Random Coordinate Langevin Monte Carlo

Zhiyan Ding ZDING49@MATH.WISC.EDU and
Qin Li QINLI@MATH.WISC.EDU
Department of Mathematics
University of Wisconsin-Madison
Madison, WI 53706, USA

Jianfeng Lu JIANFENG@MATH.DUKE.EDU
Department of Mathematics, Department of Physics, and Department of Chemistry
Duke University
Durham, NC 27708 USA

Stephen J. Wright SWRIGHT@CS.WISC.EDU
Department of Computer Sciences and Wisconsin Institute for Discovery
University of Wisconsin-Madison
Madison, WI 53706, USA

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

Langevin Monte Carlo (LMC) is a popular Markov chain Monte Carlo sampling method. One drawback is that it requires the computation of the full gradient at each iteration, an expensive operation if the dimension of the problem is high. We propose a new sampling method: Random Coordinate LMC (RC-LMC). At each iteration, a single coordinate is randomly selected to be updated by a multiple of the partial derivative along this direction plus noise, while all other coordinates remain untouched. We investigate the total complexity of RC-LMC and compare it with the classical LMC for log-concave probability distributions. We show that when the gradient of the log-density is Lipschitz, RC-LMC is less expensive than the classical LMC if the log-density is highly skewed for high dimensional problems. Further, when both the gradient and the Hessian of the log-density are Lipschitz, RC-LMC is always cheaper than the classical LMC, by a factor proportional to the square root of the problem dimension. In the latter case, we use an example to demonstrate that our estimate of complexity is sharp with respect to the dimension.

Keywords: Bayesian inference, Random coordinate descent, Langevin Monte Carlo

1. Introduction

Monte Carlo sampling plays an important role in machine learning (Andrieu et al., 2003) and Bayesian statistics. Sampling is essential to such applications as atmospheric science (Fabian, 1981), epidemiology (Li et al., 2020), and petroleum engineering (Nagarajan et al., 2007). It is often needed in data assimilation (Reich, 2011), volume computation (Vempala, 2010) and bandit optimization (Russo et al., 2018).

In many of these applications, the dimension of the problem is extremely high. For example, for weather prediction, one measures the current state temperature and moisture level to infer the flow of the air, before running the Navier–Stokes equations into the near future (Evensen, 2009). In

a global numerical weather prediction model, the degrees of freedom in the air flow can be as high as 10^9. Another example is from epidemiology. One measures the everyday new infection cases to infer the transmission rate in different regions. In a county-level model of the us, the parameter to be inferred has dimension at least $3,141$ (Li et al., 2020).

In this work, we focus on Monte Carlo sampling of log-concave probability distributions on $\mathbb{R}^d$, meaning the probability density can be written as $p(x) \propto e^{-f(x)}$ where $f(x)$ is a convex function. The goal is to generate (approximately) i.i.d. samples according to the target probability distribution with density $p(x)$. Several sampling frameworks have been proposed in the literature, including importance sampling and sequential Monte Carlo (Geweke, 1989; Neal, 2001; Del Moral et al., 2006); ensemble methods (Reich, 2011; Iglesias et al., 2013; Ding et al., 2020a; Liu, 2017; Chewi et al., 2020); Markov chain Monte Carlo (MCMC) (Roberts and Rosenthal, 2004), including Metropolis-Hasting based MCMC (MH-MCMC) (Metropolis et al., 1953; Hastings, 1970; Roberts and Tweedie, 1996); Gibbs samplers (Geman and Geman, 1984; Casella and George, 1992); and Hamiltonian Monte Carlo (Neal, 1993; Duane et al., 1987; Lee and Vempala, 2018). Langevin Monte Carlo (LMC) (Rossky et al., 1978; Parisi, 1981; Roberts and Tweedie, 1996) is a popular MCMC method that has received intense attention in recent years due to the recent progress in the non-asymptotic analysis of its convergence properties (Durmus and Moulines, 2017; Dalalyan, 2017; Dalalyan and Karagulyan, 2019; Durmus et al., 2019).

Denoting by x^m the location of the sample at m-th iteration, LMC updates by setting

$$x^{m+1} = x^m - \nabla f(x^m)h + \sqrt{2h}\xi_d^m \,, \tag{1}$$

where h is the time stepsize and ξ_d^m is drawn i.i.d. from $\mathcal{N}(0, I_d)$, where I_d denotes identity matrix of size $d \times d$. LMC can be viewed as the Euler-Maruyama discretization of the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_{d,t} \,, \tag{2}$$

where $B_{d,t}$ is a d-dimensional Brownian motion. It is well known that under suitable conditions, the distribution of X_t converges exponentially fast to the target distribution (see e.g., (Markowich and Villani, 1999)). Since (1) approximates the SDE (2) with an $\mathcal{O}(h)$ discretization error, the probability distribution of x^m produced by LMC (1) converges exponentially to the target distribution up to a discretization error (Dalalyan and Karagulyan, 2019).

A significant drawback of LMC is that the algorithm requires the evaluation of the full gradient at each iteration. This could be potentially very expensive in most practical problems. Indeed, when the analytical expression of the gradient is not available, each partial derivative component in the gradient needs to be computed separately, either through finite differencing or automatic differentiation (Baydin et al., 2017), so that the total number of evaluations of gradient components can be as many as d times the number of iterations. In the weather prediction and epidemiology problems discussed above, f stands for the map from the parameter space of measured quantities via the underlying partial differential equations (PDEs), and each dimensional partial derivative calls for one forward and one adjoint PDE solve. Thus, $2d$ PDE solves are required in general at each iteration. Another example comes from the study of directed graphs with multiple nodes. Denote the nodes by $\mathcal{N} = \{1, 2, \ldots, d\}$ and directed edges by $\mathcal{E} \subset \{(i, j) : i, j \in \mathcal{N}\}$, and suppose there is a scalar variable x_i associated with each node. When the function f has the form

$f(x) = \sum_{(i,j)\in\mathcal{E}} f_{ij}(x_i, x_j)$, the partial derivative of f with respect to x_i is given by

$$\frac{\partial f}{\partial x_i} = \sum_{j:(i,j)\in\mathcal{E}} \frac{\partial f_{ij}}{\partial x_i}(x_i, x_j) + \sum_{l:(l,i)\in\mathcal{E}} \frac{\partial f_{li}}{\partial x_i}(x_l, x_i)\,.$$

Note that the number of terms in the summations equals the number of edges that touch node i, the expected value of which is about $2/d$ times the total number of edges in the graph. Meanwhile, evaluation of the full gradient would require evaluation of both partial derivatives of each f_{ij} for *all* edges in the graph. Hence, the cost difference between these two operations is a factor of order d.

In this paper, we study how to modify the updating strategies of LMC to reduce the numerical cost, with the focus on reducing dependence on d. In particular, we will develop and analyze a method called Random Coordinate Langevin Monte Carlo (RC-LMC). This idea is inspired by the random coordinate descent (RCD) algorithm from optimization (Nesterov, 2012). RCD is a version of Gradient Descent (GD) in which one coordinate (or a block of coordinates) is selected at random for updating, often by adding a negative multiple of the gradient component corresponding to this coordinate. In optimization, RCD can be significantly cheaper than GD, especially when the objective function is skewed and the problem dimension is high. In RC-LMC, we use the same basic strategy: At iteration m, a single coordinate of x^m is randomly selected for updating according to a certain random selection process, while all others are left unchanged.

Although each iteration of RC-LMC is cheaper than conventional LMC, more iterations are required to achieve the target accuracy, and delicate analysis is required to obtain bounds on the total cost. As in the optimization setting, the savings of RC-LMC by comparison with LMC depend on the structure of the dimensional Lipschitz constants. Under the assumption that there is a factor-of-d difference in per-iteration costs, we compare our results with current results for classical LMC (Dalalyan and Karagulyan, 2019; Durmus et al., 2019) and conclude the following. (Here the notation $\widetilde{O}(\cdot)$ omits possible log terms.)

1. (Theorem 3) When the gradient of f is Lipschitz but the Hessian is not, RC-LMC requires $\widetilde{O}(d^2/\epsilon^2)$ iterations to attain an ϵ-accurate solution. This order is the same as that for LMC. However, the constant depends on the Lipschitz structure of the gradient ∇f. In particular in Remark 5, we compare the numerical cost of RC-LMC and LMC, and show that if f is skewed and the dimension of the problem is high, RC-LMC outperforms. We furthermore show that the optimal numerical cost in this setting is achieved when the probability of choosing the i-th direction is proportional to the i-th directional Lipschitz constant.

2. (Theorem 6) When both the gradient and the Hessian of f are Lipschitz, RC-LMC requires $\widetilde{O}(d^{3/2}/\epsilon)$ iterations to achieve ϵ-accuracy. This cost is strictly smaller than the cost for the classical LMC ($\widetilde{O}(d^2/\epsilon)$ in this setting), meaning RC-LMC saves a factor of at least $d^{1/2}$ regardless of the stiffness structure of f, as discussed in Remark 8.

3. (Proposition 9) We show by means of an example that the $\widetilde{O}(d^{3/2}/\epsilon)$ complexity bound for RC-LMC is sharp when both the gradient and the Hessian of f are Lipschitz.

We make three additional remarks. (a) Throughout the paper we assume that one element of the gradient is available at an expected cost of approximately $1/d$ of the cost of the full gradient evaluation. Although this property is intuitive, and often holds in many situations (such as the graph-based example presented above), it does not hold for all problems (Wright, 2015). (b) Besides replacing

gradient evaluation by coordinate algorithms, one might also improve the dimension dependence of LMC by utilizing a more rapidly convergent method for the underlying SDEs than (2). One such possibility is to use underdamped Langevin dynamics, see e.g., (Rossky et al., 1978; Dalalyan and Riou-Durand, 2018; Cheng et al., 2018; Eberle et al., 2019; Shen and Lee, 2019; Cao et al., 2019), which can also be combined with coordinate sampling (Ding et al., 2020b). For clarity of presentation, we focus only on LMC in this work. (c) The cost of full gradient evaluation can be reduced by using stochastic gradient (Welling and Teh, 2011) or MALA-in-Gibbs sampling (Tong et al., 2020). Both of these methods require specific forms of the objective function that we do not consider here.

We summarize the notations and assumptions on f in Section 2, where we also recall theoretical results for the classical LMC method for the later comparison. In Section 3 we present the RC-LMC algorithm. The main theoretical results are presented in Section 4 and numerical experiments are shown in Section 5. Proofs of the main results are deferred to the Appendix.

2. Notations, assumptions and previous results

Here we unify notations and assumptions and summarize and discuss the classical results on LMC. The Wasserstein distance defined here quantifies the distance between two probability distributions:

$$W(\mu,\nu) = \left(\inf_{(X,Y)\in\Gamma(\mu,\nu)} \mathbb{E}|X-Y|^2 \right)^{1/2}, \tag{3}$$

where $\Gamma(\mu,\nu)$ is the set of distribution of $(X,Y) \in \mathbb{R}^{2d}$ whose marginal distributions, for X and Y respectively, are μ and ν. The distributions in $\Gamma(\mu,\nu)$ are called the *couplings* of μ and ν.

Here and in the sequel, we use $|\cdot|$ to denote the Euclidean norm of a vector.

We assume that f is strongly convex, so that p is strongly log-concave. We obtain results under two different assumptions: First, Lipschitz continuity of the gradient of f (Assumption 1) and second, Lipschitz continuity of the Hessian of f (Assumption 2 together with Assumption 1).

Assumption 1 *The function f is twice differentiable, f is μ-strongly convex for some $\mu > 0$ and its gradient ∇f is L-Lipschitz. That is, for all $x, x' \in \mathbb{R}^d$, we have*

$$f(x) - f(x') - \nabla f(x')^\top (x-x') \geq \frac{\mu}{2}|x-x'|^2, \tag{4}$$

and

$$|\nabla f(x) - \nabla f(x')| \leq L|x-x'|. \tag{5}$$

It is an elementary consequence of (4) that

$$(\nabla f(x') - \nabla f(x))^\top (x'-x) \geq \mu|x'-x|^2, \quad \text{for all } x, x' \in \mathbb{R}^d. \tag{6}$$

Since each coordinate direction plays a distinct role in RC-LMC, we distinguish the Lipschitz constants in each such direction. When Assumption 1 holds, partial derivatives in all coordinate directions are also Lipschitz. Denoting them as L_i for each $i = 1, 2, \ldots, d$, we have

$$|\partial_i f(x + t\boldsymbol{e}_i) - \partial_i f(x)| \leq L_i|t| \tag{7}$$

for any $x \in \mathbb{R}^d$ and any $t \in \mathbb{R}$. We further denote $L_{\max} := \max_i L_i$ and define condition numbers:

$$\kappa = L/\mu \geq 1, \quad \kappa_i = L_i/\mu \geq 1, \quad \kappa_{\max} = \max_i \kappa_i. \tag{8}$$

As shown in (Wright, 2015), we have

$$L_i \le L_{\max} \le L \le dL_{\max}, \quad \kappa_i \le \kappa_{\max} \le \kappa \le d\kappa_{\max}\,. \tag{9}$$

These assumptions together imply that the spectrum of the Hessian is bounded above and below for all x, specifically, $\mu I_d \preceq \nabla^2 f(x) \preceq L I_d$, meaning all eigenvalues of $\nabla^2 f$ are in $[\mu, L]$. Furthermore, $[\nabla^2 f(x)]_{ii} \le L_i \le L_{\max}$ for all $x \in \mathbb{R}^d$.

Both upper and lower bounds of L in term of $L_{\max}$ in (9) are tight. If $\nabla^2 f$ is a diagonal matrix, then $L_{\max} = L$, both being the largest diagonal element of $\nabla^2 f$, so that $\kappa_{\max} = \kappa$. (This is the case in which all coordinates are independent of each other, for example, $f = \sum_i \lambda_i x_i^2$.) On the other hand, if $\nabla^2 f = \mathsf{e} \cdot \mathsf{e}^\top$ where $\mathsf{e} \in \mathbb{R}^d$ satisfies $\mathsf{e}_i = 1$ for all i, then $L = dL_{\max}$ and $\kappa = d\kappa_{\max}$. In this situation, f is highly skewed, with $f = (\sum_i x_i)^2/2$.

Assumption 2 *The function f is three times differentiable and $\nabla^2 f$ is H-Lipschitz, that is*

$$\|\nabla^2 f(x) - \nabla^2 f(x')\|_2 \le H|x - x'|, \quad \textit{for all } x, x' \in \mathbb{R}^d. \tag{10}$$

When this assumption holds, we further define H_i to satisfy

$$|\partial_{ii} f(x + t\boldsymbol{e}_i) - \partial_{ii} f(x)| \le H_i |t|\,, \tag{11}$$

for any $i = 1, 2, \ldots, d$, all $x \in \mathbb{R}^d$, and all $t \in \mathbb{R}$, where $\partial_{ii} f$ is $[\nabla^2 f(x)]_{ii}$, the (i, i) diagonal entry of the Hessian matrix $\nabla^2 f$. Known results for classical LMC can be summarized as follows.

Theorem 1 (Durmus et al. (2019, Theorem 9), Dalalyan and Karagulyan (2019, Theorem 5)) *Let q_m be the probability distribution of the m-th iteration of LMC (1), and p be the target distribution. Using the notation $W_m := W(q_m, p)$, we have the following:*

- *Under Assumption 1, setting $h \le 1/L$, we have*

$$W_m \le \exp\left(-\mu h m/2\right) W_0 + 2(\kappa h d)^{1/2}\,; \tag{12}$$

- *Under Assumptions 1 and 2, setting $h < 2/(\mu + L)$, we have*

$$W_m \le \exp\left(-\mu h m\right) W_0 + \frac{Hhd}{2\mu} + 3\kappa^{3/2}\mu^{1/2} h d^{1/2}\,. \tag{13}$$

This theorem yields stopping criteria for the number of iterations M to achieve a user-defined accuracy of ϵ. When the gradient of f is Lipschitz, to achieve ϵ-accuracy, we can require both terms on the right hand side of (12) to be smaller than $\epsilon/2$, which occurs when

$$h = \Theta(\epsilon^2/d\kappa)\,, \quad M = \Theta\left(\frac{1}{\mu h}\log\left(\frac{W_0}{\epsilon}\right)\right) = \Theta\left(\frac{d\kappa}{\mu\epsilon^2}\log\left(\frac{W_0}{\epsilon}\right)\right), \tag{14}$$

leading to a cost of $\widetilde{O}(d^2\kappa/(\mu\epsilon^2))$ evaluations of gradient components (when we assume that each full gradient can be obtained at the cost of d individual components of the gradient). (The notation $A = \Theta(B)$ indicates that $c_{\text{lo}} B \le A \le c_{\text{hi}} B$ for some positive constants c_{lo} and c_{hi}, when B is sufficiently large.) When both the gradient and the Hessian are Lipschitz, we can achieve ϵ-accuracy by requiring all three terms on the right hand side of (13) to be smaller than $\epsilon/3$. Assuming that $d \gg 1$ and all other constants are $O(1)$, the choices

$$h = \Theta(\epsilon\mu/(dH + d^{1/2}L^{3/2}))\,, \quad M = \Theta\left(\frac{dH + d^{1/2}L^{3/2}}{\mu^2\epsilon}\log\left(\frac{W_0}{\epsilon}\right)\right), \tag{15}$$

yield a cost of $\widetilde{O}(d^2 H/(\mu^2\epsilon))$ evaluations of gradient components.

3. Random Coordinate Langevin Monte Carlo

In our proposed Random Coordinate Langevin Monte Carlo (RC-LMC) method, one coordinate is chosen at random and updated at each iteration, while the other components of x are unchanged. Specifically, denoting by r^m the index of the random coordinate chosen at iteration m, we obtain $x_{r^m}^{m+1}$ according to a single-coordinate version of (1) and set $x_i^{m+1} = x_i^m$ for $i \neq r^m$.

The coordinate index r^m can be chosen uniformly from $\{1, 2, \ldots, d\}$; but we will consider more general possibilities. Let ϕ_i be the probability of component i being chosen, we denote the distribution from which r^m is drawn by Φ, where

$$\Phi := \{\phi_1, \phi_2, \ldots, \phi_d\}, \quad \text{where } \phi_i > 0 \text{ for all } i \text{ and } \textstyle\sum_{i=1}^d \phi_i = 1. \tag{16}$$

The stepsize may depend on the choice of coordinate; we denote the stepsizes by $\{h_1, h_2, \ldots, h_d\}$ and assume that they do not change across iterations. In this paper, we choose h_i to be inversely dependent on probabilities ϕ_i, as follows:

$$h_i = \frac{h}{\phi_i}, \quad i = 1, 2, \ldots, d, \tag{17}$$

where $h > 0$ is a parameter that can be viewed as the expected stepsize. In Section 4.2-4.3, we will find the optimal form of Φ under different scenarios. The initial iterate x^0 is drawn from a distribution q_0, which can be any distribution that is easy to draw from (the normal distribution, for example). We present the complete method in Algorithm 1.

Algorithm 1 Random Coordinate Langevin Monte Carlo (RC-LMC)

Input: Coordinate distribution $\Phi := \{\phi_1, \phi_2, \ldots, \phi_d\}$; parameter $h > 0$ and stepsize set $\{h_1, h_2, \ldots, h_d\}$ defined in (16)–(17); M (stop index).

Sample x^0 from an initial distribution q_0

for $m = 0, 1, 2, \ldots M-1$ **do**

1. Draw $r^m \in \{1, \ldots, d\}$ according to probability distribution Φ;
2. Draw ξ^m from $\mathcal{N}(0, 1)$;
3. Update x^{m+1} by

$$x_i^{m+1} = \begin{cases} x_i^m - h_i \partial_i f(x^m) + \sqrt{2h_i}\, \xi^m, & i = r^m \\ x_i^m, & i \neq r^m. \end{cases} \tag{18}$$

end

return x^M

When we compare (18) with the classical LMC (1), we see that only one random coordinate is updated per iteration, meaning:

$$\nabla f(x^m) \to \partial_{r^m} f(x^m) \boldsymbol{e}_{r^m}, \quad \xi_d^m \to \xi^m \boldsymbol{e}_{r^m}$$

where $\boldsymbol{e}_i$ is the unit vector for i-th direction and ξ^m is drawn from $\mathcal{N}(0, 1)$.

We note that, as will be shown below, the optimal choice of Φ depends on the directional Lipschitz constants. These constants, however, are usually not available. One may need to compute it

on-the-fly. Let L_i^0 be the approximate local directional Lipshitz constant at the initial point:

$$L_i^0 = \frac{\partial_i f(x^0 + h\boldsymbol{e}_i) - \partial_i f(x^0)}{h}, \quad \forall\, 1 \le i \le d,$$

and one updates L_i at each step.:

$$L_{r^m}^{m+1} = \max\left\{ L_{r^m}^m, \left| \frac{\partial_{r^m} f(x^{m+1}) - \partial_{r^m} f(x^m)}{x_{r^m}^{m+1} - x_{r^m}^m} \right| \right\}.$$

Φ and $\{h_i\}$ can be updated correspondingly as well. This adaptive version of the algorithm, termed ARC-LMC, is summarized in Appendix A.

As other LMC sampling methods, RC-LMC has a continuous-time counterpart: Define the elapsed time at m-th iteration as

$$T^m := \sum_{n=0}^{m-1} h_{r^n}, \quad \text{with } T^0 := 0, \tag{19}$$

then for $t \in (T^m, T^{m+1}]$, the updating formula (18) in the algorithm can be viewed as the Euler-Maruyama discretization to the following coordinate SDE:

$$\begin{cases} X_{r^m}(t) = X_{r^m}(T^m) - \displaystyle\int_{T^m}^t \partial_{r^m} f(X(s))\,\mathrm{d}s + \sqrt{2}\int_{T^m}^t \mathrm{d}B_s, \\ X_i(t) = X_i(T^m), \quad \forall i \neq r^m, \end{cases} \tag{20}$$

where B_t is a 1-dimensional Brownian motion. Note that each time point, only one coordinate is changed in the SDE (20). We will show in Section 4.1 that this SDE preserves the invariant measure, that is, $X(t) \sim p$ for any $t > 0$ if $X(0) \sim p$. Thus, under very mild conditions, the distribution of $X(t)$ converges to p even if that of $X(0)$ does not follow p. Consequently, x^m, viewed as the discrete counterpart of $X(T^m)$, has a distribution density converging to p exponentially fast in m.

4. Main results

In Section 4.1 we examine the stationary distribution of the underlying SDE (20), which will be used in the convergence proof of RC-LMC. The main results on the convergence of RC-LMC algorithm are presented in Section 4.2 and 4.3 under two different assumptions. Section 4.4 shows that when both Assumption 1 and 2 are satisfied, our bound is tight with respect to d and ϵ. We sketch the proof ideas of the results in Section 4.5, deferring the technical proofs to the appendix.

4.1. Convergence of the coordinate SDE

To study the convergence of the coordinate SDE (20), we use notation $X^m = X(T^m)$ and denote the probability filtration by $\mathcal{F}^m = \{x^0, r^{n\le m}, B_{s\le T^m}\}$. Then $\{X^m\}_{m=0}^\infty$ is a Markov chain and the following proposition (proved in Appendix B) shows its geometric ergodicity.

Proposition 2 *Let $X^m = X(T^m)$ solve the stochastic differential equation (20), then $p(x)$ is a stationary probability density of the Markov chain $\{X^m\}$.*

Though not the main goal of the current paper, we can prove further that the distribution density of X^m converges to the target distribution exponentially fast under mild assumptions; see Proposition 10. Since the samples x^m generated by the algorithm can be viewed as discrete version of X^m, our algorithm could be expected to converge as well, up to a discretization error. We show in the upcoming two subsections that this is indeed the case; we present there the non-asymptotic convergence rate and the complexity of the algorithm.

4.2. Convergence of RC-LMC. Case 1: Lipschitz gradient

Under Assumption 1, we have the following result, proved in Appendix C.

Theorem 3 *Assume that f satisfies Assumption 1, and $h_i = h/\phi_i$ with $h \leq \frac{\mu \min\{\phi_i\}}{8L^2}$.*

Let q_m be the probability distribution of x^m computed in (18), let p be the target distribution, and denote $W_m := W(q_m, p)$. Then we have

$$W_m \leq \exp\left(-\frac{\mu h m}{4}\right) W_0 + \frac{5h^{1/2}}{\mu}\sqrt{\sum_{i=1}^{d}\frac{L_i^2}{\phi_i}}\,. \tag{21}$$

We make a few comments here. (1) The requirement on h is rather weak. When both μ and L are moderate (both $O(1)$ constants), the requirement is essentially $h \lesssim 1/d$. (2) The estimate (21) consists of two terms. The first is an exponentially decaying term and the second comes from the variance of random coordinate selection. If we assume all Lipschitz constants L_i to be $O(1)$, this remainder term is roughly $O(h^{1/2}d)$. (3) The theorem suggests a stopping criterion: Assuming again that $L_i = O(1)$ for all i, we obtain $W_M \leq \epsilon$ by setting $h < \epsilon^2/d^2$ and $M = \widetilde{O}(d^2/\epsilon^2)$. In terms of ϵ and d dependence, this puts M at the same order as (14), as in the classical LMC.

Theorem 3 holds for all choices of $\{\phi_i\}$ satisfying (16). By using the explicit formula (21), we can choose $\{\phi_i\}$ to minimize the right-hand side of the bound. Nesterov (2012) proposed distributions Φ that depend on the dimensional Lipschitz constants L_i, $i = 1, 2, \ldots, d$ from (7). For $\alpha \in \mathbb{R}$, we can let $\phi_i(\alpha) \propto L_i^\alpha$, specifically,

$$\phi_i(\alpha) := \frac{L_i^\alpha}{\sum_j L_j^\alpha}\,, \quad \text{and} \quad \Phi(\alpha) := \{\phi_1(\alpha), \phi_2(\alpha), \ldots, \phi_d(\alpha)\}\,. \tag{22}$$

Note that when $\alpha = 0$, we have the uniform distribution $\phi_i(0) = 1/d$ for all i. When $\alpha > 0$, the directions that with larger Lipschitz constants are chosen with higher probability. Since $h_i = h/\phi_i$, one uses smaller stepsizes for stiffer directions. (On the other hand, when $\alpha < 0$, the directions with larger Lipschitz constants are less likely to be chosen, and the stepsizes are larger in stiffer directions, a situation that is not favorable and should be avoided.)

The following corollary discusses various choices of α and the corresponding computational cost.

Corollary 4 *Under the same conditions as in Theorem 3, with $\phi_i = \phi_i(\alpha)$ defined in (22), the number of iterations M required to attain $W_M \leq \epsilon$ is $M = \Theta\left(\frac{K_{2-\alpha}K_\alpha}{\mu\epsilon^2}\log\left(\frac{W_0}{\epsilon}\right)\right)$, where $K_\alpha = \sum_{i=1}^d \kappa_i^\alpha$. This cost is optimized when $\alpha = 1$, for which choice we have*

$$M = \Theta\left(\frac{(\sum_i \kappa_i)^2}{\mu\epsilon^2}\log\left(\frac{W_0}{\epsilon}\right)\right)\,. \tag{23}$$

See proof in Appendix C. We note that the initial error W_0 enters only through a log term.

Remark 5 *We now compare the numerical cost of RC-LMC and LMC in Case 1.*

• *Optimal sampling: According to Corollary 4, the optimal sampling strategy is achieved when $\alpha = 1$, meaning $\phi_i \propto L_i$. In this case, we compare (23) with (14), adjusting (14) by a factor of d to account for the higher cost per iteration. RC-LMC has more favorable computational cost if*

$$d^2\kappa \geq \left(\sum_i \kappa_i\right)^2 .$$

Considering $\kappa_i \leq \kappa_{\max} \leq \kappa \leq d\kappa_{\max}$, as presented in (9), this is guaranteed if $\kappa \geq \kappa_{\max}^2$. In the regime when $\kappa \sim d\kappa_{\max}$ this holds so long as $d > \kappa_{\max}$, meaning the dimension of the problem is high. In the regime of $\kappa_{\max} \sim \kappa$, RC-LMC still outperforms when κ_i decreases rapidly. One example of such a case is the separable function $f(x) = dx_1^2 + \sum_{i=2}^d x_i^2$ with $d \gg 1$.

• *Uniform sampling: Uniform sampling means $\phi_i = 1/d$ for all i, with $\alpha = 0$ in Corollary 4. This leads to a cost of $\Theta\left(\frac{\sum \kappa_i^2}{\mu\epsilon^2} \log\left(\frac{W_0}{\epsilon}\right)\right)$. Comparing with (14) adjusted by a factor of d, we see that RC-LMC still has a more favorable computational cost if*

$$d^2\kappa \geq \sum_i \kappa_i^2 .$$

As in the optimal case, this happens when f is highly skewed.

Under Assumption 1, Dalalyan and Karagulyan (2019) obtains an estimate of $\widetilde{O}(d^2\kappa^2/(\mu\epsilon^2))$ for the cost of the classical LMC. This is weaker than the optimal cost of LMC obtained in Durmus et al. (2019). It is not clear whether the latter proof technique can be adapted to the coordinate setting to obtain an improved estimate, so we followed the strategy of Dalalyan and Karagulyan (2019). Compared with the latter result for LMC, our estimate for the cost of RC-LMC is always cheaper, since $\kappa^2 \geq \kappa_{\max}^2$.

4.3. Convergence of RC-LMC. Case 2: Lipschitz Hessian

We now assume that Assumption 1 and 2 hold, that is, both the gradient and the Hessian of f are Lipschitz continuous. In this setting, we obtain the following improved convergence estimate. The proof can be found in Appendix D.

Theorem 6 *Assume f satisfies Assumptions 1 and 2 and let $h_i = h/\phi_i$, with $h \leq \frac{\mu \min\{\phi_i\}}{8L^2}$.*

Denoting by $q_m(x)$ the probability density function of x^m computed from (18) and by p the target distribution, and letting $W_m := W(q_m, p)$, we have:

$$W_m \leq \exp\left(-\frac{\mu h m}{4}\right) W_0 + \frac{3h}{\mu}\sqrt{\sum_{i=1}^d \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}} . \tag{24}$$

We see again two terms in the bound, an exponentially decaying term and a variance term. Assuming all Lipschitz constants are $O(1)$, the variance term is of $O(hd^{3/2})$. By comparing with Theorem 3, we see that ϵ error can be achieved with the looser stepsize requirement $h \lesssim \frac{\epsilon}{d^{3/2}}$.

By choosing $\{\phi_i\}$ to optimize the bound in Theorem 6, we obtain the following corollary.

Corollary 7 *Under the same conditions as in Theorem 6, the optimal choice of $\{\phi_i\}$ is to set:*

$$\phi_i = \frac{\left(L_i^3 + H_i^2\right)^{1/3}}{\sum_{i=1}^d \left(L_i^3 + H_i^2\right)^{1/3}}.$$

For this choice, the number of iterations M required to guarantee $W_M \leq \epsilon$ satisfies

$$M = \Theta\left(\frac{\left(\sum_{i=1}^d \left(L_i^3 + H_i^2\right)^{1/3}\right)^{3/2}}{\mu^2 \epsilon} \log\left(\frac{W_0}{\epsilon}\right)\right). \tag{25}$$

If μ, κ_i, and H_i are all $O(1)$ constants, the total cost is $\widetilde{O}(d^{3/2}/\epsilon)$, regardless of the choice of $\{\phi_i\}$.

We emphasize that when f satisfies Assumption 1 and 2, the discretization error that measures the difference between X^m and x^m is smaller than in the case in which only Assumption 1 holds. A similar observation is made in Dalalyan and Karagulyan (2019); it explains a faster non-asymptotic convergence rate in terms of d and ϵ.

Remark 8 *We now compare RC-LMC with LMC in Case 2 using Theorem 6 and Corollary 7. Note the cost of LMC is revealed by (15) adjusted by a factor of d to account for the higher cost per iteration, we see that the cost of RC-LMC with optimal sampling, seen in (25), is always smaller. Furthermore, if one uses uniform sampling by setting $\phi_i = 1/d$, then by (24), the cost is*

$$M = \Theta\left(\frac{d\left(\sum_{i=1}^d \left(L_i^3 + H_i^2\right)\right)^{1/2}}{\mu^2 \epsilon} \log\left(\frac{W_0}{\epsilon}\right)\right).$$

This is still cheaper than LMC. Suppose L and H are all constants of $O(1)$, then the cost of RC-LMC is roughly $\widetilde{O}(d^{3/2}/\epsilon)$, while the classical LMC requires $\widetilde{O}(d^2/\epsilon)$, according to Dalalyan and Riou-Durand (2018)—a factor of $d^{1/2}$ in savings, regardless of the structure of f.

4.4. Tightness of the complexity bound

When both the gradient and the Hessian are Lipschitz, the estimate $\widetilde{O}(d^{3/2}/\epsilon)$ obtained in Corollary 7 is tight. The following proposition is proven in Appendix E.

Proposition 9 *Let $\phi_i = 1/d$ for all i, and set the initial distribution and the target distribution as*

$$q_0(x) = \frac{1}{(4\pi)^{d/2}} \exp(-|x - \mathsf{e}|^2/4), \quad p(x) = \frac{1}{(2\pi)^{d/2}} \exp(-|x|^2/2), \tag{26}$$

where $\mathsf{e} \in \mathbb{R}^d$ satisfies $\mathsf{e}_i = 1$ for all i. Let q_m be the probability distribution of x^m generated by Algorithm 1, and denote $W_m := W(q_m, p)$. We then have that

$$W_m \geq \exp(-2mh)\frac{\sqrt{d}}{3} + \frac{d^{3/2}h}{6}, \quad m \geq 1. \tag{27}$$

In particular, to have $W_M \leq \epsilon$, one needs at least $M = \widetilde{O}(d^{3/2}/\epsilon)$.

4.5. Sketch and discussion of the proof

Our proof of the convergence of the RC-LMC algorithm follows the coupling approach seen in previous works on convergence of LMC-type algorithms (Dalalyan and Riou-Durand, 2018; Dalalyan and Karagulyan, 2019; Cheng et al., 2018).

To explain this claim, we let $\tilde{x}(t)$ satisfy the coordinate SDE (20) with $\tilde{x}^0$ drawn from the target distribution induced by p, and let $\tilde{x}^m = \tilde{x}(T^m)$. According to Proposition 2, the distribution of $\tilde{x}^m$ is given by p for all m, and thus

$$W(q_m, p) \leq \mathbb{E}\left|x^m - \tilde{x}^m\right|^2.$$

For this reason, it suffices to bound the difference: $\Delta^m = x^m - \tilde{x}^m$, and to establish the decay of its L^2 norm in m. This analysis is carried out in Appendix C.

Unlike the existing results in literature, we encounter a special technical difficulty here. In the previous papers studying LMC-type algorithms, the associated SDEs typically enjoy the contraction property. For example, it is shown that two different trajectories that follow SDE (2) with different initial data will contract in time. Therefore, in (Dalalyan and Riou-Durand, 2018; Dalalyan and Karagulyan, 2019), the authors only need to control the discretization errors. This contraction property, however, is not available in our setup. The continuous version of RC-LMC, the coordinate SDE (20) investigated here, cannot be shown to contract, because only one random coordinate is updated per step, preventing us to utilize the convexity property of f.

For this reason, we switch to derive the contraction property on the discrete level directly. This usually requires a much more delicate analysis. In the end, we find that for RC-LMC, the contraction property holds true only *on average*, instead of component-wisely. Namely, if we denote x^m and y^m two solutions to the algorithm (18) with x^0 drawn from q_0 and y^0 drawn from the target p, we can only show:

$$\mathbb{E}_{r^m}\left(|x^{m+1} - y^{m+1}|^2\right) < |x^m - y^m|^2.$$

instead of $|x^{m+1}_{r^m} - y^{m+1}_{r^m}|^2 < |x^m_{r^m} - y^m_{r^m}|^2$. This contraction property on the algorithm level, incorporated with a careful control over the discretization error, allows us to compare x^m with $\tilde{x}^m$, the solution to the underlying SDE. In the process of controlling the error, one needs to analyze each coordinate of Δ^m separately, and combine these estimates by taking expectation of coordinate according to Φ. Due to the difference of directional parameters such as L_i and h_i, a careful analysis is needed to establish the optimal contraction rate.

5. Numerical examples

We provide some numerical results in this section. Since it is extremely challenging to estimate the Wasserstein distance between two distributions in high dimensions, we demonstrate instead the convergence of an estimated expectation for a given observable. Denoting by $\{x^{(i),M}\}_{i=1}^N$ the list of N samples, with each of them computed through Algorithm 1 (or the adaptive version Algorithm 2) independently with M iterations, we define the error as follows:

$$\text{Error}_{M,N} = \left|\frac{1}{N}\sum_{i=1}^{N}\psi(x^{(i),M}) - \mathbb{E}_{p_{\mathrm{X}}}(\psi)\right|, \tag{28}$$

where ψ is a real-valued function, $|\cdot|$ is the absolute value, and $\mathbb{E}_p(\psi)$ is the expectation of ψ under the target distribution p. As $h \to 0$ and $Mh \to \infty$, we have $W_M \to 0$, and $x^{(i),M}$ can

be regarded as approximately sampled from p. According to the central limit theorem, we have $\lim_{h\to 0, Mh\to\infty} \text{Error}_{M,N} = O(1/\sqrt{N})$.

We consider first a problem with the graph structure, discussed in Section 1. Set

$$g(z) = \sum_{1\le i,j\le d} g_{ij}(z_i, z_j)\,, \quad \text{where} \quad g_{ij}(z_i, z_j) = \begin{cases} \dfrac{(z_i - z_j)^2}{4d}, & \text{if } i > j \\ \dfrac{(z_i + z_j)^2}{4d}, & \text{if } i < j \\ \dfrac{(z_i)^2}{2d} + \cos(z_i), & \text{if } i = j \end{cases}.$$

We then set the target distribution function to be

$$p_{\mathsf{X}}(x) \propto \exp\left(-g\left(\mathcal{L}(x)\right)\right)\,,$$

where $\mathcal{L}(x) = (\mathsf{x}\Gamma, x_{11}, \ldots, x_d)$, $\mathsf{x} = (x_1, x_2, \ldots, x_{10})$ is the list of first 10 entries, and $\Gamma = \mathsf{T} + \frac{d}{10} I$. Here I is the 10×10 identity matrix and T is a random matrix whose entries are i.i.d. standard Gaussian random variables. This example has an ill-conditioned f. The Lipschitz constants are $\mathbb{E}(L_{1\le i\le 10}) = \frac{3d^2}{200}$ and $L_{i\ge 10} = \frac{3}{2}$. Thus $\mu = \frac{1}{2}$, and $\mathbb{E}(\kappa_{1\le i\le 10}) = \frac{3d^2}{100}$. When $d \gg 1$, we have $\sum_{j=1}^d \kappa_i^p \ll d\kappa^p$ for $p \ge 1$.

In the simulation we set $d = 100$, $N = 10^5$, and let $\psi(x) = |\mathsf{x}|$. Initially, all particles are drawn from $\mathcal{N}(0.5\mathsf{e}_d, I_d)$, where e_d is a vector in $\mathbb{R}^d$ and all entries equal to 1 and I_d is the $d \times d$ identity matrix. The result is plotted in Figure 1. To run (A)RC-LMC, we use time stepsize $h = 6 \times 10^{-5}$. For comparison we also run LMC, however, due to the cost difference per iteration, there is no standard choice of h for LMC for a fair comparison. Since $d = 100$ in this example, the per-iteration cost of LMC is about 100 times of that of (A)RC-LMC, we first experiment LMC with $h = 6 \times 10^{-3}$ (blue (diamond) line). It is clear that (A)RC-LMC, presented by the (purple dotted) green dashed line achieves a lower error than LMC with the same amount of cost. We then test LMC with different choices of h, hoping to find its best performance. With smaller h, the error plateau is also lower, meaning the error will eventually saturate at a lower value, but the decay rate of error with respect to the cost also decrease, as one can see by comparing the blue (diamond), red (circle), and yellow (plus) lines in Figure 1, all produced by LMC with different values of h. None of them are competitive with (A)RC-LMC regarding the level of error at the same cost.

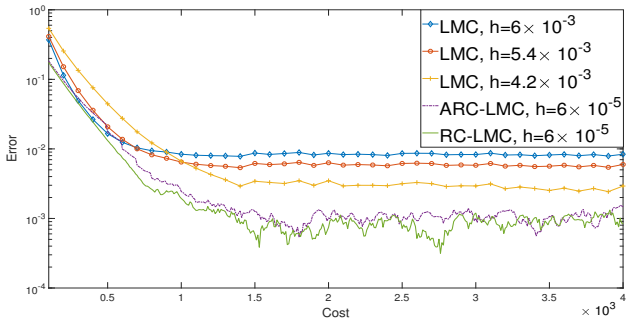

Figure 1: The decay of error with respect to the cost (the number ∂f calculations).

Acknowledgments

Q.L. acknowledges support from Vilas Early Career award. The research of Z.D. and Q.L is supported in part by NSF via grant DMS-1750488. The work of J.L. is supported in part by NSF via grant DMS-2012286 and CCF-1934964. The work of Z.D., Q.L. and S.W. is supported in part by NSF via grant DMS-2023239. S.W. also acknowledges support from NSF Awards 1628384, 1740707, 1839338, and 1934612, and Subcontract 8F-30039 from Argonne National Laboratory.

References

C. Andrieu, N. Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43, 01 2003.

A. Baydin, B. Pearlmutter, A. Radul, and J. Siskind. Automatic differentiation in machine learning: A survey. *J. Mach. Learn. Res.*, 18(1):5595–5637, 01 2017.

Y. Cao, J. Lu, and L. Wang. On explicit L^2-convergence rate estimate for Underdamped Langevin dynamics. *arXiv:1908.04746*, 2019.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3): 167–174, 1992.

X. Cheng, N. Chatterji, P. Bartlett, and M. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 300–323, 07 2018.

S. Chewi, T. Gouic, C. Lu, T. Maunu, and P. Rigollet. SVGD as a kernelized wasserstein gradient flow of the chi-squared divergence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

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

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

A. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic Langevin diffusions. *arXiv*, abs/1807.09382, 2018.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Z. Ding, Q. Li, and J. Lu. Ensemble Kalman inversion for nonlinear problems: Weights, consistency, and variance bounds. *Foundations of Data Science*, 2020a.

Z. Ding, Q. Li, J. Lu, and S. J. Wright. Random coordinate Underdamped Langevin Monte Carlo. *arXiv*, abs/2010.11366, 2020b.

S. Duane, A. Kennedy, B. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195 (2):216 – 222, 1987.

A. Durmus and É. Moulines. Non-asymptotic convergence analysis for the Unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20:73:1–73:46, 2019.

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*, 47(4):1982–2010, 07 2019.

G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag Berlin Heidelberg, 2009.

P. Fabian. Atmospheric sampling. *Advances in Space Research*, 1(11):17 – 27, 1981.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, 11 1984.

J. Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

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

M. Iglesias, K. Law, and A. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 03 2013.

Y. T. Lee and S. Vempala. Convergence rate of Riemannian Hamiltonian Monte Carlo and faster polytope volume computation. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, page 1115–1121, 2018.

R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490): 489–493, 2020.

Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, pages 3115–3123, 2017.

P. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, pages 1–29, 1999.

J. Mattingly, A. Stuart, and D. Higham. Ergodicity for SDEs and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic Processes and their Applications*, 101(2):185 – 232, 2002.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

N. Nagarajan, M. Honarpour, and K. Sampath. Reservoir-fluid sampling and characterization — key to efficient reservoir management. *Journal of Petroleum Technology*, 59, 08 2007.

R. Neal. Probabilistic inference using Markov Chain Monte Carlo methods. *Technical Report CRG-TR-93-1. Dept. of Computer Science, University of Toronto.*, 1993.

R. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 03 2011.

G. Roberts and J. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1, 04 2004.

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

P. Rossky, J. Doll, and H. Friedman. Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

D. Russo, B. Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 01 2018.

R. Shen and Y. T. Lee. The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems*, pages 2100–2111, 2019.

X. Tong, M. Morzfeld, and Y. Marzouk. MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure. *SIAM Journal on Scientific Computing*, 42(3):A1765–A1788, 2020.

S. Vempala. Recent progress and open problems in algorithmic convex geometry. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 8, pages 42–64, 2010.

M. Welling and Y. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming, Series B*, 151(1):3–34, 2015.

Appendix A. The adaptive version of RC-LMC

The adaptive version of RC-LMC updates the approximation to the dimensional Lipschitz constants. The configuration of Φ and the stepsizes h_i are thus changed accordingly. We summarize it in Algorithm 2.

Algorithm 2 Adaptive Random Coordinate Langevin Monte Carlo (ARC-LMC)

Input: Stepsize $h > 0$; M (stop index).

Sample x^0 from an initial distribution q_0.

Define L_i^0 for $1 \le i \le d$:

$$L_i^0 = \left| \frac{\partial_i f(x^0 + h\mathbf{e}_i) - \partial_i f(x^0)}{h} \right| .$$

Define ϕ_i^0 and h_i^0 for $1 \le i \le d$:

$$\phi_i^0 = \frac{L_i^0}{\sum_{j=1}^d L_j^0}, \quad h_i^0 = \frac{h}{\phi_i^0} .$$

for $m = 0, 1, 2, \dots M-1$ **do**

1. Draw $r^m \in \{1, \dots, d\}$ according to probability distribution $\Phi^m = \{\phi_1^m, \phi_2^m, \cdots, \phi_d^m\}$;
2. Draw ξ^m from $\mathcal{N}(0,1)$;
3. Update x^{m+1} by

$$x_i^{m+1} = \begin{cases} x_i^m - h_i^m \partial_i f(x^m) + \sqrt{2h_i^m}\, \xi^m, & i = r^m \\ x_i^m, & i \neq r^m \end{cases} . \tag{29}$$

4. Update L^{m+1} for $1 \le i \le d$

$$L_i^{m+1} = \begin{cases} \max\left\{ L_{r^m}^m, \left| \frac{\partial_{r^m} f(x^{m+1}) - \partial_{r^m} f(x^m)}{x_{r^m}^{m+1} - x_{r^m}^m} \right| \right\}, & i = r^m \\ L_i^m, & i \neq r^m \end{cases} . \tag{30}$$

5. Update ϕ_i^{m+1} and h_i^{m+1} for $1 \le i \le d$

$$\phi_i^{m+1} = \frac{L_i^{m+1}}{\sum_i^d L_i^{m+1}}, \quad h_i^{m+1} = \frac{h}{\phi_i^{m+1}} .$$

end

return x^M

Appendix B. Proof of Proposition 2 and further discussions

We recall the SDE (20):

$$\begin{cases} X_{r^m}(t) = X_{r^m}(T^m) - \displaystyle\int_{T^m}^{t} \partial_{r^m} f(X(s))\,\mathrm{d}s + \sqrt{2}\int_{T^m}^{t} \mathrm{d}B_s\,, \\ X_i(t) = X_i(T^m)\,, \quad \forall i \neq r^m\,, \end{cases} \tag{31}$$

where r^m is randomly selected from $1, \ldots, d$.

Proof [Proof of Proposition 2]

To prove Proposition 2, we assume the distribution of X^m is Π and we need to prove:

For any choice of r^m, the conditional distribution of X^{m+1} is also Π.

Without loss of generality, we consider $r^m = 1$. Under this condition, we have the following.

- The distribution of $X_{2\le j\le d}(t)$ between $[T^m, T^{m+1}]$ is preserved.
- For fixed $z_2, z_3, \ldots, z_d$, the stationary density of SDE

$$\mathrm{d}z = -\partial_1 f(z, z_2, z_3, \ldots, z_d)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_s\,, \tag{32}$$

  is $\frac{\exp(-f(z,z_2,\ldots,z_d)}{\int \exp(-f(z,z_2,\ldots,z_d)\,\mathrm{d}z}$. This implies that the conditional distribution of $X_1(t)$ with fixed $X_{2\le j\le d}(t)$ is also preserved.

Combining these two points, we find that under condition $r^m = 1$, the conditional distribution of X^{m+1} is Π.

Since the statement holds true for all other values of r^m as well, and thus Π is the stationary distribution and Proposition 2 holds. ■

We note that in Proposition 2 we do not claim the convergence of the Markov chain. With more technical derivation, we can in fact show:

Proposition 10 *Denote Π^m the probability distribution of X^m and Π be the probability distribution induced by $p(x)$, then under the conditions of Proposition 2, and assuming the second moment of Π^0 is finite, then there are constants $R > 0$ and $r > 1$, independent of m, such that for any $m \ge 0$, we have:*

$$d_{TV}\left(\Pi^m, \Pi\right)\,\mathrm{d}x \le Rr^{-m}\,. \tag{33}$$

The proof for this result is rather technical, and is only remotely related to the core of the current paper, and thus is omitted from this paper. We mostly follow the theory presented in (Mattingly et al., 2002), and justify the Lyapunov condition together with minorization condition called for there.

We also note that according to Mattingly et al. (2002), the constants R and r do not depend on m, but their dependence on other parameters such as h, d, and L is hard to trace. This contrasts with the results in Dalalyan and Karagulyan (2019) for the classical Langevin dynamics, where clear dependence on all parameters can be established. The new complication comes mainly from the complicated coordinate selection process. We should reiterate, however, this convergence result only concerns the SDE, the continuous version of RC-LMC, so the explicit dependence of the convergence rate is not sought after.

Appendix C. Proof of Theorem 3

The proof of this theorem requires us to design a reference solution to explicitly bound $W(q_m, p)$. Let $\tilde{x}^0$ be a random vector drawn from target distribution induced by p, so that $W_2^2(q_0, p) = \mathbb{E}|x^0 - \tilde{x}^0|^2$. We then require $\tilde{x}$ to solve the following SDE: for $t \in (T^m, T^{m+1}]$, with T^m defined in (19):

$$\begin{cases} \tilde{x}_{r^m}(t) = \tilde{x}_{r^m}(T^m) - \displaystyle\int_{T^m}^{t} \partial_{r^m} f(\tilde{x}(s))\,\mathrm{d}s + \sqrt{2}\int_{T^m}^{t} \mathrm{d}B_s\,, \\ \tilde{x}_i(t) = \tilde{x}_i(T^m), \quad i \neq r^m\,. \end{cases} \tag{34}$$

If we use the same Brownian motion as in (29), we have

$$\tilde{x}^{m+1} = \tilde{x}^m + \left[-\int_{T^m}^{T^{m+1}} \partial_{r^m} f(\tilde{x}(s))\,\mathrm{d}s + \sqrt{2h_{r^m}}\xi^m \right] \boldsymbol{e}_{r^m}\,, \tag{35}$$

where $\boldsymbol{e}_{r^m}$ is the unit vector in r^m direction. According to Proposition 2, the distribution of $\tilde{x}(t)$ is preserved to be p for all t. Therefore, by the definition $W_m = W(q_m, p)$, we have

$$W_m^2 \le \mathbb{E}|\Delta^m|^2 = \mathbb{E}|x^m - \tilde{x}^m|^2\,,$$

where

$$\Delta^m := \tilde{x}^m - x^m\,. \tag{36}$$

Bounding W_m now amounts to evaluating $\mathbb{E}|\Delta^m|^2$. Under Assumption 1, we have the following result.

Proposition 11 *Suppose the assumptions of Theorem 3 are satisfied and let* $\{x^m\}$, $\{\tilde{x}^m\}$, *and* $\{\Delta^m\}$ *be defined in* (29), (34), *and* (36), *respectively. Then, we have*

$$\mathbb{E}|\Delta^{m+1}|^2 \le \left(1 - \frac{h\mu}{2}\right)\mathbb{E}|\Delta^m|^2 + \frac{10h^2}{\mu}\sum_{i=1}^{d}\frac{L_i^2}{\phi_i}\,. \tag{37}$$

The proof of this proposition appears in Appendix C.1. The proof for Theorem 3 is now immediate.

Proof [Proof of Theorem 3] By iterating (37), we obtain

$$\mathbb{E}|\Delta^m|^2 \le \left(1 - \frac{h\mu}{2}\right)^m \mathbb{E}|\Delta^0|^2 + \frac{20h}{\mu^2}\sum_{i=1}^{d}\frac{L_i^2}{\phi_i}\,,$$

and since $h\mu/2 \in (0, 1)$, we have

$$\mathbb{E}|\Delta^m|^2 \le \exp\left(-\frac{\mu h m}{2}\right)\mathbb{E}|\Delta^0|^2 + \frac{20h}{\mu^2}\sum_{i=1}^{d}\frac{L_i^2}{\phi_i}\,. \tag{38}$$

By construction, we have $W^2(q_0, p) = \mathbb{E}|\Delta^0|^2$ and $W^2(q_m, p) \le \mathbb{E}|\Delta^m|^2$. By taking the square root of both sides and using $a^2 \le b^2 + c^2 \Rightarrow a \le b + c$ for any nonnegative a, b, and c, we arrive at (21). ■

The proof for Corollary 4 is also obvious.

Proof [Proof of Corollary 4] To ensure that $W_m \leq \epsilon$, we set the two terms on the right hand side of (21) to be smaller than $\epsilon/2$, which implies that

$$h = O\left(\frac{\mu^2\epsilon^2}{100\sum_{i=1}^d \frac{L_i^2}{\phi_i(\alpha)}}\right) \quad \text{and} \quad m \geq \frac{4}{\mu h}\log\left(\frac{2W_0}{\epsilon}\right). \tag{39}$$

By using the definition of $\phi_i(\alpha)$ according to (22), we obtain

$$\sum_{i=1}^d \frac{L_i^2}{\phi_i(\alpha)} = \left(\sum_{i=1}^d \frac{L_i^2}{L_i^\alpha}\right)\left(\sum_{j=1}^d L_j^\alpha\right) = \mu^2 K_{2-\alpha}K_\alpha\,,$$

which implies that $m = \widetilde{O}\left(\left(K_{2-\alpha}K_\alpha\right)/(\mu\epsilon^2)\right)$. Furthermore, $\alpha = 1$ gives the optimal cost, because:

$$K_{2-\alpha}K_\alpha = \left(\sum \kappa_i^\alpha\right)\left(\sum \kappa_i^{2-\alpha}\right) \geq \left(\sum_i \kappa_i\right)^2 = K_1^2\,,$$

due to Hölder's inequality. ■

C.1. Proof of Proposition 11

To prove Proposition 11, we first present the following lemma.

Lemma 12 *Under the conditions of Proposition 11, for $m \geq 0$ and $i = 1, 2, \ldots, d$, we have*

$$\begin{aligned}\mathbb{E}|\Delta_i^{m+1}|^2 \leq \left(1 + h\mu + \frac{h^2\mu^2}{\phi_i}\right)\mathbb{E}|\Delta_i^m|^2 - 2h\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right)\right] \\ + \frac{3h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right|^2 + \left(\frac{2h^3L_i^3}{\mu\phi_i^2} + \frac{8h^2L_i^2}{\mu\phi_i}\right).\end{aligned} \tag{40}$$

Proof In the m-th time step, we have

$$\mathbb{P}(r^m = i) = \phi_i, \quad \mathbb{P}(r^m \neq i) = 1 - \phi_i\,,$$

so that

$$\begin{aligned}\mathbb{E}|\Delta_i^{m+1}|^2 &= \phi_i\mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m = i\right) + (1-\phi_i)\,\mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m \neq i\right) \\ &= \phi_i\mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m = i\right) + (1-\phi_i)\,\mathbb{E}\,|\Delta_i^m|^2\,.\end{aligned} \tag{41}$$

We now analyze the first term on the right hand side under condition $r^m = i$. By definition of Δ_i^{m+1}, we have

$$\begin{aligned}
\Delta_i^{m+1} &= \Delta_i^m + (\tilde{x}_i^{m+1} - \tilde{x}_i^m) - (x_i^{m+1} - x_i^m) \\
&= \Delta_i^m + \left(- \int_{T^m}^{T^m + h_i} \partial_i f(\tilde{x}(s))\, \mathrm{d}s + \sqrt{2h_i}\xi_m \right) - \left(- \int_{T^m}^{T^m + h_i} \partial_i f(x^m)\, \mathrm{d}s + \sqrt{2h_i}\xi_m \right) \\
&= \Delta_i^m - \int_{T^m}^{T^m + h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(x^m) \right) \mathrm{d}s \\
&= \Delta_i^m - \int_{T^m}^{T^m + h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m) + \partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right) \mathrm{d}s \\
&= \Delta_i^m - h_i \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right) - \int_{T^m}^{T^m + h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m) \right) \mathrm{d}s \\
&= \Delta_i^m - h_i \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right) - V^m \,,
\end{aligned} \tag{42}$$

where we have defined

$$V^m := \int_{T^m}^{T^m + h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m) \right) \mathrm{d}s \,. \tag{43}$$

By Young's inequality, we have

$$\begin{aligned}
&\mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m = i \right) \\
&= \mathbb{E}\left(|\Delta_i^{m+1} + V^m - V^m|^2 \mid r^m = i \right) \\
&\le (1+a)\, \mathbb{E}\left(|\Delta_i^{m+1} + V^m|^2 \mid r^m = i \right) + \left(1 + \frac{1}{a} \right) \mathbb{E}\left(|V^m|^2 \mid r^m = i \right) \,,
\end{aligned} \tag{44}$$

where $a > 0$ is a parameter to be specified later.

For the first term on the right hand side of (44), we have

$$\begin{aligned}
&\mathbb{E}\left(|\Delta_i^{m+1} + V^m|^2 \mid r^m = i \right) \\
&= \mathbb{E} |\Delta_i^m - h_i \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right)|^2 \\
&= \mathbb{E} |\Delta_i^m|^2 - 2h_i \mathbb{E}\left[\Delta_i^m \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right) \right] + h_i^2 \mathbb{E} \left| \partial_i f(\tilde{x}^m) - \partial_i f(x^m) \right|^2 \,.
\end{aligned} \tag{45}$$

Note that the second term will essentially become the second line in (40), and the third term will become the third line in (40) (upon the proper choice of a). For very small h, this term is negligible.

For the second term on the right-hand side of (44), we recall the definition (43) and obtain

$$\begin{aligned}
\mathbb{E}\left(|V^m|^2\middle|r^m=i\right) &\overset{\text{(I)}}{\leq} h_i\int_{T^m}^{T^m+h_i}\mathbb{E}\left(|\partial_i f(\tilde{x}(s))-\partial_i f(\tilde{x}^m)|^2\middle|r^m=i\right)\mathrm{d}s\\
&\overset{\text{(II)}}{\leq} h_i L_i^2\int_{T^m}^{T^m+h_i}\mathbb{E}\left(|\tilde{x}(s)-\tilde{x}^m|^2\middle|r^m=i\right)\mathrm{d}s\\
&= h_i L_i^2\int_{T^m}^{T^m+h_i}\mathbb{E}\left(\left|\int_{T^m}^{s}\partial_i f(\tilde{x}(t))\,\mathrm{d}t+\sqrt{2}(B_s-B_{T^m})\right|^2\middle|r^m=i\right)\mathrm{d}s\\
&\overset{\text{(III)}}{\leq} 2h_i^2L_i^2\int_{T^m}^{T^m+h_i}\int_{T^m}^{s}\mathbb{E}\left(|\partial_i f(\tilde{x}(t))|^2\middle|r^m=i\right)\mathrm{d}t\,\mathrm{d}s\\
&\quad+4h_i^2L_i^2\int_{T^m}^{T^m+h_i}\mathbb{E}|\xi^m|^2\,\mathrm{d}s\\
&\overset{\text{(IV)}}{=} h_i^4L_i^2\mathbb{E}\left(|\partial_i f(\tilde{x}^m)|^2\right)+4h_i^3L_i^2\\
&\overset{\text{(V)}}{=} h_i^4L_i^2\mathbb{E}_p|\partial_i f|^2+4h_i^3L_i^2\overset{\text{(VI)}}{\leq} h_i^4L_i^3+4h_i^3L_i^2\,,
\end{aligned}\tag{46}$$

where (II) comes from L-Lipschitz condition (7), (I) and (III) come from the use of Young's inequality and Jensen's inequality when we move the $|\cdot|^2$ from outside to inside of the integral, and (IV) and (V) hold true because $\tilde{x}(t)\sim p$ for all t. In (VI) we use $\mathbb{E}_p|\partial_i f|^2\leq L_i$ using (Dalalyan and Karagulyan, 2019, Lemma 3).

By substituting (45) and (46) into the right hand side of (44), we obtain

$$\begin{aligned}
&\mathbb{E}\left(|\Delta_i^{m+1}|^2\mid r^m=i\right)\\
&\leq(1+a)\mathbb{E}|\Delta_i^m|^2-2h_i(1+a)\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]\\
&\quad+h_i^2(1+a)\mathbb{E}\left|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right|^2+\left(1+\frac{1}{a}\right)\left(h_i^4L_i^3+4h_i^3L_i^2\right).
\end{aligned}\tag{47}$$

By substituting (47) into (41), we have

$$\begin{aligned}
\mathbb{E}|\Delta_i^{m+1}|^2&\leq(1+a\phi_i)\,\mathbb{E}|\Delta_i^m|^2-2(1+a)h\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]\\
&\quad+\frac{(1+a)h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right|^2+\left(1+\frac{1}{a}\right)\left(\frac{h^4L_i^3}{\phi_i^3}+\frac{4h^3L_i^2}{\phi_i^2}\right),
\end{aligned}\tag{48}$$

where we have used $h_i\phi_i=h$.

Now, we need to choose a value of $a>0$ appropriate to establish (40). By comparing the two formulas, we see the need to set

$$a\phi_i=h\mu\quad\Rightarrow\quad a=h_i\mu=\frac{h\mu}{\phi_i}\leq 1\,.$$

since $h\leq\min\{\phi_i\}/\mu$. It follows that $1+\frac{1}{a}\leq\frac{2\phi_i}{h\mu}$. By substituting into (48), we obtain

$$\begin{aligned}
\mathbb{E}|\Delta_i^{m+1}|^2&\leq(1+h\mu)\,\mathbb{E}|\Delta_i^m|^2-2h\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]\\
&\quad-\frac{2h^2\mu}{\phi_i}\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]+\frac{2h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right|^2\\
&\quad+\left(\frac{2h^3L_i^3}{\mu\phi_i^2}+\frac{8h^2L_i^2}{\mu\phi_i}\right).
\end{aligned}\tag{49}$$

We conclude the lemma by using the following Cauchy-Schwartz inequality to control the third term on the right hand side of this expression:

$$-\frac{2h^2\mu}{\phi_i}\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right] \leq \frac{h^2\mu^2}{\phi_i}\mathbb{E}|\Delta_i^m|^2+\frac{h^2}{\phi_i}\mathbb{E}|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)|^2\,.$$

■

Proposition 11 is obtained by simply summing all components in the lemma.

Proof [Proof of Proposion 11] Noting

$$\mathbb{E}|\Delta^{m+1}|^2=\sum_{i=1}^d \mathbb{E}|\Delta_i^{m+1}|^2\,,$$

we bound the right hand side by (40) and get

$$\begin{aligned}\mathbb{E}|\Delta^{m+1}|^2 \leq \left(1+h\mu+\frac{h^2\mu^2}{\min\{\phi_i\}}\right)\mathbb{E}|\Delta^m|^2-2h\mathbb{E}\left\langle\Delta^m,\nabla f(\tilde{x}^m)-\nabla f(x^m)\right\rangle \\ +\frac{3h^2}{\min\{\phi_i\}}\mathbb{E}\left|\nabla f(\tilde{x}^m)-\nabla f(x^m)\right|^2+\left(\frac{2h^3}{\mu}\sum_{i=1}^d\frac{L_i^3}{\phi_i^2}+\frac{8h^2}{\mu}\sum_{i=1}^d\frac{L_i^2}{\phi_i}\right)\,.\end{aligned} \tag{50}$$

The second and third terms on the right-hand side can be bounded in terms of $\mathbb{E}|\Delta^m|^2$:

- By convexity, we have

$$\mathbb{E}\left\langle\Delta^m,\nabla f(\tilde{x}^m)-\nabla f(x^m)\right\rangle \geq \mu\mathbb{E}|\Delta^m|^2\,. \tag{51}$$

- As the gradient is L-Lipschitz, we have

$$\mathbb{E}\left|\nabla f(\tilde{x}^m)-\nabla f(x^m)\right|^2 \leq L^2\mathbb{E}|\Delta^m|^2\,. \tag{52}$$

By substituting (51) and (52) into (50) and using $\mu \leq L$, we obtain

$$\mathbb{E}|\Delta^{m+1}|^2 \leq \left(1-h\mu+\frac{4h^2L^2}{\min\{\phi_i\}}\right)\mathbb{E}|\Delta^m|^2+\left(\frac{2h^3}{\mu}\sum_{i=1}^d\frac{L_i^3}{\phi_i^2}+\frac{8h^2}{\mu}\sum_{i=1}^d\frac{L_i^2}{\phi_i}\right)\,. \tag{53}$$

If we take h sufficiently small, the coefficient in front of $\mathbb{E}|\Delta^m|^2$ is strictly smaller than 1, ensuring the decay of the error. Indeed, by setting $h \leq \frac{\mu\min\{\phi_i\}}{8L^2}$, we have

$$\frac{4h^2L^2}{\min\{\phi_i\}} \leq \frac{h\mu}{2}, \quad \text{and} \quad \frac{hL_i}{\phi_i} \leq \frac{\mu}{8L} \leq 1\,,$$

which leads to the iteration formula (37). ■

Appendix D. Proof of Theorem 6

Theorem 6 is based on the following proposition.

Proposition 13 *Suppose the assumptions of Theorem 6 and let* $\{x^m\}$, $\{\tilde{x}^m\}$, *and* $\{\Delta_m\}$ *be defined as in* (29), (34), *and* (36), *respectively. Then we have*

$$\mathbb{E}|\Delta^{m+1}|^2 \leq \left(1 - \frac{h\mu}{2}\right) \mathbb{E}|\Delta^m|^2 + \frac{4h^3}{\mu} \sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}. \tag{54}$$

We prove this result in Appendix D.1. The proof of the theorem is now immediate.

Proof [Proof of Theorem 6] Use (54) iteratively, we have

$$\begin{aligned}\mathbb{E}|\Delta^{m+1}|^2 &\leq \left(1 - \frac{h\mu}{2}\right)^m \mathbb{E}|\Delta^0|^2 + \frac{8h^2}{\mu^2} \sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2} \\ &\leq \exp\left(-\frac{\mu h m}{2}\right) \mathbb{E}|\Delta^0|^2 + \frac{8h^2}{\mu^2} \sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}.\end{aligned}$$

Using $W^2(q_0, p) = \mathbb{E}|\Delta^0|^2$ and $W^2(q_m, p) \leq \mathbb{E}|\Delta^m|^2$, we take the square root on both sides, we obtain (24). ■

The proof of Corollary 7 is also immediate.

Proof [Proof of Corollary 7] Use (24), to ensure $W_m \leq \epsilon$, we set two terms on the right hand side of (24) to be smaller than $\epsilon/2$, which implies that

$$h = O\left(\frac{\epsilon\mu}{\sqrt{\sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}}}\right), \quad m \geq \frac{4}{\mu h} \log\left(\frac{2W_0}{\epsilon}\right). \tag{55}$$

To find optimal choice of ϕ_i, we need to minimize

$$\sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}$$

under constraint $\sum_i^d \phi_i = 1$ and $\phi_i > 0$. Introducing a Lagrange multiplier $\lambda \in \mathbb{R}$, define the Lagrangian function as follows:

$$F(\phi_1, \phi_2, \ldots, \phi_d, \lambda) = \sum_{i=1}^{d} \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2} + \lambda\left(\sum_{i=1}^{d} \phi_i - 1\right).$$

By setting $\partial F/\partial\phi_i = 0$ for all i, and substituting into the constraint $\sum_i^d \phi_i = 1$ to find the appropriate value of λ, we find that the optimal $(\phi_1, \phi_2, \ldots, \phi_d)$ satisfies

$$\phi_i = \frac{\left(L_i^3 + H_i^2\right)^{1/3}}{\sum_{i=1}^{d} \left(L_i^3 + H_i^2\right)^{1/3}}, \quad i = 1, 2, \ldots, d.$$

By substituting into (55), we obtain (25). ■

D.1. Proof of Proposition 13

The strategy of the proof for this proposition is almost identical to that of the previous section. The reference solution $\tilde{x}$ is defined as in (34). We will use the following lemma:

Lemma 14 *Under the conditions of Proposition 13, for $m \geq 0$ and $i = 1, 2, \ldots, d$, we have*

$$\begin{aligned}\mathbb{E}|\Delta_i^{m+1}|^2 \leq & \left(1 + h\mu + \frac{h^2\mu^2}{\phi_i}\right) \mathbb{E}|\Delta_i^m|^2 - 2h\mathbb{E}\left[\Delta_i^m \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right)\right] \\ & + \frac{3h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right|^2 + \frac{4h^3\left(L_i^3 + H_i^2\right)}{\phi_i^2 \mu}\,. \end{aligned} \tag{56}$$

Proof In the m-th time step, we have

$$\mathbb{P}(r^m = i) = \phi_i, \quad \mathbb{P}(r^m \neq i) = 1 - \phi_i\,,$$

meaning that

$$\begin{aligned}\mathbb{E}|\Delta_i^{m+1}|^2 &= \phi_i \mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m = i\right) + (1-\phi_i)\,\mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m \neq i\right) \\ &= \phi_i \mathbb{E}\left(|\Delta_i^{m+1}|^2 \mid r^m = i\right) + (1-\phi_i)\,\mathbb{E}\,|\Delta_i^m|^2\,. \end{aligned} \tag{57}$$

To bound the first term in (41) we use the definition of Δ_i^{m+1}. Under the condition $r^m = i$, we have, with the same derivation as in (42):

$$\begin{aligned}\Delta_i^{m+1} &= \Delta_i^m - h_i\left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right) - \int_{T^m}^{T^m + h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m)\right) \mathrm{d}s \\ &= \Delta_i^m - h_i\left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right) - V^m\,, \end{aligned} \tag{58}$$

where we denoted $V^m = \int_{T^m}^{T^m+h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m)\right) \mathrm{d}s$.

However, different from (46), since f has higher regularity, we can find a tighter bound for the integral. Denote

$$U^m = \int_{T^m}^{T^m+h_i} \left(\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m) - \sqrt{2}\int_{T^m}^{s} \partial_{ii} f(\tilde{x}(z))\,\mathrm{d}B_z\right) \mathrm{d}s \tag{59}$$

and

$$\Phi^m = \sqrt{2}\int_{T^m}^{T^m+h_i} \int_{T^m}^{s} \partial_{ii} f(\tilde{x}(z))\,\mathrm{d}B_z\,\mathrm{d}s\,. \tag{60}$$

Then (58) can be written as

$$\Delta_i^{m+1} = \Delta_i^m - h_i\left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right) - \Phi^m - U^m\,, \tag{61}$$

which implies, according to Young's inequality, that, for any a:

$$\begin{aligned}&\mathbb{E}\left(|\Delta_i^{m+1}|^2 \big| r^m = i\right) = \mathbb{E}\left(|\Delta_i^{m+1} + U^m - U^m|^2 \big| r^m = i\right) \\ \leq &(1+a)\mathbb{E}\left(|\Delta_i^{m+1} + U^m|^2 \big| r^m = i\right) + \left(1 + \frac{1}{a}\right)\mathbb{E}\left(|U^m|^2 \big| r^m = i\right)\,. \end{aligned} \tag{62}$$

Both terms on the right-hand side of (62) are small. We now control the first term. Plug in the definition (61), we have:

$$\mathbb{E}\left(|\Delta_i^{m+1}+U^m|^2 \mid r^m=i\right)=\mathbb{E}\left(|\Delta_i^m-h_i\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)-\Phi^m|^2\middle|r^m=i\right). \tag{63}$$

Noting that

$$\begin{aligned}&\mathbb{E}\left(\left(\Delta_i^m-h_i\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right)\cdot\Phi^m\right)\\ =&\sqrt{2}\int_{T^m}^{T^m+h_i}\mathbb{E}\left[\int_{T^m}^{s}\left(\Delta_i^m-h_i\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right)\cdot\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\right]\mathrm{d}s=0\end{aligned}$$

because

$$\mathbb{E}\left[\int_{T^m}^{s}\left(\Delta_i^m-h_i\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right)\cdot\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\right]=0\,,$$

according to the property of Itô's integral, we can discard the cross terms with Φ^m in (63) to obtain

$$\begin{aligned}\mathbb{E}\left(|\Delta_i^{m+1}+U^m|^2 \mid r^m=i\right)=&\mathbb{E}|\Delta_i^m|^2-2h_i\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]\\ &+h_i^2\mathbb{E}\left|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right|^2+\mathbb{E}\left(|\Phi^m|^2\middle|r^m=i\right).\end{aligned} \tag{64}$$

For the last term of (64), we have the following control:

$$\begin{aligned}\mathbb{E}\left(|\Phi^m|^2\middle|r^m=i\right)=&\mathbb{E}\left(2\left|\int_{T^m}^{T^m+h_i}\int_{T^m}^{s}\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\,\mathrm{d}s\right|^2\middle|r^m=i\right)\\ \overset{\text{(I)}}{\leq}&2\mathbb{E}\left[\left(\int_{T^m}^{T^m+h_i}\mathrm{d}s\right)\left(\int_{T^m}^{T^m+h_i}\left|\int_{T^m}^{s}\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\right|^2\mathrm{d}s\right)\middle|r^m=i\right]\\ \leq&2h_i\int_{T^m}^{T^m+h_i}\mathbb{E}\left(\left|\int_{T^m}^{s}\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\right|^2\middle|r^m=i\right)\mathrm{d}s\\ \overset{\text{(II)}}{=}&2h_i\int_{T^m}^{T^m+h_i}\int_{T^m}^{s}\mathbb{E}\left(|\partial_{ii}f(\tilde{x}(z))|^2\middle|r^m=i\right)\mathrm{d}z\,\mathrm{d}s\\ \overset{\text{(III)}}{=}&h_i^3\mathbb{E}_p|\partial_{ii}f|^2=h_i^3L_i^2\,,\end{aligned}$$

where we use Hölder's inequality in I and $\tilde{x}(t)\sim p$ for all t in III. In II, we use the following property of Itô's integral:

$$\mathbb{E}\left(\left|\int_{T^m}^{s}\partial_{ii}f(\tilde{x}(z))\,\mathrm{d}B_z\right|^2\middle|r^m=i\right)=\int_{T^m}^{s}\mathbb{E}\left(|\partial_{ii}f(\tilde{x}(z))|^2\middle|r^m=i\right)\mathrm{d}z\,.$$

By substituting into (64), we obtain

$$\begin{aligned}\mathbb{E}\left(|\Delta_i^{m+1}+U^m|^2 \mid r^m=i\right)\leq&\mathbb{E}|\Delta_i^m|^2-2h_i\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right)\right]\\ &+h_i^2\mathbb{E}\left|\partial_i f(\tilde{x}^m)-\partial_i f(x^m)\right|^2+h_i^3L_i^2\end{aligned} \tag{65}$$

To bound the second term on the right-hand side of (62), we first note that f is three times continuously differentiable, and (11) implies $\|\partial_{iii}f\|_\infty\leq H_i$. Take $\mathrm{d}t$ on both sides of (34), under condition $r^m=i$, we first have

$$\mathrm{d}\tilde{x}_i(t)=-\partial_i f(\tilde{x}(s))\,\mathrm{d}s+\sqrt{2}\,\mathrm{d}B_s\,. \tag{66}$$

According to Itô's formula, we obtain

$$\partial_i f(\tilde{x}(t)) - \partial_i f(\tilde{x}^m) = \int_{T^m}^{t} \partial_{ii} f(\tilde{x}(s))\,\mathrm{d}\tilde{x}_i(s) + \int_{T^m}^{t} \partial_{iii} f(\tilde{x}(s))\,\mathrm{d}s\,. \tag{67}$$

Substituting (66) into (67), we have

$$\begin{aligned} &\partial_i f(\tilde{x}(t)) - \partial_i f(\tilde{x}^m) - \sqrt{2}\int_{T_m}^{t} \partial_{ii} f(\tilde{x}(s))\,\mathrm{d}B_s \\ &= \int_{T^m}^{t} -\partial_{ii} f(\tilde{x}(s))\partial_i f(\tilde{x}(s)) + \partial_{iii} f(\tilde{x}(s))\,\mathrm{d}s\,. \end{aligned} \tag{68}$$

By substituting into (59), we obtain

$$\begin{aligned} &\mathbb{E}\left(|U^m|^2 \mid r^m = i\right) \\ &\overset{\text{(I)}}{\leq} h_i \int_{T^m}^{T^m+h_i} \mathbb{E}\left(\left|\partial_i f(\tilde{x}(s)) - \partial_i f(\tilde{x}^m) - \sqrt{2}\int_{T^m}^{s} \partial_{ii} f(\tilde{x}(z))\,\mathrm{d}B_r\right|^2 \middle| r^m = i\right) \mathrm{d}s \\ &\overset{\text{(II)}}{=} h_i \int_{T^m}^{T^m+h_i} \mathbb{E}\left(\left|\int_{T^m}^{s} \left(-\partial_{ii} f(\tilde{x}(z))\partial_i f(\tilde{x}(z)) + \partial_{iii} f(\tilde{x}(z))\right)\mathrm{d}z\right|^2 \middle| r^m = i\right) \mathrm{d}s \\ &\overset{\text{(III)}}{\leq} h_i^2 \int_{T^m}^{T^m+h_i} \int_{T^m}^{s} \mathbb{E}\left(\left|\partial_{ii} f(\tilde{x}(z))\partial_i f(\tilde{x}(z)) + \partial_{iii} f(\tilde{x}(z))\right|^2 \middle| r^m = i\right) \mathrm{d}z\,\mathrm{d}s \\ &\overset{\text{(IV)}}{\leq} 2h_i^2 \int_{T^m}^{T^m+h_i} \int_{T^m}^{s} \mathbb{E}\left(\left|\partial_{ii} f(\tilde{x}(z))\partial_i f(\tilde{x}(z))\right|^2 \middle| r^m = i\right) \mathrm{d}z\,\mathrm{d}s \\ &\quad + 2h_i^2 \int_{T^m}^{T^m+h_i} \int_{T^m}^{s} \mathbb{E}\left(\left|\partial_{iii} f(\tilde{x}(z))\right|^2 \middle| r^m = i\right) \mathrm{d}z\,\mathrm{d}s \\ &\overset{\text{(V)}}{\leq} h_i^4 \left(L_i^3 + H_i^2\right)\,. \end{aligned} \tag{69}$$

In the derivation, (II) comes from plugging in (68), and (I) and (III) come from the use of Jensen's inequality, (V) comes from the use of Lipschitz continuity in the first and the second derivative ((7) and (11) in particular), and the fact that $\tilde{x}(t) \sim p$ for all t. Note also $\mathbb{E}_p|\partial_i f|^2 \leq L_i$ by (Dalalyan and Karagulyan, 2019, Lemma 3).

By plugging (65) and (69) into (57) and (62), we obtain

$$\begin{aligned} \mathbb{E}|\Delta_i^{m+1}|^2 \leq\ & (1+a\phi_i)\,\mathbb{E}|\Delta_i^m|^2 - 2(1+a)h\mathbb{E}\left[\Delta_i^m\left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right)\right] \\ & + \frac{(1+a)h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right|^2 + \frac{(1+a)h^3 L_i^2}{\phi_i^2} + \left(1+\frac{1}{a}\right)\frac{h^4\left(L_i^3 + H_i^2\right)}{\phi_i^3}\,, \end{aligned} \tag{70}$$

where we use $h_i\phi_i = h$. Comparing with (56), we need to set

$$a = h_i\mu = \frac{h\mu}{\phi_i} < 1\,,$$

where we use $h < \frac{\mu \min\{\phi_i\}}{8L^2}$. This leads to $1 + \frac{1}{a} \leq \frac{2\phi_i}{h\mu}$. By substituting into (48), we obtain

$$\mathbb{E}|\Delta_i^{m+1}|^2 \leq \left(1 + h\mu + \frac{h^2\mu^2}{\phi_i}\right) \mathbb{E}|\Delta_i^m|^2 - 2h\mathbb{E}\left[\Delta_i^m \left(\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right)\right] + \frac{3h^2}{\phi_i}\mathbb{E}\left|\partial_i f(\tilde{x}^m) - \partial_i f(x^m)\right|^2 + \frac{2h^3 L_i^2}{\phi_i^2} + \frac{2h^3\left(L_i^3 + H_i^2\right)}{\phi_i^2 \mu}.$$

Noting $L_i/\mu > 1$, we conclude the lemma. ■

The proof of Proposition 13 is obtained by summing up all components and applying Lemma 14.

Proof [Proof of Proposition 13] Noting that

$$\mathbb{E}|\Delta^{m+1}|^2 = \sum_{i=1}^d \mathbb{E}|\Delta_i^{m+1}|^2,$$

we substitute using (56) to obtain

$$\mathbb{E}|\Delta^{m+1}|^2 \leq \left(1 + h\mu + \frac{h^2\mu^2}{\min\{\phi_i\}}\right) \mathbb{E}|\Delta^m|^2 - 2h\mathbb{E}\left\langle \Delta^m, \nabla f(\tilde{x}^m) - \nabla f(x^m)\right\rangle + \frac{3h^2}{\min\{\phi_i\}}\mathbb{E}\left|\nabla f(\tilde{x}^m) - \nabla f(x^m)\right|^2 + \frac{4h^3}{\mu}\sum_{i=1}^d \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}. \tag{71}$$

The second and third terms in the right-hand side of this bound can be controlled by $\mathbb{E}|\Delta^m|^2$, as follows. By convexity, we have

$$\mathbb{E}\left\langle \Delta^m, \nabla f(\tilde{x}^m) - \nabla f(x^m)\right\rangle \geq \mu\mathbb{E}|\Delta^m|^2. \tag{72}$$

By the L-Lipschitz property, we have

$$\mathbb{E}\left|\nabla f(\tilde{x}^m) - \nabla f(x^m)\right|^2 \leq L^2\mathbb{E}|\Delta^m|^2. \tag{73}$$

By substituting (72) and (73) into (50), and using $\mu < L$, we have

$$\mathbb{E}|\Delta^{m+1}|^2 \leq \left(1 - h\mu + \frac{4h^2L^2}{\min\{\phi_i\}}\right) \mathbb{E}|\Delta^m|^2 + \frac{4h^3}{\mu}\sum_{i=1}^d \frac{\left(L_i^3 + H_i^2\right)}{\phi_i^2}. \tag{74}$$

Since $h < \frac{\mu \min\{\phi_i\}}{8L^2}$, we obtain (54). ■

Appendix E. Proof of Proposition 9

We now give the proof to Proposition 9 to show the convergence rate we obtain is sharp.

Proof [Proof of Proposition 9] For this special target distribution p, the objective function is $f(x) = \sum_{i=1}^d \frac{|x_i|^2}{2}$. With $\alpha = 0$ and $\phi_i = 1/d$, we have: $x_i^{m+1} = x_i^m$ for all $i \neq r^m$ and

$$x_{r^m}^{m+1} = (1 - dh)x_{r^m}^m + \sqrt{2dh}\xi^m.$$

Therefore for all $i = 1, 2, \ldots, d$, we have

$$\begin{aligned}\mathbb{E}|x_i^{m+1}|^2 &= \frac{1}{d}\mathbb{E}\left(|x_i^{m+1}|^2 \middle| r^m = i\right) + \left(1 - \frac{1}{d}\right)\mathbb{E}\left(|x_i^{m+1}|^2 \mid r^m \neq i\right) \\ &= \frac{1}{d}\mathbb{E}\left(|(1-dh)x_i^m + \sqrt{2dh}\xi^m|^2 \middle| r^m = i\right) + \left(1 - \frac{1}{d}\right)\mathbb{E}\left(|x_i^m|^2\right) \\ &= \left(1 - 2h + dh^2\right)\mathbb{E}|x_i^m|^2 + 2h \end{aligned} \tag{75}$$

where we use $\mathbb{E}_\xi \left|x_i^m - dhx_i^m + \sqrt{2dh}\xi^m\right|^2 = (1-dh)^2|x_i^m|^2 + 2dh$ in the last equation. By summing (75) over i, we obtain

$$\mathbb{E}|x^{m+1}|^2 = \left(1 - 2h + dh^2\right)\mathbb{E}|x^m|^2 + 2dh\,.$$

Using it iteratively, and considering $\mathbb{E}|x^0|^2 = 3d$, we have:

$$\begin{aligned}\mathbb{E}|x^m|^2 &\geq 3d\left(1 - 2h + dh^2\right)^m + \left(1 - \left(1 - 2h + dh^2\right)^m\right)\frac{2dh}{2h - dh^2} \\ &= d\left(1 - 2h + dh^2\right)^m + \frac{2d}{2 - dh} + 2d\left(1 - \frac{1}{2 - dh}\right)\left(1 - 2h + dh^2\right)^m \\ &\geq d\left(1 - 2h\right)^m + \frac{2d}{2 - dh}\,,\end{aligned}$$

where we use $dh \leq 1$ in the last inequality.

Since

$$W(q_m, p) \geq \left(\int |x|^2 q_m(x)\,\mathrm{d}x\right)^{1/2} - \left(\int |x|^2 p(x)\,\mathrm{d}x\right)^{1/2} = \left(\int |x|^2 q_m(x)\,\mathrm{d}x\right)^{1/2} - \sqrt{d}\,,$$

we have

$$\begin{aligned}W(q_m, p) \geq \left(\int |x|^2 q_m(x)\,\mathrm{d}x\right)^{1/2} - \sqrt{d} &\geq \frac{d\left(1 - 2h\right)^m + \frac{2d}{2-dh} - d}{\sqrt{d\left(1 - 2h\right)^m + \frac{2d}{2-dh}} + \sqrt{d}} \\ &\geq \frac{\sqrt{d}}{3}\left(1 - 2h\right)^m + \frac{d^{3/2}h}{6} \\ &\geq \exp\left(-2mh\right)\frac{\sqrt{d}}{3} + \frac{d^{3/2}h}{6}\,,\end{aligned}$$

where in the last inequality we use

$$\sqrt{d\left(1 - 2h\right)^m + \frac{2d}{2 - dh}} + \sqrt{d} \leq 3\sqrt{d}.$$

Therefore, we finally prove (27). ■