the optimal value of σ for each ϵ and n pair, for each algorithm. Thus, we examine a range of σ values, running long trajectories for each, and then empirically estimate the one with the maximum MSD_n . See Figure 4.2 for examples of such a computations. In this way we are able to compare performance across methods. In these experiments inverse temperature is set to $\beta = 5$ and 10^8 iterations are performed in each run. The starting point for these runs is $x_0 = (0, 0, \dots, 0)^T \in \mathbb{R}^n$. For the independence

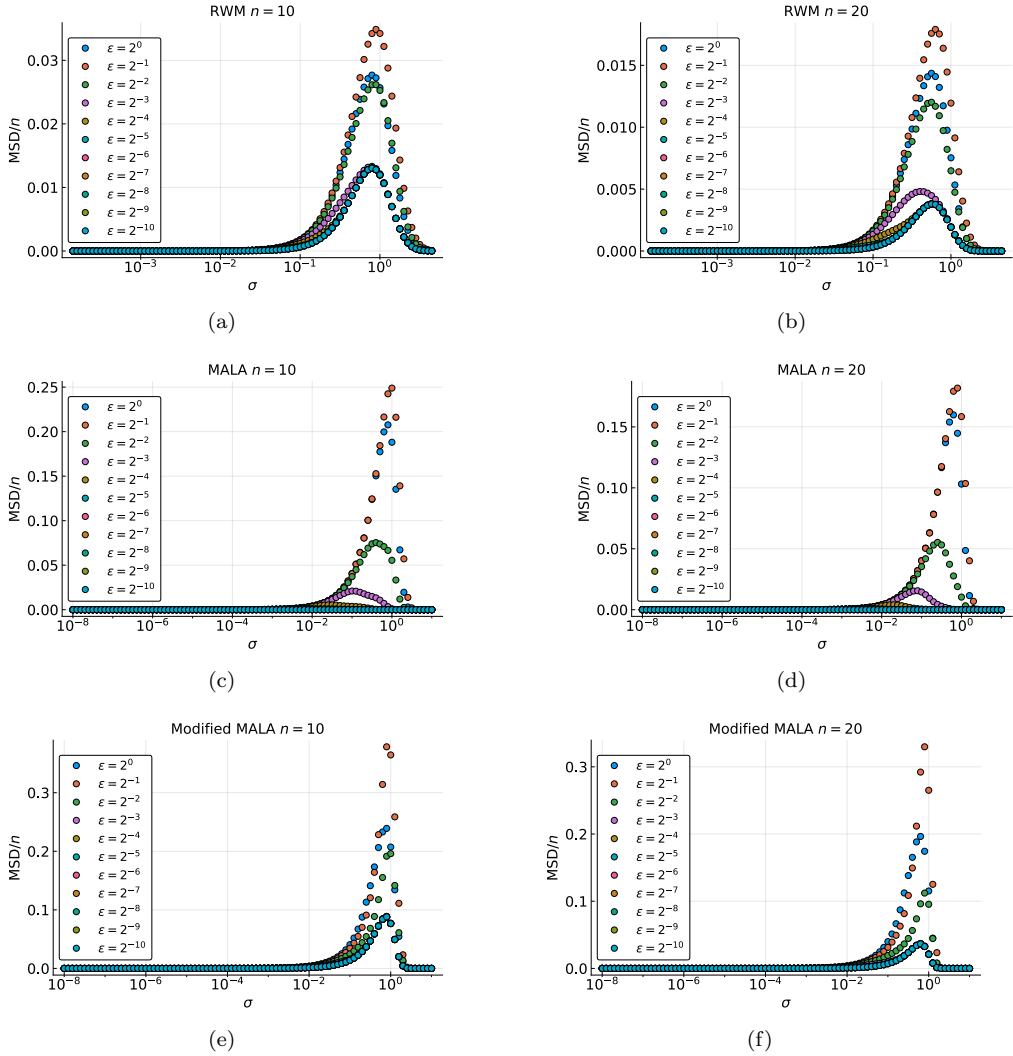


FIG. 4.2. Example computations of the MSD for (4.1) for a variety of σ and ϵ values. The empirical maximum, for each (σ, ϵ) pair is interpreted to be the optimal choice. Note that the peak of MSD_n for MALA falls off rapidly with ϵ (color online).

sampler, we combine numerical quadrature and interpolation to approximate the inverse cumulative distribution function of the distribution $e^{-\beta v_0(x_i)}$. This allows us to perform inverse function sampling. Consequently, there are no free parameters to tune for this strategy.

Having estimated the optimal σ , we can then compare performance between the algorithms, as a function of ϵ and n . This is shown in Figure 4.3. This indicates that $\sigma \propto \epsilon$ for MALA, provided n is large enough. In this scaling, Theorem 2.3 does not apply. However, the numerical experiments reveal that mixing, as measured by MSD, is impeded because of the size of σ . We thus conclude from the numerical experiment that MALA fails to satisfy (1.23), implying that it will also fail to be globally robust, even with optimal σ . Also note that the mean acceptance rates across a range of ϵ and

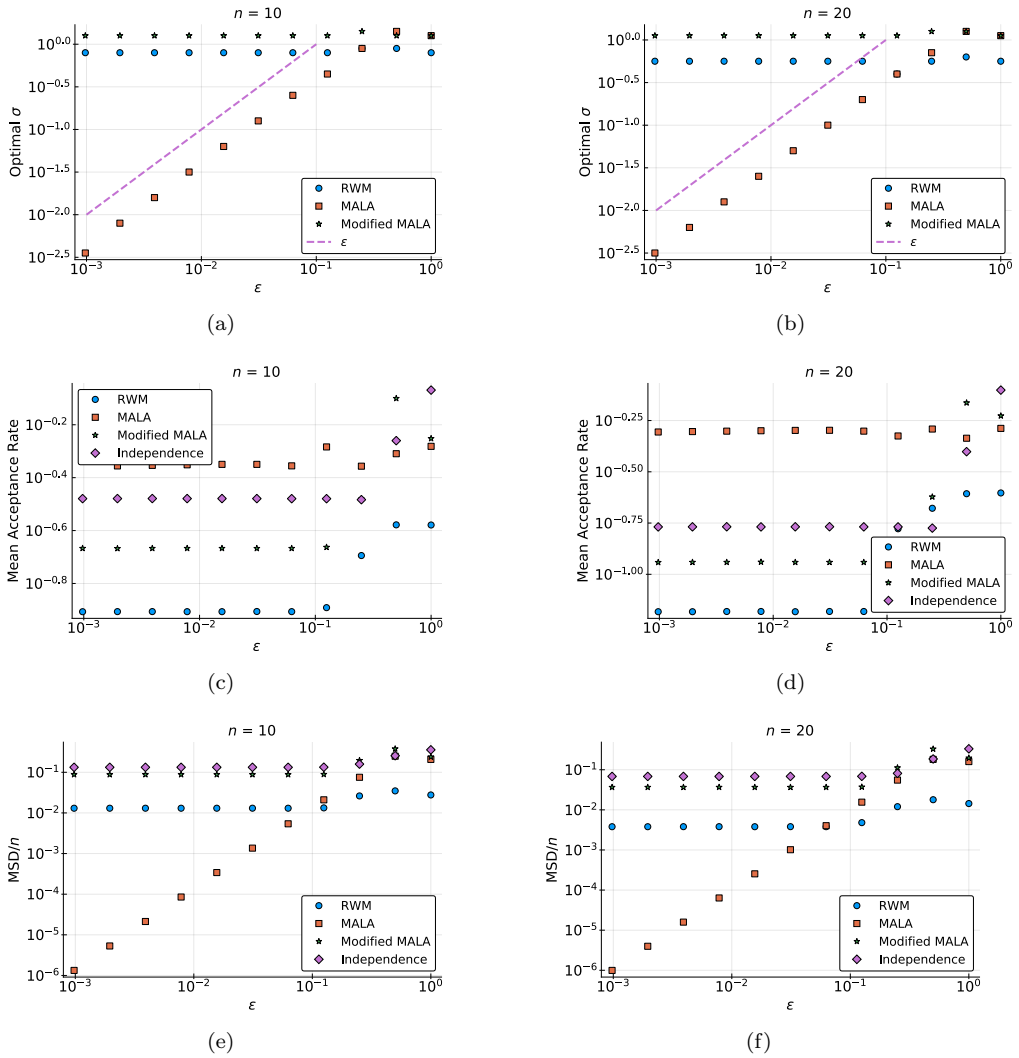


FIG. 4.3. Comparison of performance for (4.1) in different dimensions for a range of ϵ . Observe that the optimal $\sigma \propto \epsilon$, as predicted. Also note that while MALA maintains a comparatively higher acceptance rate, because its step size is shrinking, the mixing rate, as measured by the MSD vanishes with ϵ (color online).

n deviate from the idealized $n \rightarrow \infty$ RWM and MALA values.

Since RWM, Modified MALA, and the Independence sampler are all globally robust to $\epsilon \rightarrow 0$ in the sense of condition (1.20), we examined amplification in performance at different ϵ and n , as measured by

$$\text{Performance Amplification} = \frac{\text{Optimal MSD for Alternative Method}}{\text{Optimal MSD for RWM}}. \quad (4.2)$$

This is shown in Table 4.1. The alternative methods always beat RWM, and, there is a greater improvement in higher dimension, though the performance improvement saturates as $\epsilon \rightarrow 0$. The independence sampler method typically outperforms the Modified MALA method.

(a) Modified MALA Sampler

ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	2.43	6.73	9.59	15.71	4.38
2^{-6}	2.43	6.74	9.61	15.88	15.16
2^{-7}	2.43	6.74	9.62	15.57	20.93
2^{-8}	2.43	6.73	9.60	15.40	25.07
2^{-9}	2.43	6.73	9.62	15.72	20.36
2^{-10}	2.43	6.74	9.61	15.48	22.51

(b) Independence Sampler

ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	2.20	10.15	17.91	40.36	12.59
2^{-6}	2.20	10.17	17.94	39.56	57.39
2^{-7}	2.20	10.17	17.88	40.07	70.85
2^{-8}	2.20	10.17	17.91	39.26	67.93
2^{-9}	2.20	10.17	17.94	39.98	67.61
2^{-10}	2.20	10.17	17.92	39.12	73.55

TABLE 4.1. Ratios of the MSD at optimal σ for Modified MALA and Independence sampler to the RWM for the rough harmonic landscape, (4.1). As $\epsilon \rightarrow 0$, the amplification in performance saturates, but increases with dimension. The Independence sampler method appears to give better performance in moderate to high dimensions.

(a) Modified MALA Sampler

ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	0.31	4.90	6.93	3.84	1.22
2^{-6}	0.30	4.89	6.98	11.45	1.80
2^{-7}	0.29	4.89	6.96	5.22	1.01
2^{-8}	0.29	4.89	6.97	11.69	15.40
2^{-9}	0.29	4.89	7.03	1.00	1.00
2^{-10}	0.29	4.90	6.94	11.64	17.30

(b) Independence Sampler

ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	6.30	330.89	589.73	441.69	62.95
2^{-6}	6.31	330.06	594.21	1317.64	262.12
2^{-7}	6.31	330.42	592.44	434.52	0.00
2^{-8}	6.32	330.32	594.40	1343.84	2150.46
2^{-9}	6.32	330.32	598.89	0.00	0.00
2^{-10}	6.31	330.81	592.33	1335.96	2287.34

TABLE 4.2. Ratios of the MSD at optimal σ for Modified MALA and Independence sampler to the RWM for the rough double-well landscape, (4.3). As $\epsilon \rightarrow 0$, the amplification in performance saturates, but increases with dimension. The Independence sampler method appears to give better performance in moderate to high dimensions. In contrast to the harmonic potential, both the Modified MALA sampler and the Independence sampler show signs of decreased performance in high dimensions.

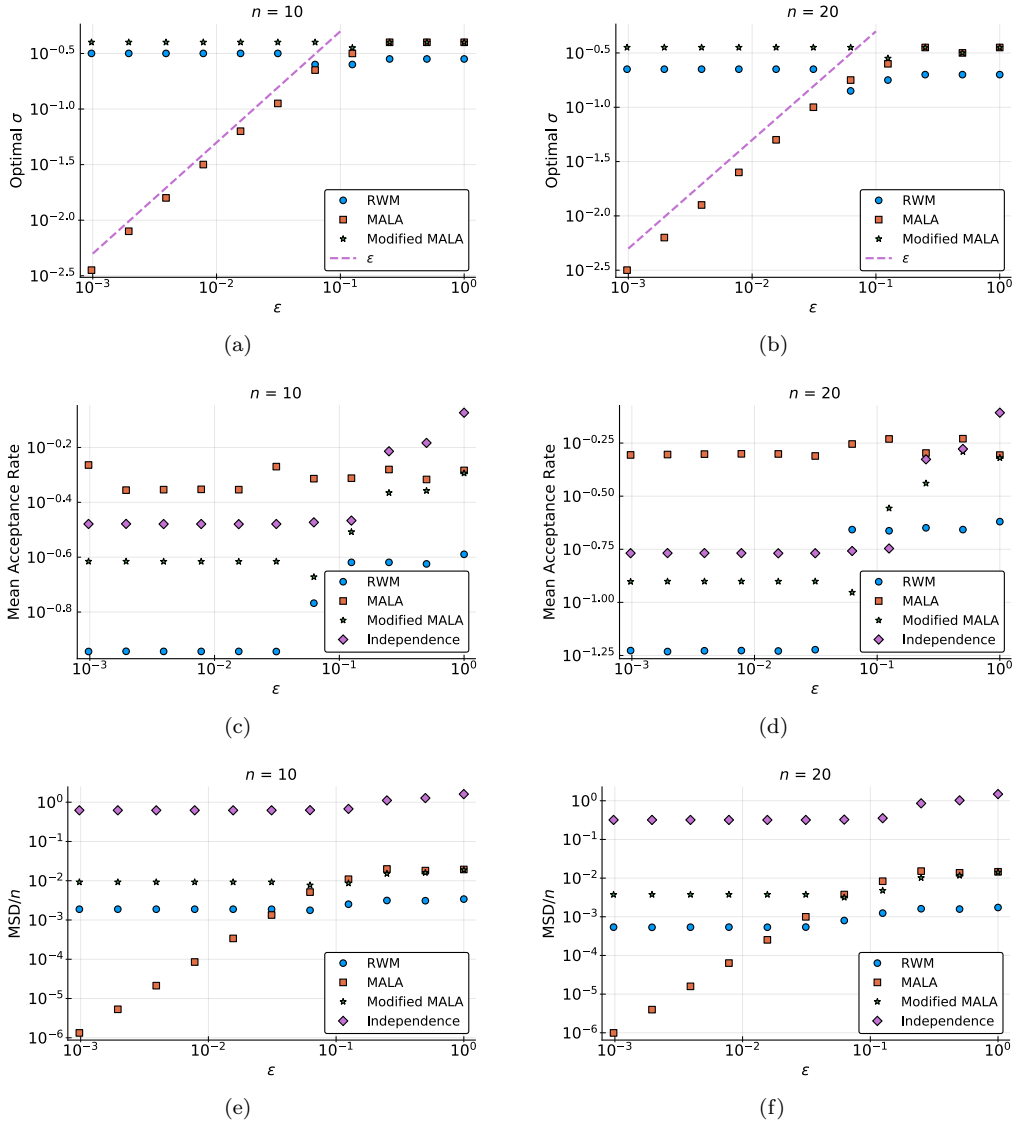


FIG. 4.4. Comparison of performance for (4.3) in different dimensions for a range of ϵ . Observe that the optimal $\sigma \propto \epsilon$, as predicted. Also note that while MALA maintains a comparatively higher acceptance rate, because its step size is shrinking, the mixing rate, as measured by the MSD vanishes with ϵ (color online).

4.2. Rough double-well potential. We repeat the experiment from Section 4.1 for a more challenging problem of sampling from the distribution given by the potential

$$v_\epsilon(x_i) = (x_i^2 - 1)^2 + \frac{1}{8} \cos(x_i/\epsilon), \quad (4.3)$$

with parameters otherwise the same as in the harmonic case. The starting point for these runs is $x_0 = (-1, -1, \dots, -1)^T \in \mathbb{R}^n$. Results here are shown in Figure 4.4 and Table 4.2. These are similar to the experiments for the harmonic potential, but there are some serious differences and some cases where there is little or no performance gain.

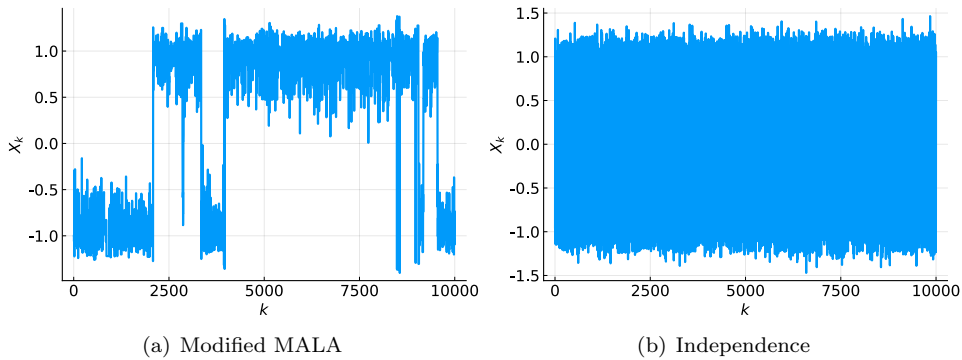


FIG. 4.5. Sample paths corresponding to the energy landscapes in Figure 1.1(b). These were generated using MALA with $\sigma = 1$. Compare with standard MALA sampling in Figure 1.2(b) (color online).

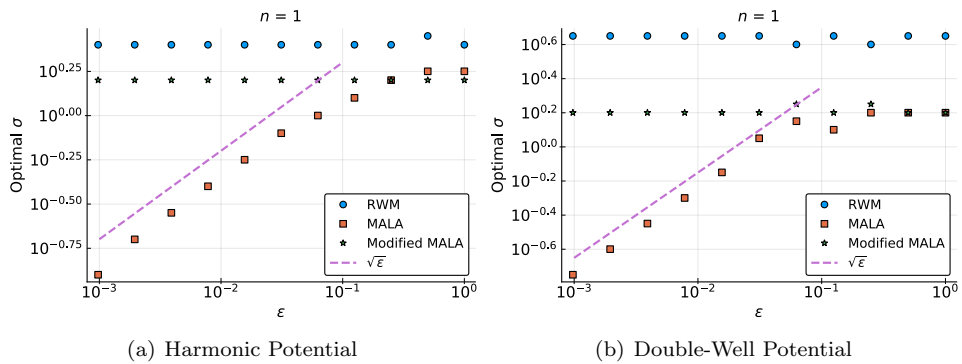


FIG. 4.6. Optimal σ values for the rough harmonic and double-well potentials in $n = 1$. In contrast to results in higher dimensions (see Figures 4.3 and 4.4), for MALA, $\sigma \propto \sqrt{\epsilon}$ for $n = 1$ (color online).

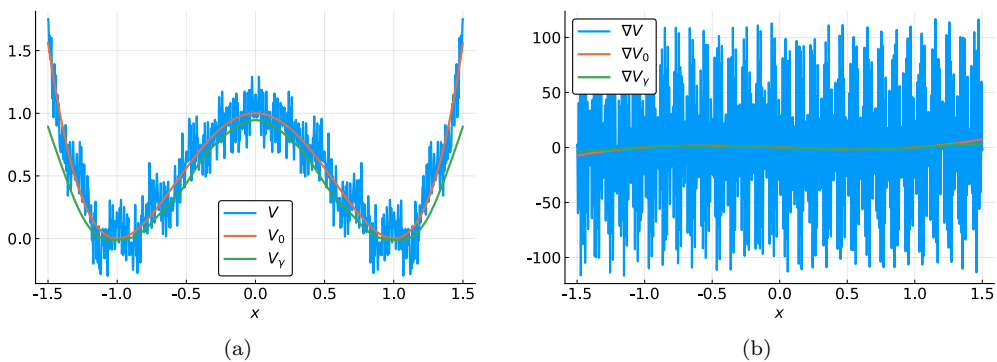


FIG. 4.7. A particular realization of the energy landscape given by (4.4), along with its leading order, long range component $V_0(x) = (x^2 - 1)^2$, and the local entropy approximation, V_γ , computed using (3.4) with $\beta = 5$ and $\gamma = 0.05$ (color online).

(a) Modified MALA Sampler					
ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	0.31	4.91	6.95	3.88	1.22
2^{-6}	0.30	4.89	6.99	11.49	1.73
2^{-7}	0.29	4.89	6.98	11.54	7.48
2^{-8}	0.29	4.89	6.98	11.43	16.84
2^{-9}	0.29	4.88	6.99	11.67	17.11
2^{-10}	0.29	4.88	6.96	11.60	15.67

(b) Independence Sampler					
ϵ	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
2^{-5}	6.31	331.09	591.19	439.82	72.42
2^{-6}	6.31	330.56	594.66	1339.65	191.19
2^{-7}	6.32	330.24	594.54	1337.39	976.85
2^{-8}	6.32	330.13	594.19	1311.59	2562.82
2^{-9}	6.32	329.43	595.04	1355.55	2566.64
2^{-10}	6.31	329.40	593.38	1325.89	2199.09

TABLE 4.3. Ratios of the MSD at optimal σ for Modified MALA and Independence Sampler to the RWM for the rough double-well landscape, (4.3) with $x_0 = (0, 0, \dots, 0)^T \in \mathbb{R}^n$. Compare with Table 4.2.

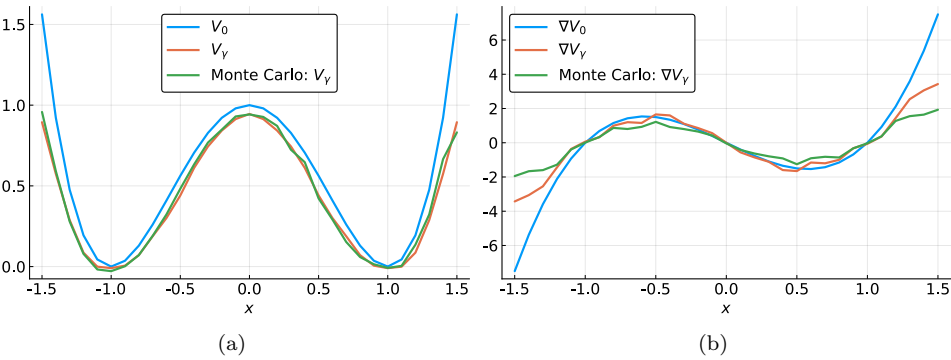


FIG. 4.8. Monte Carlo estimates of V_γ and ∇V_γ for the landscapes in Figure 4.7 (color online).

First, in comparison to the harmonic problem, the performance amplification of the independence sampler is significantly larger in the double-well problem in dimensions $n = 10$ and $n = 20$. We attribute this to the ability of the independence sampler to switch between the left and right super basins, a behavior that is entirely absent from the harmonic problem and one which RWM and the Modified MALA method are incapable of reproducing.

The second difference is that both the Modified MALA sampler and the independence sampler have significant performance issues in dimensions $n = 50$ and $n = 100$ values of ϵ including 2^{-7} and 2^{-9} . Examining the actual trajectories, in the case of the Modified MALA, the optimal σ is so small that the contribution of $\frac{\sigma^2}{2} \nabla V_0$ is negligible.

Consequently, it is trending towards RWM resulting in no performance amplification. This is why the ratio is unity. For the independence sampler, the trajectory is stagnant. This is partially a function of the choice of initial conditions for the trajectories. If, we instead start the independence sampler at $x_0 = (0, 0, \dots, 0)^T \in \mathbb{R}^n$, instead of Table 4.2 (b), we obtain Table 4.3. This variation requires a more detailed discussion which we provide in Section 5. Despite these problems in the higher dimension, there is clearly improvement upon RWM in more modest dimension over a broad range of ϵ .

In Figure 4.5, we repeat the experiment from Figure 1.2 with $n = 1$, but for the Modified MALA and the Independence sampler schemes. Computing at $\sigma = 1$, we see better mixing than in Figure 1.2, for the same rough energy landscape.

4.3. Results in dimension one. We briefly consider the behavior in dimension one for the harmonic potential and the double well. For both energy landscapes, as shown in Figure 4.6, the optimal $\sigma \propto \sqrt{\epsilon}$, not $\sigma \propto \epsilon$. Thus, while the performance of MALA will also degrade in $n = 1$, a different scaling appears here. This is consistent with the harmonic problem analyzed in Section 2.4.

4.4. Local entropy approximations. We briefly consider the possibility of using local entropy, introduced in Section 3.3 with (3.4) and (3.5). This approach may be of use in problems where no straightforward scale separation, of the type found in (1.2), is present in the energy landscape. As motivation we consider the potential in dimension one

$$V(x) = \underbrace{(x^2 - 1)^2}_{V_0(x)} + \sum_{j=1}^M c_j \cos(k_j x), \quad (4.4)$$

where the c_j and k_j are random, from a particular distribution. Indeed, taking $M = 10$ and $c_j \sim U(-.1, .1)$, $\log(k_j) \sim U(10^1, 10^3)$, we obtain the landscape shown in Figure 4.7, along with the leading contribution, V_0 , and V_γ obtained through a numerical quadrature at $\beta = 5$ and $\gamma = 0.05$. Clearly, the local entropy approximation eliminates the fine scale roughness found in the original potential.

In general, Monte Carlo approximations of V_γ and ∇V_γ will be needed, as a quadrature will be impractical in high dimensions. An example of this is shown in Figure 4.8. In Figure 4.8(a), we compare V_0 , V_γ and a Monte Carlo estimate of V_γ computed using (3.7) with $N_s = 10^2$.

In Figure 4.8(b), we compare ∇V_0 , ∇V_γ , and a Monte Carlo estimate of V_γ , computed using (3.9). In this latter figure we take $N_s = 10^4$, and each sample is obtained by taking $N_{\delta t} = 4$ time steps with $\delta t = 1$ in a variant of MALA that exactly linearly integrates the Ornstein-Uhlenbeck component of (3.8). Obviously, there are many options for how Monte Carlo estimates of V_γ and ∇V_γ can be obtained. We will return to this in the discussion.

5. Discussion

We have examined a class of rough energy landscapes where the performance of MALA can be driven to zero at a fixed dimension. When σ is inadequately scaled with ϵ , Proposition 2.2 and Theorem 2.3 reveal that MALA will fail to be globally robust, (1.20). Even if σ is optimally scaled, according to the empirical estimates, the numerical simulations indicate that it also suffers.

There are several outstanding questions on MALA that merit investigation. First, it would be desirable to develop a rigorous understanding of why the optimal σ scaling is $\sigma \propto \sqrt{\epsilon}$ in dimension one, while it is $\sigma \propto \epsilon$ for n sufficiently large. Next, there is an

analysis of why the spectral gap collapses at the optimal scaling as $\epsilon \rightarrow 0$; recall that our result in Theorem 2.3 does not apply to either scaling. Another, related, challenge is how to quantify the size of the small gradient set, (2.36), outside of the separable case, as $\epsilon \rightarrow 0$, or, alternatively, how to avoid analyzing the small set. Separately, there is the question of how σ should be scaled for MALA before it has reached stationarity.

We have also demonstrated that RWM along with the Modified MALA and the Independence sampler are insensitive to the roughness of the landscape and are globally robust. While modified MALA and the Independence sampler require a smoothed energy landscape for proposals, RWM is a viable option without additional information. Indeed, Corollary 2.1 tells us that for rough energy landscapes like (1.2), with roughness bounded uniformly in ϵ , if the proposal of the MCMC scheme is ϵ -independent, then the performance will be insensitive to ϵ . Corollary 2.3 indicates that other methods that have a weak ϵ -dependence, such as the tamed MALA, will also exhibit robustness to the roughness.

As the results in Section 4 show, there still appear to be challenges in high dimensions. The degradation of the Independence sampler shown in Table 4.2 is a consequence of an unusually difficult starting point, $x_0 = (-1, -1, \dots, -1)^T \in \mathbb{R}^n$, as the following calculation shows.

For $\epsilon = 2^{-7}$, $\cos(-1/\epsilon) = -0.69$ and at $\epsilon = 2^{-9}$, $\cos(-1/\epsilon) = -1.00$. Consequently, every single component of the initial guess is close to a global minimum of the energy landscape and almost every proposal will be at a position in state space in which every coordinate has higher energy. Starting at any $X_0 = x_0$,

$$\begin{aligned} \mathbb{E}[|X_1 - X_0|^2]/n &= \mathbb{E}[|X_1 - X_0|^2]/n = \mathbb{E}[|X_1^P - X_0|^2 \wedge e^{R(X_0, X_1^P)}]/n \\ &\leq \sqrt{\mathbb{E}[|X_1^P - X_0|^4]/n^2} \sqrt{\mathbb{E}[1 \wedge e^{R(X_0, X_1^P)}]}. \end{aligned} \quad (5.1)$$

Under our assumptions on $U = V_0$, the fourth moment term is bounded by a constant independent of both n and ϵ . In the above expressions the expectation is over the proposal. The exponent in the acceptance probability is

$$R(X_0, X_1^P) = \underbrace{R_n(x, y)}_{x=X_0, y=X_1^P} = \sum_{i=1}^n r(x_i, y_i) = \sum_{i=1}^n r_i. \quad (5.2)$$

With $x_1 = x_2 = \dots x_n = -1$, the r_i 's are thus i.i.d. Furthermore, since

$$r(x, y) = v_1(x/\epsilon) - v_1(y/\epsilon) \Rightarrow |r(x, y)| \leq \text{osc } v_1 < \infty. \quad (5.3)$$

Suppose $\mu_r \equiv \mathbb{E}[r_i] < 0$; $\mathbb{E}[R_n] = n\mu_r$. Then using Hoeffding's inequality, [29],

$$\begin{aligned} \mathbb{E}[1 \wedge e^{R_n(x, y)}] &= \mathbb{E}[1 \wedge e^{R_n(x, y)} 1_{R_n - n\mu_r < n|\mu_r|/2}] + \mathbb{E}[1 \wedge e^{R_n(x, y)} 1_{R_n - n\mu_r \geq n|\mu_r|/2}] \\ &\leq e^{-n|\mu_r|/2} + \mathbb{P}(R_n - n\mu_r \geq n|\mu_r|/2) \\ &< e^{-n|\mu_r|/2} + \exp\left(-\frac{n}{8} \frac{|\mu_r|^2}{(\text{osc } v_1)^2}\right). \end{aligned} \quad (5.4)$$

Consequently the mean acceptance and the sampling performance will both be driven to zero exponentially fast as $n \rightarrow \infty$ once $\mu_r < 0$. Checking this numerically with $x_1 = -1$, $\beta = 5$, $\epsilon = 2^{-9}$, and $v_0 = \beta \cdot (x^2 - 1)^2$, $v_1 = \beta \cdot \frac{1}{8} \cos(x)$,

$$\begin{aligned} \mu_r &= v_1(x_1/\epsilon) - \mathbb{E}[v_1(y_1/\epsilon)] = v_1(x_1/\epsilon) - \int v_1(y_1/\epsilon) z_0^{-1} e^{-v_0(y_1)} dy_1 \\ &= -0.623. \end{aligned} \quad (5.5)$$

In fact, due to the oscillatory nature of the integral, $\mu_r \approx v_1(x_1/\epsilon)$. With such a value of μ_r , it is straightforward to see, from (5.4), that the performance will rapidly be driven to zero for large enough n . In contrast, in Table 4.3, we see better results when $x_1 = 0$. Repeating the above computation, $\mu_r = 0.625 > 0$, making it resistant to the previous pathology.

This phenomenon does not plague RWM because, in contrast to the independence sampler, RWM has a parameter, σ , that can be tuned to maintain an $O(1)$ mean acceptance probability as $n \rightarrow \infty$. Indeed, the scaling $\sigma^2 = O(n^{-1})$ for RWM, discussed in Section 2.3, ensures this. We conjecture that a Metropolis within Gibbs sampler, by which only a subset of the n coordinates are altered in each step of the sampler, will alleviate this problem in the Independence sampler.

Generating the samples from the smooth distribution for the Independence sampler is also a challenge in the general case. We believe this can be easily accomplished using MALA or HMC. These other samplers should be well behaved on the smooth energy landscape, allowing for the straightforward construction of approximately independent proposals for the rough landscape.

Smoothed energy landscapes might be available through a known decomposition like (1.2). We also conjecture that the optimal choice of approximate landscapes for potentials like (1.2) corresponds to the homogenized energy landscapes discussed in [3–5]. Unfortunately, computing such smoothed energies requires solving an elliptic PDE in a space of the same dimension as the considered state space; a Monte Carlo estimator of the solution may partially overcome this difficulty. Alternatively, physical knowledge of the system may motivate some choice for a surrogate smoothed landscape.

When these options are not available, the local entropy approximation is another possibility. The challenge to using local entropy, which we do not further develop here, is that unless the problem is in a very low dimension, auxiliary sampling algorithms must be formulated and tuned to first estimate V_γ and ∇V_γ . This task would involve determining a sample size, a sampling strategy, and some form of parallelization in order to outperform simpler alternatives like RWM.

Finally, the weakness of MALA in the presence of roughness can be seen as the MCMC manifestation of stiffness. We conjecture that it is a generic problem in gradient based MCMC methods, including HMC. Indeed, the magnitude of ∇V in HMC will constrain the time step of, for instance, the Verlet method used in the Hamiltonian flow subroutine. Thus, the number of force calls per HMC step will tend to increase with roughness degrading the overall performance. This has been partially addressed in [21], where the authors considered potentials of the form (1.15) and rigorously established that HMC suffers from scaling issues.

Acknowledgements. The work of PP was partially supported by DARPA project W911NF-15-2-0122 and GS was supported by US National Science Foundation Grant DMS-1818716. The authors thank UCLA IPAM for hosting them during the beginning of this project. The authors also thank M. Luskin, S. Osher, J. Mattingly, and N. Bou-Rabee for helpful discussions.

Work reported here was run on hardware supported by Drexel’s University Research Computing Facility.

Appendix A. Details of the mean square displacement computation. In this section, we give details of the derivation of (2.22). Differentiating (2.21) with

respect to σ

$$M' = \mathbb{E}_\mu \left[\int |x - y|^2 (F'(R) \partial_\sigma R g(y; x, \sigma) + F(R) \partial_\sigma g(y; x, \sigma)) dy \right] \quad (\text{A.1})$$

$$F'(r) = F(r)(1 - F(r)) \quad (\text{A.2})$$

$$\partial_\sigma R = \frac{\sigma}{4} (|\nabla V(x)|^2 - |\nabla V(y)|^2), \quad (\text{A.3})$$

$$\begin{aligned} \partial_\sigma g &= -\frac{n}{\sigma} g \\ &+ \frac{1}{\sigma^3} \left(|y - x + \frac{\sigma^2}{2} \nabla V(x)|^2 - 2(y - x + \frac{\sigma^2}{2} \nabla V(x)) \cdot \frac{\sigma^2}{2} \nabla V(x) \right) g. \end{aligned} \quad (\text{A.4})$$

Consequently,

$$\begin{aligned} M' &= \frac{\sigma}{4} \mathbb{E} [|x - y|^2 F(1 - F) (|\nabla V(x)|^2 - |\nabla V(y)|^2)] - \frac{n}{\sigma} M \\ &+ \frac{1}{\sigma^3} \mathbb{E} [|x - y|^2 F |y - x + \frac{\sigma^2}{2} \nabla V(x)|^2] \\ &- \frac{2}{\sigma^3} \mathbb{E} [|x - y|^2 F (y - x + \frac{\sigma^2}{2} \nabla V(x)) \cdot \frac{\sigma^2}{2} \nabla V(x)]. \end{aligned} \quad (\text{A.5})$$

Assuming that the optimal σ occurs at a finite value, the first order condition $M' = 0$ will hold. The expression (A.5), at the optimal value, can then be expressed as

$$M = \frac{1}{\sigma^2 n} \mathbb{E} \left[|x - y|^2 F \left(|x - y|^2 - |\frac{\sigma^2}{2} \nabla V(x)|^2 F - |\frac{\sigma^2}{2} \nabla V(y)|^2 (1 - F) \right) \right]. \quad (\text{A.6})$$

This calculation makes use of the identity

$$\begin{aligned} &|y - x + \frac{\sigma^2}{2} \nabla V(x)|^2 - 2(y - x + \frac{\sigma^2}{2} \nabla V(x)) \cdot \frac{\sigma^2}{2} \nabla V(x) \\ &= |y - x|^2 - |\frac{\sigma^2}{2} \nabla V(x)|^2. \end{aligned} \quad (\text{A.7})$$

Since $M = \mathbb{E}[|x - y|^2 F]$, we can re-write (A.6) to get (2.22).

Next, note that for the Barker proposal, $F(r) = (1 + e^{-r})^{-1} = 1 - F(-r)$ so

$$F(y, x) = F(R(y, x)) = 1 - F(-R(y, x)) = 1 - F(R(x, y)) = 1 - F(x, y).$$

Additionally, recall that the Metropolis method satisfies detailed balance. Therefore,

$$\begin{aligned} &\mathbb{E}[|x - y|^2 F(x, y)(1 - F(x, y)) |\nabla V(y)|^2] \\ &= \int |x - y|^2 F(x, y)(1 - F(x, y)) |\nabla V(y)|^2 g(y; x, \sigma) dy \mu(dx) \\ &= \int |x - y|^2 F(y, x)(1 - F(x, y)) |\nabla V(y)|^2 g(x; y, \sigma) dx \mu(dy) \\ &= \int |x - y|^2 F(y, x)^2 |\nabla V(y)|^2 g(x; y, \sigma) dx \mu(dy), \end{aligned}$$

and we conclude

$$\mathbb{E} [|x - y|^2 F (F |\nabla V(x)|^2 + (1 - F) |\nabla V(y)|^2)] = 2 \mathbb{E} [|x - y|^2 F^2 |\nabla V(x)|^2].$$

Using this result in (A.6) gives us (2.22).

Appendix B. Details of computations in dimension one. In this section, we provide a derivation of (2.26). We denote $g_x(x)$ the Gaussian density for $N(0, \epsilon)$ and

$g_y(y|x)$ the Gaussian density for $N((1 - \sigma^2\epsilon^{-1}/2)x, \sigma^2)$. Then

$$\begin{aligned} A_1 &= \mathbb{E}[1 \wedge e^{R(x,y)}] \\ &= \int_{R(x,y)>0} g_x(x)g_y(y|x)dxdy + \int_{R(x,y)\leq 0} e^{R(x,y)}g_x(x)g_y(y|x)dxdy, \end{aligned} \quad (\text{B.1})$$

and we observe that

$$e^{R(x,y)}g_x(x)g_y(y|x) = g_x(y)g_y(x|x). \quad (\text{B.2})$$

Furthermore, we note that the set $R(x, y) > 0$ corresponds to $|x| > |y|$ and $R(x, y) < 0$ corresponds to $|x| < |y|$. We can thus use the symmetry $(x, y) \mapsto (y, x)$ to reduce (B.1) to

$$\begin{aligned} A_1 &= 2 \int_{R(x,y)>0} g_x(x)g_y(y|x)dxdy \\ &= \int_{x=-\infty}^{\infty} \int_{y=-|x|}^{|x|} 2g_x(x)g_y(y|x)dxdy \\ &= \underbrace{\int_{x=-\infty}^0 \int_{y=x}^0 (\dots)dydx}_I + \underbrace{\int_{x=-\infty}^0 \int_{y=0}^{-x} (\dots)dydx}_{II} \\ &\quad + \underbrace{\int_{x=0}^{\infty} \int_{y=-x}^0 (\dots)dydx}_{III} + \underbrace{\int_{x=0}^{\infty} \int_{y=0}^x (\dots)dydx}_{IV}. \end{aligned} \quad (\text{B.3})$$

Since the integrand, $2(y-x)^2g_x(x)g_y(y|x)$ is invariant to $(x, y) \mapsto (-x, -y)$, $I = IV$ and $II = III$. Thus we have

$$A_1 = 4 \int_{x=0}^{\infty} \int_{y=-x}^x g_x(x)g_y(y|x)dxdy. \quad (\text{B.4})$$

Using Mathematica and making the change of variables, $\xi = x/\sqrt{\epsilon}$ and $\delta = \epsilon^{-1}\sigma^2/2$,

$$A_1 = \frac{2}{\pi} \arctan\left((2\delta^{-1})^{3/2}\right). \quad (\text{B.5})$$

Analogously,

$$\begin{aligned} \text{MSD}_1 &= \mathbb{E}[(y-x)^2 1 \wedge e^{R(x,y)}] = \int_{R(x,y)>0} (y-x)^2 g_x(x)g_y(y|x)dxdy \\ &\quad + \int_{R(x,y)\leq 0} (y-x)^2 e^{R(x,y)}g_x(x)g_y(y|x)dxdy. \end{aligned} \quad (\text{B.6})$$

As in the case of the computation of A_1 , we use the symmetry

$$(x-y)^2 e^{R(x,y)}g_x(x)g_y(y|x) = (x-y)^2 g_x(y)g_y(x|x), \quad (\text{B.7})$$

to reduce (B.6) to

$$\text{MSD}_1 = 2 \int_{R(x,y)>0} (y-x)^2 g_x(x)g_y(y|x)dxdy$$

$$= \int_{x=-\infty}^{\infty} \int_{y=-|x|}^{|x|} 2(y-x)^2 g_x(x) g_y(y|x) dx dy. \quad (\text{B.8})$$

This is the split into four integrals, as in (B.3). Since the integrand, $2(y-x)^2 g_x(x) g_y(y|x)$ is invariant to $(x, y) \mapsto (-x, -y)$, $I = IV$ and $II = III$. Thus

$$\text{MSD}_1 = 4 \int_{x=0}^{\infty} \int_{y=-x}^x g_x(x) g_y(y|x) dx dy. \quad (\text{B.9})$$

Using Mathematica with $\xi = x/\sqrt{\epsilon}$ and $\delta = \epsilon^{-1}\sigma^2/2$, $\text{MSD}_n = \epsilon m(\delta)$ we have

$$m(\delta) = \frac{2\delta}{\pi(4 + \delta(-2 + \delta))} \left\{ (8 + \delta^3) \arctan\left(\sqrt{\frac{8}{\delta^3}}\right) - 2\sqrt{2}\delta^{3/2} \right\}. \quad (\text{B.10})$$

This function is plotted in Figure B.1.

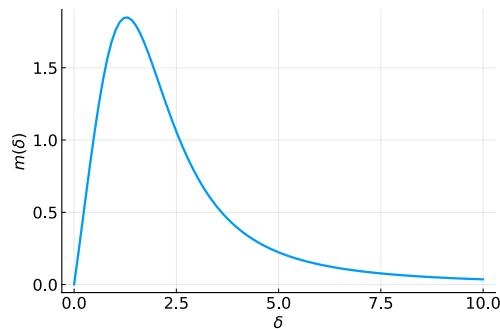


FIG. B.1. Function (B.10). Note that it has a single maximum at $\delta_\star = 1.2779727440041808$. At this value the acceptance rate is 0.70.

REFERENCES

- [1] E. Pollak, A. Auerbach, and P. Talkner, *Observations on rate theory for rugged energy landscapes*, Biophys. J., **95**:4258–4265, 2008. 1
- [2] M. Hu and J.-D. Bao, *Diffusion crossing over a barrier in a random rough metastable potential*, Phys. Rev. E, **97**:062143, 2018. 1
- [3] A. Duncan, S. Kalliadasis, G. Pavliotis, and M. Pradas, *Noise-induced transitions in rugged energy landscapes*, Phys. Rev. E, **94**(3):032107, 2016. 1, 5
- [4] G.B. Arous and H. Owhadi, *Multiscale homogenization with bounded ratios and anomalous slow diffusion*, Commun. Pure Appl. Math., **56**(1):80–113, 2003. 1, 5
- [5] H. Owhadi, *Anomalous slow diffusion from perpetual homogenization*, Ann. Probab., **31**(4):1935–1969, 2003. 1, 5
- [6] J.S. Rosenthal, *Asymptotic variance and convergence rates of nearly-periodic Markov chain Monte Carlo algorithms*, J. Am. Stat. Assoc., **98**(461):169–177, 2003. 1, 3, 1.2, 2.5
- [7] G.O. Roberts, A. Gelman, and W.R. Gilks, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Probab., **7**(1):110–120, 1997. 1.1, 1.1, 1.1
- [8] G.O. Roberts and J.S. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*, J. Royal Stat. Soc. Ser. B, **60**(1):255–268, 1998. 1.1, 1.1, 1.1
- [9] A. Beskos, G. Roberts, and A. Stuart, *Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions*, Ann. Appl. Probab., **19**(3):863–898, 2009. 1.1, 1.1, 1.1
- [10] A. Beskos, G. Roberts, A. Thiery, and N. Pillai, *Asymptotic analysis of the random-walk Metropolis algorithm on ridged densities*, Ann. Probab., **28**(5):2966–3001, 2018. 1.1, 1.1, 1.1

- [11] J. Kuntz, M. Ottobre, and A.M. Stuart, *Diffusion limit for the random walk Metropolis algorithm out of stationarity*, [arXiv:arXiv:1405.4896](#), [1.1](#), [1.1](#)
- [12] J. Kuntz, M. Ottobre, and A.M. Stuart, *Non-stationary phase of the MALA algorithm*, Stoch. PDE: Anal. Comp., **6**(3):446–499, 2018. [1.1](#), [1.1](#)
- [13] J. Kuntz, M. Ottobre, and A.M. Stuart, *Non-stationary phase of the MALA algorithm*, Stoch. PDE: Anal. Comp., **6**:446–499, 2018. [1.1](#), [1.1](#)
- [14] M. Ottobre, N.S. Pillai, F.J. Pinski, and A.M. Stuart, *A function space HMC algorithm with second order Langevin diffusion limit*, Bernoulli, **22**(1):60–106, 2016. [1.1](#)
- [15] N. Bou-Rabee and J. Sanz-Serna, *Geometric integrators and the Hamiltonian Monte Carlo method*, Acta Numer., **27**:113–206, 2018. [1.1](#), [1.1](#), [1.1](#), [iv](#)
- [16] B. Jourdain, T. Lelièvre, and B. Miasojedow, *Optimal scaling for the transient phase of Metropolis–Hastings algorithms: The longtime behavior*, Bernoulli, **20**:1930–1978, 2014. [1.1](#), [1.1](#)
- [17] B. Jourdain, T. Lelièvre, and B. Miasojedow, *Optimal scaling for the transient phase of the random walk Metropolis algorithm: The mean-field limit*, Ann. Appl. Probab., **25**:2263–2300, 2015. [1.1](#), [1.1](#)
- [18] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart, *Optimal tuning of the hybrid Monte Carlo algorithm*, Bernoulli, **19**(5A):1501–1534, 2013. [1.1](#)
- [19] C.C. Potter, R.H. Swendsen, *0.234: The myth of a universal acceptance ratio for Monte Carlo simulations*, Phys. Procedia, **68**:120–124, 2015. [1.1](#)
- [20] J. Yang, G.O. Roberts, and J.S. Rosenthal, *Optimal scaling of random-walk Metropolis algorithms on general target distributions*, Stoch. Process. Appl., **130**(10):6094–6132, 2020. [1.1](#)
- [21] S. Livingstone and G. Zanella, *On the robustness of gradient-based MCMC algorithms*, [arXiv:1908.11812](#), 2019. [1.1](#), [1.1](#), [1.2](#), [2.5](#), [5](#)
- [22] A. Durmus, E. Moulines, and M. Pereyra, *Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau*, SIAM J. Imag. Sci., **11**(1):473–506, 2018. [1.1](#)
- [23] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier, *Deep relaxation: partial differential equations for optimizing deep neural networks*, Res. Math. Sci., **5**(3):30, 2018. [v](#), [3.3](#)
- [24] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, *Entropy-SGD: Biasing gradient descent into wide valleys*, J. Stat. Mech. Theory Exp., **2019**(12):124018, 2019. [v](#), [3.3](#)
- [25] N. Bou-Rabee and E. Vanden-Eijnden, *Pathwise accuracy and ergodicity of Metropolized integrators for SDEs*, Commun. Pure Appl. Math., **63**(5):655–696, 2010. [2.1](#)
- [26] G. Roberts and R. Tweedie, *Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms*, Biometrika, **83**(1):95–110, 1996. [2.1](#)
- [27] M. Hutzenthaler, A. Jentzen, and P.E. Kloeden, *Strong convergence of an explicit numerical method for SDEs with non-globally Lipschitz continuous coefficients*, Ann. Appl. Probab., **22**(4):1611–1641, 2012. [2.1](#)
- [28] G. Zanella, M. Bé dard, and W.S. Kendall, *A Dirichlet form approach to MCMC optimal scaling*, Stoch. Process. Appl., **127**(12):4053–4082, 2017. [2.5](#)
- [29] R. Vershynin, *High-dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, **47**, 2018. [5](#)