Sampling from a log-concave distribution with Projected Langevin Monte Carlo

Sébastien Bubeck, Ronen Eldan, Joseph Lehec

Sampling from a log-concave distribution with Projected Langevin Monte Carlo

Sébastien Bubeck * Ronen Eldan † Joseph Lehec ‡

August 8, 2016

Abstract

We extend the Langevin Monte Carlo (LMC) algorithm to compactly supported measures via a projection step, akin to projected Stochastic Gradient Descent (SGD). We show that (projected) LMC allows to sample in polynomial time from a log-concave distribution with smooth potential. This gives a new Markov chain to sample from a log-concave distribution. Our main result shows in particular that when the target distribution is uniform, LMC mixes in $\widetilde{O}(n^7)$ steps (where n is the dimension). We also provide preliminary experimental evidence that LMC performs at least as well as hit-and-run, for which a better mixing time of $\widetilde{O}(n^4)$ was proved by Lovász and Vempala.

1 Introduction

Let $K \subset \mathbb{R}^n$ be a convex body such that $0 \in K$, K contains a Euclidean ball of radius r, and K is contained in a Euclidean ball of radius R. Denote $\mathcal{P}_K$ for the Euclidean projection on K. Let $f : K \to \mathbb{R}$ be a L-Lipschitz and β-smooth convex function, that is f is differentiable and statisfies $\forall x, y \in K, |\nabla f(x) - \nabla f(y)| \leq \beta |x - y|$, and $|\nabla f(x)| \leq L$. We are interested in the problem of sampling from the probability measure μ on $\mathbb{R}^n$ whose density with respect to the Lebesgue measure is given by:

$$\frac{d\mu}{dx} = \frac{1}{Z} \exp(-f(x)) \mathbb{1}\{x \in K\}, \quad \text{where} \quad Z = \int_{y \in K} \exp(-f(y)) dy.$$

In this paper we study the following Markov chain, which depends on a parameter $\eta > 0$, and where $\xi_1, \xi_2, \ldots$ is an i.i.d. sequence of standard Gaussian random variables in $\mathbb{R}^n$:

$$\overline{X}_{k+1} = \mathcal{P}_K \left(\overline{X}_k - \frac{\eta}{2} \nabla f(\overline{X}_k) + \sqrt{\eta} \xi_k \right), \tag{1}$$

with $\overline{X}_0 = 0$.

*Microsoft Research; sebubeck@microsoft.com.

†Weizmann Institute; roneneldan@gmail.com.

‡Université Paris-Dauphine; lehec@ceremade.dauphine.fr.

Recall that the total variation distance between two measures μ, ν is defined as $\mathrm{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$ where the supremum is over all measurable sets A. With a slight abuse of notation we sometimes write $\mathrm{TV}(X, \nu)$ where X is a random variable distributed according to μ. The notation $v_n = \widetilde{O}(u_n)$ (respectively $\widetilde{\Omega}$) means that there exists $c \in \mathbb{R}, C > 0$ such that $v_n \leq C u_n \log^c(u_n)$ (respectively $\geq$). We also say $v_n = \widetilde{\Theta}(u_n)$ if one has both $v_n = \widetilde{O}(u_n)$ and $v_n = \widetilde{\Omega}(u_n)$. Our main result is the following:

Theorem 1 *Assume that $r = 1$ and let $\varepsilon > 0$. Then one has $\mathrm{TV}(\overline{X}_N, \mu) \leq \varepsilon$ provided that $\eta = \widetilde{\Theta}(R^2/N)$ and that N satisfies the following: if μ is uniform then*

$$N = \widetilde{\Omega}\left(\frac{R^6 n^7}{\varepsilon^8}\right),$$

and otherwise

$$N = \widetilde{\Omega}\left(\frac{R^6 \max(n, RL, R\beta)^{12}}{\varepsilon^{12}}\right).$$

1.1 Context and related works

There is a long line of works in theoretical computer science proving results similar to Theorem 1, starting with the breakthrough result of Dyer et al. [1991] who showed that the lattice walk mixes in $\widetilde{O}(n^{23})$ steps. The current record for the mixing time is obtained by Lovász and Vempala [2007], who show a bound of $\widetilde{O}(n^4)$ for the hit-and-run walk. These chains (as well as other popular chains such as the ball walk or the Dikin walk, see e.g. Kannan and Narayanan [2012] and references therein) all require a *zeroth-order oracle* for the potential f, that is given x one can calculate the value $f(x)$. On the other hand our proposed chain (1) works with a *first-order oracle*, that is given x one can calculate the value of $\nabla f(x)$. The difference between zeroth-order oracle and first-order oracle has been extensively studied in the optimization literature (e.g., Nemirovski and Yudin [1983]), but it has been largely ignored in the literature on polynomial-time sampling algorithms. We also note that hit-and-run and LMC are the only chains which are rapidly mixing from any starting point (see Lovász and Vempala [2006]), though they have this property for seemingly very different reasons. When initialized in a corner of the convex body, hit-and-run might take a long time to take a step, but once it moves it escapes very far (while a chain such as the ball walk would only do a small step). On the other hand LMC keeps moving at every step, even when initialized in a corner, thanks for the projection part of (1).

Our main motivation to study the chain (1) stems from its connection with the ubiquitous *stochastic gradient descent* (SGD) algorithm. In general this algorithm takes the form $x_{k+1} = \mathcal{P}_K\left(x_k - \eta \nabla f(x_k) + \varepsilon_k\right)$ where $\varepsilon_1, \varepsilon_2, \ldots$ is a centered i.i.d. sequence. Standard results in approximation theory, such as Robbins and Monro [1951], show that if the variance of the noise $\mathrm{Var}(\varepsilon_1)$ is of smaller order than the step-size η then the iterates (x_k) converge to the minimum of f on K (for a step-size decreasing sufficiently fast as a function of the number of iterations). For the specific noise sequence that we study in (1), the variance is exactly equal to the step-size, which is why the chain deviates from its standard and well-understood behavior. We also note that other regimes where SGD does not converge to the minimum of f have been studied in the optimization literature, such as the constant step-size case investigated in Pflug [1986], Bach and

Moulines [2013].

The chain (1) is also closely related to a line of works in Bayesian statistics on Langevin Monte Carlo algorithms, starting essentially with Tweedie and Roberts [1996]. The focus there is on the unconstrained case, that is $K = \mathbb{R}^n$. In this simpler situation, a variant of Theorem 1 was proven in the recent paper Dalalyan [2014]. The latter result is the starting point of our work. A straightforward way to extend the analysis of Dalalyan to the constrained case is to run the unconstrained chain with an additional potential that diverges quickly as the distance from x to K increases. However it seems much more natural to study directly the chain (1). Unfortunately the techniques used in Dalalyan [2014] cannot deal with the singularities in the diffusion process which are introduced by the projection. As we explain in Section 1.2 our main contribution is to develop the appropriate machinery to study (1).

In the machine learning literature it was recently observed that Langevin Monte Carlo algorithms are particularly well-suited for large-scale applications because of the close connection to SGD. For instance Welling and Teh [2011] suggest to use mini-batch to compute approximate gradients instead of exact gradients in (1), and they call the resulting algorithm SGLD (Stochastic Gradient Langevin Dynamics). It is conceivable that the techniques developed in this paper could be used to analyze SGLD and its refinements introduced in Ahn et al. [2012]. We leave this as an open problem for future work. Another interesting direction for future work is to improve the polynomial dependency on the dimension and the inverse accuracy in Theorem 1 (our main goal here was to provide the simplest polynomial-time analysis).

1.2 Contribution and paper organization

As we pointed out above, Dalalyan [2014] proves the equivalent of Theorem 1 in the unconstrained case. His elegant approach is based on viewing LMC as a discretization of the diffusion process $dX_t = dW_t - \frac{1}{2}\nabla f(X_t)$, where (W_t) is a Brownian motion. The analysis then proceeds in two steps, by deriving first the mixing time of the diffusion process, and then showing that the discretized process is 'close' to its continuous version. In Dalalyan [2014] the first step is particularly clean as he assumes α-strong convexity for the potential, which in turns directly gives a mixing time of order $1/\alpha$. The second step is also rather simple once one realizes that LMC can be viewed as the diffusion process $d\overline{X}_t = dW_t - \frac{1}{2}\nabla f(X_{\eta\lfloor\frac{t}{\eta}\rfloor})$. Using Pinsker's inequality and Girsanov's formula it is then a short calculation to show that the total variation distance between $\overline{X}_t$ and X_t is small.

The constrained case presents several challenges, arising from the *reflection* of the diffusion process on the boundary of K, and from the lack of curvature in the potential (indeed the constant potential case is particularly important for us as it corresponds to μ being the uniform distribution on K). Rather than a simple Brownian motion with drift, LMC with projection can be viewed as the discretization of *reflected Brownian motion with drift*, which is a process of the form $dX_t = dW_t - \frac{1}{2}\nabla f(X_t)dt - \nu_t L(dt)$, where $X_t \in K, \forall t \geq 0$, L is a measure supported on $\{t \geq 0 : \ X_t \in \partial K\}$, and ν_t is an outer normal unit vector of K at X_t. The term $\nu_t L(dt)$ is referred to as the *Tanaka drift*. Following Dalalyan [2014] the analysis is again decomposed in two steps. We study the mixing time of the continuous process via a simple coupling argument, which

crucially uses the convexity of K and of the potential f. The main difficulty is in showing that the discretized process $(\overline{X}_t)$ is close to the continuous version (X_t), as the Tanaka drift prevents us from a straightforward application of Girsanov's formula. Our approach around this issue is to first use a geometric argument to prove that the two processes are close in Wasserstein distance, and then to show that in fact for a reflected Brownian motion with drift one can deduce a total variation bound from a Wasserstein bound.

The paper is organized as follows. We start in Section 2 by proving Theorem 1 for the case of a uniform distribution. We first remind the reader of Tanaka's construction (Tanaka [1979]) of reflected Brownian motion in Subsection 2.1. We present our geometric argument to bound the Wasserstein distance between (X_t) and $(\overline{X}_t)$ in Subsection 2.2, and we use our coupling argument to bound the mixing time of (X_t) in Subsection 2.3. Then in Subsection 2.4 we use properties of reflected Brownian to show that one can obtain a total variation bound from the Wasserstein bound of Subsection 2.2. We conclude the proof of the first part of Theorem 1 in Subsection 2.5. In Section 3 we generalize these arguments to an arbitrary smooth potential. Finally we conclude the paper in Section 4 with some preliminary experimental comparison between LMC and hit-and-run.

2 The constant potential case

In this section we prove Theorem 1 for the case where μ is uniform, that is $\nabla f = 0$. First we introduce some useful notation. For a point $x \in \partial K$ we say that ν is an outer unit normal vector at x if $|\nu| = 1$ and

$$\langle x - x', \nu \rangle \geq 0, \quad \forall x' \in K.$$

For $x \notin \partial K$ we say that 0 is an outer unit normal at x. Let $\|\cdot\|_K$ be the gauge of K defined by

$$\|x\|_K = \inf\{t \geq 0;\ x \in tK\}, \quad x \in \mathbb{R}^n,$$

and h_K the support function of K by

$$h_K(y) = \sup\{\langle x, y\rangle;\ x \in K\}, \quad y \in \mathbb{R}^n.$$

Note that h_K is also the gauge function of the polar body of K. Finally we denote $m = \int |x| \mu(dx)$, and $M = \mathbb{E}\left[\|\theta\|_K\right]$, where θ is uniform on the sphere $\mathbb{S}^{n-1}$.

2.1 The Skorokhod problem

Let $T \in \mathbb{R}_+ \cup \{+\infty\}$ and $w\colon [0,T) \to \mathbb{R}^n$ be a piecewise continuous path with $w(0) \in K$. We say that $x\colon [0,T) \to \mathbb{R}^n$ and $\varphi\colon [0,T) \to \mathbb{R}^n$ solve the Skorokhod problem for w if one has $x(t) \in K, \forall t \in [0,T)$,

$$x(t) = w(t) + \varphi(t), \quad \forall t \in [0,T),$$

and furthermore φ is of the form

$$\varphi(t) = -\int_0^t \nu_s\, L(ds), \quad \forall t \in [0,T),$$

where ν_s is an outer unit normal at $x(s)$, and L is a measure on $[0,T]$ supported on the set $\{t \in [0,T) : \ x(t) \in \partial K\}$.

The path x is called the *reflection* of w at the boundary of K, and the measure L is called the *local time* of x at the boundary of K. Skorokhod showed the existence of such a a pair (x, φ) in dimension 1 in Skorokhod [1961], and Tanaka extended this result to convex sets in higher dimensions in Tanaka [1979]. Furthermore Tanaka also showed that the solution is unique, and if w is continuous then so is x and φ. In particular the reflected Brownian motion in K, denoted (X_t), is defined as the reflection of the standard Brownian motion (W_t) at the boundary of K (existence follows by continuity of W_t). Observe that by Itô's formula, for any smooth function g on $\mathbb{R}^n$,

$$g(X_t) - g(X_0) = \int_0^t \langle \nabla g(X_s), dW_s \rangle + \frac{1}{2} \int_0^t \Delta g(X_s)\, ds - \int_0^t \langle \nabla g(X_s), \nu_s \rangle\, L(ds). \tag{2}$$

To get a sense of what a solution typically looks like, let us work out the case where w is piecewise constant (this will also be useful to realize that LMC can be viewed as the solution to a Skorokhod problem). For a sequence $g_1 \dots g_N \in \mathbb{R}^n$, and for $\eta > 0$, we consider the path:

$$w(t) = \sum_{k=1}^{N} g_k \, \mathbb{1}\{t \geq k\eta\}, \qquad t \in [0, (N+1)\eta).$$

Define $(x_k)_{k=0,\dots,N}$ inductively by $x_0 = 0$ and

$$x_{k+1} = \mathcal{P}_K(x_k + g_k).$$

It is easy to verify that the solution to the Skorokhod problem for w is given by $x(t) = x_{\lfloor \frac{t}{\eta} \rfloor}$ and $\varphi(t) = -\int_0^t \nu_s \, L(ds)$, where the measure L is defined by (denoting δ_s for a dirac at s)

$$L = \sum_{k=1}^{N} |x_k + g_k - \mathcal{P}_K(x_k + g_k)| \delta_{k\eta},$$

and for $s = k\eta$,

$$\nu_s = \frac{x_k + g_k - \mathcal{P}_K(x_k + g_k)}{|x_k + g_k - \mathcal{P}_K(x_k + g_k)|}.$$

2.2 Discretization of reflected Brownian motion

Given the discussion above, it is clear that when f is a constant function, the chain (1) can be viewed as the reflection $(\overline{X}_t)$ of a discretized Brownian motion $\overline{W}_t := W_{\eta \lfloor \frac{t}{\eta} \rfloor}$ at the boundary of K (more precisely the value of $\overline{X}_{k\eta}$ coincides with the value of $\overline{X}_k$ as defined by (1)). It is rather clear that the discretized Brownian motion $(\overline{W}_t)$ is "close" to the path (W_t), and we would like to carry this to the reflected paths $(\overline{X}_t)$ and (X_t). The following lemma extracted from Tanaka [1979] allows to do exactly that.

Lemma 1 *Let w and $\overline{w}$ be piecewise continuous path and assume that (x, φ) and $(\overline{x}, \overline{\varphi})$ solve the Skorokhod problems for w and $\overline{w}$, respectively. Then for all time t we have*

$$\begin{aligned} |x(t) - \overline{x}(t)|^2 \leq\ & |w(t) - \overline{w}(t)|^2 \\ & + 2 \int_0^t \langle w(t) - \overline{w}(t) - w(s) + \overline{w}(s), \varphi(ds) - \overline{\varphi}(ds) \rangle. \end{aligned}$$

In the next lemma we control the local time at the boundary of the reflected Brownian motion (X_t).

Lemma 2 *We have, for all $t > 0$*

$$\mathbb{E}\left[\int_0^t h_K(\nu_s)\, L(ds)\right] \leq \frac{nt}{2}.$$

Proof By Itô's formula

$$d|X_t|^2 = 2\langle X_t, dW_t\rangle + n\, dt - 2\langle X_t, \nu_t\rangle\, L(dt).$$

Now observe that by definition of the reflection, if t is in the support of L then

$$\langle X_t, \nu_t\rangle \geq \langle x, \nu_t\rangle, \quad \forall x \in K.$$

In other words $\langle X_t, \nu_t\rangle \geq h_K(\nu_t)$. Therefore

$$2\int_0^t h_K(\nu_s)\, L(ds) \leq 2\int_0^t \langle X_s, dW_s\rangle + nt + |X_0|^2 - |X_t|^2.$$

The first term of the right–hand side is a martingale, so using that $X_0 = 0$ and taking expectation we get the result. ■

Lemma 3 *There exists a universal constant C such that*

$$\mathbb{E}\left[\sup_{[0,T]} \|W_t - \overline{W}_t\|_K\right] \leq C\, M\, n^{1/2}\eta^{1/2}\log(T/\eta)^{1/2}.$$

Proof Note that

$$\mathbb{E}\left[\sup_{[0,T]} \|W_t - \overline{W}_t\|_K\right] = \mathbb{E}\left[\max_{0\leq i\leq N-1} Y_i\right]$$

where

$$Y_i = \sup_{t\in[i\eta,(i+1)\eta)} \|W_t - W_{i\eta}\|_K.$$

Observe that the variables (Y_i) are identically distributed, let $p \geq 1$ and write

$$\mathbb{E}\left[\max_{i\leq N-1} Y_i\right] \leq \mathbb{E}\left[\left(\sum_{i=0}^{N-1} |Y_i|^p\right)^{1/p}\right] \leq N^{1/p}\, \|Y_0\|_p.$$

We claim that

$$\|Y_0\|_p \leq C\sqrt{p\, n\, \eta}\, M \tag{3}$$

for some constant C, and for all $p \geq 2$. Taking this for granted and choosing $p = \log(N)$ in the previous inequality yields the result (recall that $N = T/\eta$). So it is enough to prove (3). Observe that since (W_t) is a martingale, the process

$$M_t = \|W_t\|_K$$

is a sub–martingale. By Doob's maximal inequality

$$\|Y_0\|_p = \|\sup_{[0,\eta]} M_t\|_p \leq 2\|M_\eta\|_p,$$

for every $p \geq 2$. Letting γ_n be the standard Gaussian measure on $\mathbb{R}^n$ and using Khintchin's inequality we get

$$\begin{aligned}\|M_\eta\|_p &= \sqrt{\eta}\left(\int_{\mathbb{R}^n} \|x\|_K^p \, \gamma_n(dx)\right)^{1/p} \\ &\leq C\sqrt{p\eta}\int_{\mathbb{R}^n} \|x\|_K \, \gamma_n(dx)\end{aligned}$$

Lastly, integrating in polar coordinate, it is easily seen that

$$\int_{\mathbb{R}^n} \|x\|_K \, \gamma_n(dx) \leq C\sqrt{n}\, M.$$

Hence the result. ■

We are now in a position to bound the average distance between X_T and its discretization $\overline{X}_T$.

Proposition 1 *There exists a universal constant C such that for any $T \geq 0$ we have*

$$\mathbb{E}[|X_T - \overline{X}_T|] \leq C\,(\eta \log(T/\eta))^{1/4}\, n^{3/4}\, T^{1/2}\, M^{1/2}$$

Proof Applying Lemma 1 to the processes (W_t) and $(\overline{W}_t)$ at time $T = N\eta$ yields (note that $W_T = \overline{W}_T$)

$$|X_T - \overline{X}_T|^2 \leq 2\int_0^T \langle W_t - \overline{W}_t, \nu_t\rangle L(dt) - 2\int_0^T \langle W_t - \overline{W}_t, \overline{\nu}_t\rangle \overline{L}(dt)$$

We claim that the second integral is equal to 0. Indeed, since the discretized process is constant on the intervals $[k\eta, (k+1)\eta)$ the local time $\overline{L}$ is a positive combination of Dirac point masses at

$$\eta, 2\eta, \ldots, N\eta.$$

On the other hand $W_{k\eta} = \overline{W}_{k\eta}$ for all integer k, hence the claim. Therefore

$$|X_T - \overline{X}_T|^2 \leq 2\int_0^T \langle W_t - \overline{W}_t, \nu_t\rangle \, L(dt)$$

Using the inequality $\langle x, y\rangle \leq \|x\|_K \, h_K(y)$ we get

$$|X_T - \overline{X}_T|^2 \leq 2\sup_{[0,T]} \|W_t - \overline{W}_T\|_K \int_0^T h_K(\nu_t)\, L(dt).$$

Taking the square root, expectation and using Cauchy–Schwarz we get

$$\mathbb{E}\left[|X_T - \overline{X}_T|\right]^2 \leq 2\,\mathbb{E}\left[\sup_{[0,T]} \|W_t - \overline{W}_T\|_K\right] \mathbb{E}\left[\int_0^T h_K(\nu_t)\, L(dt)\right].$$

Applying Lemma 2 and Lemma 3, we get the result. ■

2.3 A mixing time estimate for the reflected Brownian motion

The reflected Brownian motion is a Markov process. We let (P_t) be the associated semi–group:

$$P_t f(x) = \mathbb{E}_x[f(X_t)],$$

for every test function f, where $\mathbb{E}_x$ means conditional expectation given $X_0 = x$. Itô's formula shows that the generator of the semigroup (P_t) is $(1/2)\Delta$ with Neumann boundary condition. Then by Stokes' formula, it is easily seen that μ (the uniform measure on K normalized to be a probability measure) is the stationary measure of this process, and is even reversible. In this section we estimate the total variation between the law of (X_t) and μ.

Given a probability measure ν supported on K, we let νP_t be the law of X_t when X_0 as law ν. The following lemma is the key result to estimate the mixing time of the process (X_t).

Lemma 4 *Let $x, x' \in K$*

$$\mathrm{TV}(\delta_x P_t, \delta_{x'} P_t) \leq \frac{|x - x'|}{\sqrt{2\pi t}}.$$

Proof Let (W_t) be a Brownian motion starting from 0 and let (X_t) be a reflected Brownian motion starting from x:

$$\begin{cases} X_0 = x \\ dX_t = dW_t - \nu_t \, L(dt) \end{cases} \tag{4}$$

where (ν_t) and L satisfy the appropriate conditions. We construct a reflected Brownian motion (X'_t) starting from x' as follows. Let

$$\tau = \inf\{t \geq 0;\ X_t = X'_t\},$$

and for $t < \tau$ let S_t be the orthogonal reflection with respect to the hyperplane $(X_t - X'_t)^\perp$. Then up to time τ, the process (X'_t) is defined by

$$\begin{cases} X'_0 = x' \\ dX'_t = dW'_t - \nu'_t \, L'(dt) \\ dW'_t = S_t(dW_t) \end{cases} \tag{5}$$

where L' is a measure supported on

$$\{t \leq \tau;\ X'_t \in \partial K\}$$

and ν'_t is an outer unit normal at X'_t for all such t. After time τ we just set $X'_t = X_t$. Since S_t is an orthogonal map (W'_t) is a Brownian motion and thus (X'_t) is a reflected Brownian motion starting from x'. Therefore

$$\mathrm{TV}(\delta_x P_t, \delta_{x'} P_t) \leq \mathbb{P}(X_t \neq X'_t) = \mathbb{P}(\tau > t).$$

Observe that on $[0, \tau)$

$$dW_t - dW'_t = (\mathrm{I} - S_t)(dW_t) = 2\langle V_t, dW_t \rangle V_t,$$

where

$$V_t = \frac{X_t - X'_t}{|X_t - X'_t|}.$$

So

$$\begin{aligned} d(X_t - X'_t) &= 2\langle V_t, dW_t\rangle V_t - \nu_t\, L(dt) + \nu'_t\, L'(dt) \\ &= 2(dB_t)\, V_t - \nu_t\, L(dt) + \nu'_t\, L'(dt), \end{aligned}$$

where

$$B_t = \int_0^t \langle V_s, dW_s\rangle, \quad \text{on } [0,\tau).$$

Observe that (B_t) is a one–dimensional Brownian motion. Itô's formula then gives

$$\begin{aligned} dg(X_t - X'_t) = {} & 2\langle \nabla g(X_t - X'_t), V_t\rangle\, dB_t - \langle \nabla g(X_t - X'_t), \nu_t\rangle\, L(dt) \\ & + \langle \nabla g(X_t - X'_t), \nu' t\rangle\, L'(dt) + 2\nabla^2 g(X_t - X'_t)(V_t, V_t)\, dt, \end{aligned}$$

for every g which is smooth in a neighborhood of $X_t - X'_t$. Now if $g(x) = |x|$ then

$$\nabla g(X_t - X'_t) = V_t$$

so

$$\begin{aligned} &\langle \nabla g(X_t - X'_t), V_t\rangle = 1 \\ &\langle \nabla g(X_t - X'_t), \nu_t\rangle \geq 0, \quad \text{on the support of } L \\ &\langle \nabla g(X_t - X'_t), \nu'_t\rangle \leq 0, \quad \text{on the support of } L'. \end{aligned} \tag{6}$$

Moreover

$$\nabla^2 g(X_t - X'_t) = \frac{1}{|X_t - Y_t|}\, P_{(X_t - Y_t)^\perp}$$

where $P_{x^\perp}$ denotes the orthogonal projection on $x^\perp$. In particular

$$\nabla^2 g(X_t - Y_t)(V_t) = 0.$$

We obtain

$$|X_t - X'_t| \leq |x - x'| + 2B_t, \quad \text{on } [0,\tau).$$

Therefore

$$\mathbb{P}(\tau > t) \leq \mathbb{P}(\tau' > t)$$

where τ' is the first time the Brownian motion (B_t) hits the value $-|x-x'|/2$. Now by the reflection principle

$$\mathbb{P}(\tau' > t) = 2\,\mathbb{P}\,(0 \leq 2\,B_t < |x - x'|) \leq \frac{|x - x'|}{\sqrt{2\pi t}}.$$

Hence the result. ■

The above result clearly implies that for a probability measure ν on K,

$$\mathrm{TV}(\delta_0 P_t, \nu P_t) \leq \frac{\int_K |x|\, \nu(dx)}{\sqrt{2\pi t}}.$$

Since μ is stationary, we obtain

$$\mathrm{TV}(\delta_0 P_t, \mu) \leq \frac{m}{\sqrt{2\pi t}} \tag{7}$$

for any $t > 0$. In other words, starting from 0, the mixing time of (X_t) is of order at most m^2. Notice also that Lemma 4 allows to bound the mixing time from any starting point: for every $x \in K$, we have

$$\mathrm{TV}(\delta_x P_t, \mu) \leq \frac{R}{\sqrt{2\pi t}},$$

where R is the diameter of K. Letting τ_{mix} be the mixing time of (X_t), namely the smallest time t for which

$$\sup_{x \in K} \{\mathrm{TV}(\delta_x P_t, \mu)\} \leq \frac{1}{e},$$

we obtain from the previous display $\tau_{mix} \leq 2R^2$. Since for any x and t we have $\mathrm{TV}(\delta_x P_t, \mu) \leq e^{-\lfloor t/\tau_{mix} \rfloor}$ (see e.g., [Levin et al., 2008, Lemma 4.12]) we obtain in particular

$$\mathrm{TV}(\delta_0 P_t, \mu) \leq e^{-\lfloor t/2R^2 \rfloor}$$

The advantage of this upon (7) is the exponential decay in t. On the other hand, since obviously $m \leq R$, inequality (7) can be more precise for a certain range of t. The next proposition sums up the results of this section.

Proposition 2 *For any $t > 0$, we have*

$$\mathrm{TV}(\delta_0 P_t, \mu) \leq C \, \min\left(m\, t^{-1/2}, e^{-t/2R^2}\right),$$

where C is a universal constant.

2.4 From Wasserstein distance to total variation

In the following lemma, which is a variation on the reflection principle, (W_t) is a Brownian motion, the notation $\mathbb{P}_x$ means probability given $W_0 = x$ and (Q_t) denotes the heat semigroup:

$$Q_t h(x) = \mathbb{E}_x[h(W_t)],$$

for every test function h.

Lemma 5 *Let $x \in K$ and let σ be the first time (W_t) hits the boundary of K. Then for all $t > 0$*

$$\mathbb{P}_x(\sigma < t) \leq 2\mathbb{P}_x(W_t \notin K) = 2Q_t(\mathbb{1}_{K^c})(x).$$

Proof Let $(\mathcal{F}_t)$ be the natural filtration of the Brownian motion. Fix $t > 0$. By the strong Markov property

$$\mathbb{P}_x(W_t \notin K \mid \mathcal{F}_\sigma) = u(\sigma, W_\sigma), \tag{8}$$

where

$$u(s, y) = \mathbb{1}\{s < t\}\, \mathbb{P}_y(W_{t-s} \notin K).$$

Let $y \in \partial K$, since K is convex it admits a supporting hyperplane H at y. Let H_+ be the halfspace delimited by H containing K. Then for any $u > 0$

$$\mathbb{P}_y(W_u \notin K) \geq \mathbb{P}_y(W_u \notin H_+) = \frac{1}{2}.$$

Equality (8) thus yields

$$\mathbb{P}_x(W_t \notin K \mid \mathcal{F}_\sigma) \geq \frac{1}{2}\mathbb{1}\{\sigma < t\},$$

almost surely. Taking expectation yields the result. ■

We also need the following elementary estimate for the heat semigroup.

Lemma 6 *For any $s \geq 0$*

$$\int_K Q_s(\mathbb{1}_{K^c})\,dx \leq \sqrt{s}\,\mathcal{H}^{n-1}(\partial K),$$

where $\mathcal{H}^{n-1}(\partial K)$ is the Hausdorff measure of the boundary of K.

Proof Let $\varphi(s) = \int_K Q_s(\mathbb{1}_{K^c})\,dx$. Then by definition of the heat semigroup and Stokes' formula

$$\varphi'(s) = \frac{1}{2}\int_K \Delta Q_s(\mathbb{1}_{K^c})\,dx = \frac{1}{2}\int_{\partial K} \langle \nabla Q_s(\mathbb{1}_{K^c})(x), \nu(x)\rangle\,\mathcal{H}^{n-1}(dx),$$

for every $s > 0$ and where $\nu(x)$ is an outer unit normal vector at point x. On the other hand an elementary computation shows that for every $s > 0$

$$|\nabla Q_s(\mathbb{1}_{K^c})| \leq s^{-1/2}, \tag{9}$$

pointwise. We thus obtain

$$|\varphi'(s)| \leq \frac{\mathcal{H}^{n-1}(\partial K)}{2\sqrt{s}},$$

for every $s > 0$. Integrating this inequality between 0 and s yields the result. ■

Proposition 3 *Let T, S be integer multiples of η. Then*

$$\mathrm{TV}(X_{T+S}, \overline{X}_{T+S}) \leq \frac{3\mathbb{E}|X_T - \overline{X}_T|}{\sqrt{S}} + \mathrm{TV}(X_T, \mu) + 4\sqrt{S}\,\mathcal{H}^{n-1}(\partial K)\,|K|^{-1}.$$

Proof We use the coupling by reflection again. Fix x and x' in K. Let (X_t) and (X'_t) be two Brownian motions reflected at the boundary of K starting from x and x' respectively, such that the underlying Brownian motions (W_t) and (W'_t) are coupled by reflection, just as in the proof of Lemma 4. Let $(\overline{X'}_t)$ be the discretization of (X'_t), namely the solution of the Skorokhod problem for the process $\left(W'_{\eta\lfloor t/\eta\rfloor}\right)$. Let S be a integer multiple of η. Obviously, if (X_t) and (X'_t) have merged before time S and in the meantime neither (X_t) nor (X'_t) has hit the boundary of K then

$$X_S = X'_S = \overline{X'}_S.$$

Therefore, letting τ be the first time $X_t = X'_t$ and σ and σ' be the first times (X_t) and (X'_t) hit the boundary of K, respectively, we have

$$\mathbb{P}(X_S \neq \overline{X}'_S) \leq \mathbb{P}(\tau > S) + \mathbb{P}(\sigma < S) + \mathbb{P}(\sigma' < S), \tag{10}$$

As we have seen before, the coupling time τ satisfies

$$\mathbb{P}(\tau > S) \leq \frac{|x - x'|}{\sqrt{2\pi S}}.$$

On the other hand Lemma 5 gives

$$\mathbb{P}(\sigma < S) \leq 2\, Q_S(\mathbb{1}_{K^c})(x),$$

and similarly for σ'. Notice also that the estimate (9) implies that

$$Q_S(\mathbb{1}_{K^c})(x') \leq Q_S(\mathbb{1}_{K^c})(x) + \frac{|x - x'|}{\sqrt{S}}.$$

Plugging everything back into (10) yields

$$\mathbb{P}(X_S \neq \overline{X}'_S) \leq \frac{3|x - x'|}{\sqrt{S}} + 4\, Q_S(\mathbb{1}_{K^c})(x). \tag{11}$$

Now let T and S be two integer multiples of η and assume that (X_t) and $(\overline{X}_t)$ start from 0 and are coupled using the same Brownian motion up to time T, and using the reflection coupling between time T and $T + S$. Then, by Markov property and (11) we get

$$\mathbb{P}(X_{T+S} \neq \overline{X}_{T+S} \mid \mathcal{F}_T) \leq \frac{3|X_T - \overline{X}_T|}{\sqrt{S}} + 2Q_S(\mathbb{1}_{K^c})(X_T).$$

Now we take expectation, and observe that by Lemma 6

$$\begin{aligned}\mathbb{E}\left[Q_S(1_{K^c})(X_T)\right] &\leq \mathrm{TV}(X_T, \mu) + \int_K Q_S(\mathbb{1}_{K^c})\, d\mu \\ &\leq \mathrm{TV}(X_T, \mu) + \sqrt{S}\, \mathcal{H}^{n-1}(\partial K)\, |K|^{-1}.\end{aligned}$$

Putting everything together we get the result. ■

2.5 Proof of the main result

Let S, T be integer multiples of η. Writing

$$\mathrm{TV}(\overline{X}_{T+S}, \mu) \leq \mathrm{TV}(\overline{X}_{T+S}, X_{T+S}) + \mathrm{TV}(X_{T+S}, \mu)$$

and using Proposition 1 and Proposition 3 yields

$$\begin{aligned}\mathrm{TV}(\overline{X}_{T+S}, \mu) \leq\ & C\, (\eta \log(T/\eta))^{1/4}\, n^{3/4}\, M^{1/2}\, T^{1/2}\, S^{-1/2} + 2\, \mathrm{TV}(X_T, \mu) \\ & + 4\, S^{1/2}\, \mathcal{H}^{n-1}(\partial K)\, |K|^{-1}.\end{aligned} \tag{12}$$

For sake of simplicity let us assume that K contains the Euclidean ball of radius 1, and let us aim at a result depending only on the diameter R of K. So we shall use the trivial estimates

$$m \leq R, \quad M \leq \frac{1}{r} \leq 1,$$

together with the less trivial but nevertheless true

$$\mathcal{H}^{n-1}(\partial K) \leq n|K|.$$

Next we use Proposition 2 to bound $\mathrm{TV}(X_T, \mu)$ and (12) becomes

$$\mathrm{TV}(\overline{X}_{T+S}, \mu) \leq C\left((\eta \log(T/\eta))^{1/4}\, n^{3/4}\, T^{1/2}\, S^{-1/2} + e^{-T/2R^2}\right) + 4\, n\, S^{1/2}.$$

Given a small positive constant ε, we have to pick S, T, η so that the right–hand side of the previous inequality equals ε. So we need to take

$$S \approx \frac{\varepsilon^2}{n^2}, \quad T \approx R^2\, \log(1/\varepsilon),$$

and to choose η so that

$$\frac{\eta}{T}\, \log\left(\frac{T}{\eta}\right) \approx \frac{\varepsilon^8}{n^7 R^6 \log(1/\varepsilon)^3}$$

Since for small ξ, ζ we have

$$\xi \log(1/\xi) \approx \zeta \quad \Leftrightarrow \quad \xi \approx \frac{\zeta}{\log(1/\zeta)},$$

and assuming that R and $1/\varepsilon$ are at most polynomial in n, we obtain

$$\eta \approx \frac{\varepsilon^8}{R^4 n^7 \log(n)^3}.$$

To sum up: Let (ξ_k) be a sequence of i.i.d. standard Gaussian vectors, choose the value of η given above and run the algorithm

$$\begin{cases} \overline{X}_0 = 0 \\ \overline{X}_{k+1} = \mathcal{P}_K\left(\overline{X}_k + \sqrt{\eta}\, \xi_{k+1}\right) \end{cases}$$

for a number of steps equal to

$$N = \frac{T+S}{\eta} \approx \frac{R^6\, n^7\, \log(n)^4}{\varepsilon^8}.$$

Then the total variation between $\overline{X}_N$ and the uniform measure on K is at most ε.

3 The general case

In the previous section we viewed LMC (for a constant function f) as a discretization of reflected Brownian motion (X_t) defined by $dX_t = dW_t - \nu_t L(dt)$ and $X_0 = 0$. In this section (X_t) is a slightly more complicated process: it is a diffusion reflected at the boundary of K. More specifically (X_t)

$$\begin{aligned} &X_t \in K, \quad \forall t \geq 0 \\ &dX_t = dW_t - \frac{1}{2}\nabla f(X_t)dt - \nu_t L(dt), \end{aligned} \tag{13}$$

where L is a measure supported on $\{t \geq 0 : \ X_t \in \partial K\}$ and ν_t is an outer unit normal at X_t for any such t. Recall the definition of LMC (1), let us couple it with the continuous process (X_t) as follows. Let (Y_t) be a process constant on each interval $[k\eta, (k+1)\eta)$ and satisfying

$$Y_{(k+1)\eta} = \mathcal{P}_K \left(Y_{k\eta} + W_{(k+1)\eta} - W_{k\eta} - \frac{\eta}{2} \nabla f(Y_{k\eta}) \right), \tag{14}$$

for every integer k. The purpose of this section is to give a bound on the total variation between X_t and its discretization Y_t.

3.1 Mixing time for the continuous process

Since ∇f is assumed to be globally Lipschitz, the existence of the reflected diffusion is insured by [Tanaka, 1979, Theorem 4.1]. Itô's formula then shows that (X_t) is a Markov process whose generator is the operator L

$$Lh = \frac{1}{2}\Delta h - \frac{1}{2}\langle \nabla f, \nabla h \rangle$$

with Neumann boundary condition. Together with Stokes' formula, one can see that the measure

$$\mu(dx) = Z\, e^{-f(x)}\, 1_K(x)\, dx$$

(where Z is the normalization constant) is the unique stationary measure of the process, and that it is even reversible.

We first show that if f is convex the mixing time estimate of the previous section remains valid. Again given a probability measure ν supported on K we let νP_t be the law of X_t when X_0 has law ν.

Lemma 7 *If f is convex then for every $x, x' \in K$*

$$\mathrm{TV}(\delta_x P_t, \delta_{x'} P_t) \leq \frac{|x' - x|}{\sqrt{2\pi t}}.$$

Proof As in the proof of Lemma 4, let (X_t) and (X'_t) be two reflected diffusions starting from x and x' and such that the underlying Brownian motions are coupled by reflection. In addition to (6), one also has

$$\langle \nabla g(X_t - X'_t), \nabla f(X_t) - \nabla f(X'_t) \rangle \geq 0,$$

by convexity of f. The argument then goes through verbatim. ■

As in section 2.3, this lemma allows us to give the following bound on the mixing time of (X_t).

Proposition 4 *For any $t > 0$*

$$\mathrm{TV}(\delta_0 P_t, \mu) \leq C \, \min \left(m\, t^{-1/2}, e^{-t/2R^2} \right),$$

where C is a universal constant.

3.2 A change of measure argument

Again let (X_t) be the reflected diffusion (13). Assume that (X_t) starts from 0 and let (Z_t) be the process

$$Z_t = W_t - \frac{1}{2}\int_0^t \nabla f(X_s)\, ds. \tag{15}$$

Observe that (X_t) solves the Skorokhod problem for (Z_t). Following the same steps as in the previous section we let

$$\overline{Z}_t = Z_{\lfloor t/\eta \rfloor \eta}$$

and we let $(\overline{X}_t)$ be the solution of the Skorokhod problem for $(\overline{Z}_t)$. In other words $(\overline{X}_t)$ is constant on intervals of the form $[k\eta, (k+1)\eta)$ and for every integer k

$$\overline{X}_{(k+1)\eta} = \mathcal{P}_K\left(\overline{X}_{k\eta} + Z_{(k+1)\eta} - Z_{k\eta}\right), \tag{16}$$

Clearly $(\overline{X}_t)$ and (Y_t) are different processes (well, unless the potential f is constant). However, we show in this subsection that using a change of measure trick similar to the one used in Dalalyan [2014], it is possible to bound the total variation distance between $\overline{X}_t$ and Y_t. Recall first the hypothesis made on the potential f

$$|\nabla f(x)| \le L, \quad |\nabla f(x) - \nabla f(y)| \le \beta |x-y|, \quad \forall x, y \in K.$$

Lemma 8 *Let T be an integer multiple of η. Then*

$$\mathrm{TV}(\overline{X}_T, Y_T) \le \frac{\sqrt{L\beta}}{2}\left(\mathbb{E}\left[\int_0^T |X_s - \overline{X}_s|\, ds\right]\right)^{1/2}.$$

Proof Write $T = k\eta$. Given a continuous path $(w_t)_{t\le k\eta}$ we define a map Q from the space of sample paths to $\mathbb{R}$ by setting $Q(w) = x_k$ where (x_i) is defined inductively as

$$\begin{aligned} x_0 &= 0 \\ x_{i+1} &= \mathcal{P}_K\left(x_i + w_{(i+1)\eta} - w_{i\eta} - \frac{\eta}{2}\nabla f(x_i)\right), \quad i \le k-1. \end{aligned}$$

Observe that with this notation we have $Y_{k\eta} = Q((W_t)_{t\le k\eta})$. On the other hand, letting (u_t) be the process

$$u_t = \frac{1}{2}\left(\nabla f(\overline{X}_t) - \nabla f(X_t)\right),$$

letting $\widetilde{W}_t = W_t + \int_0^t u_s\, ds$ and using equation (16), it is easily seen that

$$\overline{X}_{k\eta} = Q\left((\widetilde{W}_t)_{t\le k\eta}\right).$$

This yields the following inequality for the relative entropy of $\overline{X}_{k\eta}$ with respect to $Y_{k\eta}$:

$$\mathrm{H}(\overline{X}_{k\eta} \mid Y_{k\eta}) \le \mathrm{H}\left((\widetilde{W}_t)_{t\le k\eta} \mid (W_t)_{t\le k\eta}\right). \tag{17}$$

Since $\widetilde{W}$ is a Brownian motion plus a drift (observe that the process (u_t) is adapted to the natural filtration of (W_t)) it follows form Girsanov's formula, see for instance Proposition 1 in Lehec [2013], that

$$\begin{aligned}\mathrm{H}\left((\widetilde{W}_t)_{t\le k\eta} \mid (W_t)_{t\le k\eta}\right) &\le \frac{1}{2}\mathbb{E}\left[\int_0^{k\eta} |u_t|^2\, dt\right] \\ &= \frac{1}{8}\mathbb{E}\left[\int_0^{k\eta} |\nabla f(\overline{X}_t) - \nabla f(X_t)|^2\, dt\right].\end{aligned}$$

Plugging this back in (17) and using the hypothesis made on f we get

$$\mathrm{H}(\overline{X}_{k\eta} \mid Y_{k\eta}) \le \frac{L\beta}{4}\mathbb{E}\left[\int_0^{k\eta} |X_t - \overline{X}_t|\, dt\right].$$

We conclude by Pinsker's inequality. ■

The purpose of the next two subsections is to estimate the transportation and total variation distances between X_t and $\overline{X}_t$.

3.3 Estimation of the Wasserstein distance

First we extend Lemma 2 and Lemma 3 to the general case.

Lemma 9 *We have, for all $t > 0$*

$$\mathbb{E}\left[\int_0^t h_K(\nu_s)\, L(ds)\right] \le \frac{(n + RL)t}{2}.$$

Proof As in the proof of Lemma 2, Itô's formula yields

$$2\int_0^t h_K(\nu_s)\, L(ds) = 2\int_0^t \langle X_s, dW_s\rangle - \int_0^t \langle X_s, \nabla f(X_s)\rangle\, ds + nt + |X_0|^2 - |X_t|^2.$$

Assume that $X_0 = 0$, note that the first term is a martingale and observe that $|\langle X_s, \nabla f(X_s)\rangle| \le RL$ by hypothesis. Taking expectation in the previous display, we get the result. ■

Recall the definition of the process (Z_t):

$$Z_t = W_t - \frac{1}{2}\int_0^t \nabla f(X_s)\, ds,$$

and recall that $(\overline{Z}_t)$ is its discretization: $\overline{Z}_t = Z_{\eta\lfloor t/\eta\rfloor}$.

Lemma 10 *There exists a universal constant C such that*

$$\mathbb{E}\left[\sup_{[0,t]} \|Z_s - \overline{Z}_s\|_K\right] \le CMn^{1/2}\eta^{1/2}\log(t/\eta)^{1/2} + \frac{\eta L}{2r}.$$

Proof Since for every $x \in \mathbb{R}^n$

$$\|\nabla f(x)\|_K \leq \frac{1}{r} |\nabla f(x)| \leq \frac{L}{r},$$

we have

$$\begin{aligned} \|Z_t - \overline{Z}_t\|_K &\leq \|W_t - \overline{W}_t\|_K + \frac{1}{2} \int_{\lfloor t/\eta \rfloor \eta}^{t} \|\nabla f(X_t)\|_K \, dt \\ &\leq \|W_t - \overline{W}_t\|_K + \frac{\eta L}{2r}, \end{aligned}$$

for every $t > 0$. Together with Lemma 3, we get the result. ■

As in section 2.2, combining these two lemmas together yields the following estimate.

Proposition 5 *For every time T, we have*

$$\mathbb{E}\left[|X_T - \overline{X}_T|\right] \leq C \left(C_1 \, (\eta \log(T/\eta))^{1/4} \, T^{1/2} + C_2 \, \eta^{1/2} T^{1/2} \right),$$

where C is a universal constant and where

$$\begin{aligned} C_1 = C_1(K, f) &= n^{3/4} M^{1/2} + n^{1/2} R^{1/2} M^{1/2} L^{1/2} \\ C_2 = C_2(K, f) &= n^{1/2} r^{-1/2} L^{1/2} + R^{1/2} r^{-1/2} L. \end{aligned}$$

3.4 From Wasserstein distance to total variation

Unless f is constant, the diffusion (Z_t) does not satisfy Lemma 5 so we need to proceed somewhat differently from what was done in section 2.4. We start with a simple lemma showing that μ does not put too much mass close to the boundary of K.

Lemma 11 *Let $\gamma > 0$. One has*

$$\mu(\{x \in K, d(x, \partial K) \leq \gamma\}) \leq \frac{(n + RL)\gamma}{r}.$$

Proof Define

$$K_\gamma := \{x \in K; \; d(x, \partial K) \geq \gamma\}.$$

Let $\mathbb{B}^n$ be the Euclidean ball, since K contains $r\mathbb{B}^n$ and is convex we have

$$\left(1 - \frac{\gamma}{r}\right) K + \frac{\gamma}{r} r\mathbb{B}^n \subset K,$$

hence

$$\left(1 - \frac{\gamma}{r}\right) K \subset K_\gamma.$$

Clearly this implies:

$$\int_{K_\gamma} e^{-f(x)} \, dx \geq \left(1 - \frac{\gamma}{r}\right)^n \int_K e^{-f((1-\gamma/r)y)} \, dy.$$

Since f is Lipschitz with constant L one also has

$$f\left((1-\gamma/r)y\right) \leq f(y) - \frac{L\gamma|y|}{r} \leq f(y) - \frac{RL\gamma}{r}$$

for every $y \in K$. Combining the last two displays, we obtain

$$\begin{aligned}\int_{K_\gamma} \exp(-f(x))\,dx &\geq \left(1-\frac{\gamma}{r}\right)^n e^{-RL\gamma/r} \int_K e^{-f(x)}\,dx \\ &\geq \left(1 - \frac{n\gamma}{r} - \frac{RL\gamma}{r}\right) \int_K e^{-f(x)}\,dx,\end{aligned}$$

which is the result. ■

Here is a simple bound on the speed of a Brownian motion with drift.

Lemma 12 *Let (W_t) be a standard Brownian motion (starting from 0), let (v_t) an adapted drift satisfying $|v_t| \leq L$ (almost surely), and (Z_t) the process given by*

$$Z_t = W_t + \int_0^t v_s ds.$$

Then for every $t > 0$ and every $\gamma > 0$

$$\mathbb{P}\left(\sup_{s\in[0,t]} |Z_s| > \gamma\right) \leq \frac{\sqrt{nt} + Lt}{\gamma}.$$

Proof By the triangle inequality and since $|v_t| < L$, we have

$$|Z_s| \leq |W_s| + Ls,$$

for any s. Now the process $(|W_s| + Ls)$ is non–negative submartingale so by Doob's maximal inequality

$$\mathbb{P}\left(\sup_{s\in[0,t]} |Z_s| > \gamma\right) \leq \frac{\mathbb{E}\left[|W_t| + Lt\right]}{\gamma}.$$

Since $\mathbb{E}[|W_t|] \leq \sqrt{nt}$, we get the result. ■

Proposition 6 *Let T and S be integer multiples of η. We have*

$$\mathrm{TV}(X_{T+S}, \overline{X}_{T+S}) \leq C\,\left(W(T)S^{-1/2} + \mathrm{TV}(X_T, \mu) + C_3\,S^{1/4} + C_4\,S^{1/2} + C_5\,W(T)^{1/2}\right),$$

where C is a universal constant, $W(T)$ is the bound obtained in Proposition 5 and

$$\begin{aligned}C_3 &= n^{1/4}R^{1/2}r^{-1/2}L^{1/2} + n^{3/4}r^{-1/2} \\ C_4 &= R^{1/2}r^{-1/2}L + n^{1/2}r^{-1/2}L^{1/2} \\ C_5 &= R^{1/2}r^{-1/2}L^{1/2} + n^{1/2}r^{-1/2}.\end{aligned}$$

Proof The proof follows similar lines to those of the proof of Proposition 3, but the drift term requires some additional bounds which will be provided by the previous two lemmas.

We begin with fixing two points $x, x' \in K$ and we consider the two associated diffusions processes (X_t) and (X'_t), which start from the points x and x' respectively, such that the underlying Brownian motions are coupled by reflection. In other words, those processes satisfy equations (4) and (5) with the additional drift term.

In analogy with the process (Z_t), let (Z'_t) be the process

$$Z'_t = W'_s - \frac{1}{2}\int_0^t \nabla f(X'_s)\, ds,$$

let $\overline{Z}'_t = Z'_{\eta\lfloor t/\eta \rfloor}$ and let $(\overline{X}'_t)$ be the solution of the Skorokhod problem for $(\overline{Z}'_t)$. We proceed as in the proof of Proposition 3, letting τ be the coupling time of (X_t) and (X'_t) and letting σ and σ' be the first time (X_t) and (X'_t) hit the boundary of K, we have that

$$\mathbb{P}(X_S \neq \overline{X}'_S) \leq \mathbb{P}(\tau > S) + \mathbb{P}(\sigma \leq S) + \mathbb{P}(\sigma' \leq S).$$

Moreover the coupling time τ still satisfies

$$\mathbb{P}(\tau > S) \leq \frac{|x - x'|}{\sqrt{2\pi S}}.$$

Now fix $\gamma > 0$ and observe that if $d(x, \partial K) > \gamma$, then σ is at least the first time the process

$$W_t - \frac{1}{2}\int_0^t \nabla f(X_s)\, ds$$

hits the sphere centered at x of radius γ. So, by Lemma 12,

$$\mathbb{P}(\sigma \leq S) \leq \frac{\sqrt{nS} + LS}{\gamma} + \mathbb{1}_{\{d(x,\partial K) \leq \gamma\}}.$$

There is a similar inequality for σ' and we obtain

$$\begin{aligned}\mathbb{P}(X_S \neq \overline{X}'_S) &\leq \frac{|x - x'|}{\sqrt{2\pi S}} + \frac{2\sqrt{nS} + 2LS}{\gamma} + \mathbb{1}_{\{d(x,\partial K) \leq \gamma\}} + \mathbb{1}_{\{d(x',\partial K) \leq \gamma\}} \\ &\leq \frac{|x - x'|}{\sqrt{2\pi S}} + \frac{2\sqrt{nS} + 2LS}{\gamma} + 2\,\mathbb{1}_{\{d(x,\partial K) \leq 2\gamma\}} + \mathbb{1}_{\{|x-x'|) \geq \gamma\}}.\end{aligned}$$

So if T and S are two integer multiples of η, if (X_t) and $(\overline{X}_t)$ start from 0, are coupled using the same Brownian motion up to time T, and using the reflection coupling between time T and $T + S$, then we have

$$\begin{aligned}\mathbb{P}(X_{T+S} \neq \overline{X}_{T+S}) \leq \frac{\mathbb{E}\left[|X_T - \overline{X}_T|\right]}{\sqrt{2\pi S}} + \frac{2\sqrt{nS} + 2LS}{\gamma} + 2\,\mathbb{P}\left(d(X_T, \partial K) \leq 2\gamma\right) \\ + \mathbb{P}\left(|X_T - \overline{X}_T| \geq \gamma\right).\end{aligned}$$

By Lemma 11,

$$\begin{aligned}\mathbb{P}\left(d(X_T, \partial K) \le 2\gamma\right) &\le \mu\left(d(x, \partial K) \le 2\gamma\right) + \mathrm{TV}(X_T, \mu) \\ &\le \frac{2(RL+n)\gamma}{r} + \mathrm{TV}(X_T, \mu),\end{aligned}$$

and an application of Markov's inequality gives

$$\mathbb{P}(|X_T - \overline{X}_T| \ge \gamma) \le \frac{\mathbb{E}[|X_T - \overline{X}_T|]}{\gamma}.$$

Combining the last three displays together, we finally obtain

$$\begin{aligned}\mathbb{P}(X_{T+S} \neq \overline{X}_{T+S}) \le \frac{\mathbb{E}\left[|X_T - \overline{X}_T|\right]}{\sqrt{2\pi S}} + \frac{2\sqrt{nS} + 2LS}{\gamma} + \frac{4(RL+n)\gamma}{r} \\ + 2\,\mathrm{TV}(X_T, \mu) + \frac{\mathbb{E}[|X_T - \overline{X}_T|]}{\gamma}.\end{aligned}$$

Optimizing over γ and using Proposition 5 yields the desired inequality. ■

3.5 Proof of Theorem 1

This subsection contains straightforward calculations to help the reader put together the results proven above. Hereafter, to simplify notation, the constants c, C will represent positive universal constants whose value may change between different appearances.

Let T and S be integer multiples of η and write

$$\mathrm{TV}(Y_{T+S}, \mu) \le \mathrm{TV}(Y_{T+S}, \overline{X}_{T+S}) + \mathrm{TV}(\overline{X}_{T+S}, X_{T+S}) + \mathrm{TV}(X_{T+S}, \mu).$$

Again, we will not try to give an optimal result in terms of all the parameters. So assume for simplicity that K contains the Euclidean ball of radius 1 so that r is replaced by 1 in constants C_2, C_3, C_4 and C_5. Also let

$$n_\star = \max(n, RL, R\beta).$$

Keeping in mind that S shall be chosen to be rather small (hence assuming $S \le 1$), Proposition 6 is easily seen to imply that

$$\frac{1}{C}\,\mathrm{TV}(X_{T+S}, \overline{X}_{T+S}) \le W(T) S^{-1/2} + \mathrm{TV}(X_T, \mu) + n_\star\, S^{1/4} + (n_\star\, W(T))^{1/2},$$

Together with Lemma 8 and Proposition 4 we get

$$\frac{1}{C}\,\mathrm{TV}(Y_{T+S}, \mu) \le (L\beta T + n_\star)^{1/2} W(T)^{1/2} + W(T) S^{-1/2} + n_\star\, S^{1/4} + e^{-T/2R^2}.$$

Fix $\varepsilon > 0$ and choose

$$S = n_\star^{-4}\, \varepsilon^4, \quad T = R^2 \log(1/\varepsilon).$$

Then it is easy to see that it is enough to pick η small enough so that

$$W(T) < C n_\star^{-2} \varepsilon^3 \log(1/\varepsilon)^{-1},$$

to ensure $\mathrm{TV}(X_{T+S}, \mu) \leq C\varepsilon$. Now Proposition 5 clearly yields

$$W(T) < C n_* \left(\eta \log(T/\eta)\right)^{1/4} T^{1/2}.$$

Recall that $T = R^2 \log(1/\varepsilon)$ and observe that

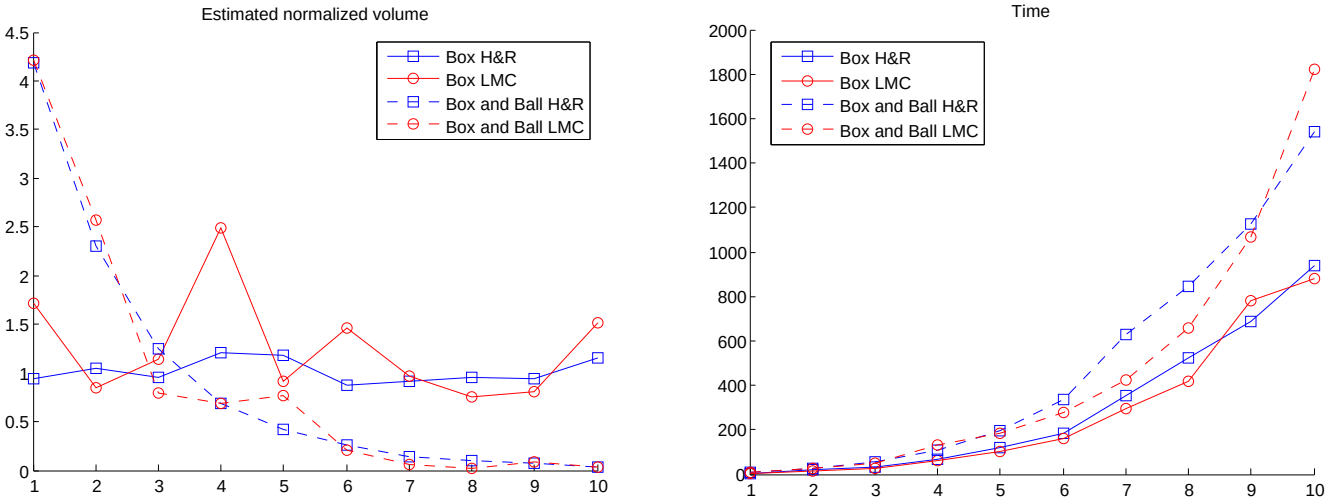
$$\eta \leq c \frac{\varepsilon^{12}}{n_\star^{12} R^4 \max(\log(n), \log(R), \log(1/\varepsilon))^7}$$

suits our purpose. Lastly for this choice of η the number of steps in the algorithm is

$$N = \frac{T+S}{\eta} \leq C \frac{n_\star^{12} R^6 \max(\log(n), \log(R), \log(1/\varepsilon))^8}{\varepsilon^{12}}.$$

4 Experiments

Comparing different Markov Chain Monte Carlo algorithms is a challenging problem in and of itself. Here we choose the following simple comparison procedure based on the volume algorithm developed in Cousins and Vempala [2014]. This algorithm, whose objective is to compute the volume of a given convex set K, procedes in phases. In each phase ℓ it estimates the mean of a certain function under a multivariate Gaussian restricted to K with (unrestricted) covariance $\sigma_\ell \mathrm{I}_n$. Cousins and Vempala provide a Matlab implementation of the entire algorithm, where in each phase the target mean is estimated by sampling from the truncated Gaussian using the hit-and-run (H&R) chain. We implemented the same procedure with LMC instead of H&R, and we choose the step-size $\eta = 1/(\beta n^2)$, where β is the smoothness parameter of the underlying log-concave distribution (in particular here $\beta = 1/\sigma_\ell^2$). The intuition for the choice of the step-size is as follows: the scaling in inverse smoothness comes from the optimization literature, while the scaling in inverse dimension squared comes from the analysis in the unconstrained case in Dalalyan [2014].



We ran the volume algorithm with both H&R and LMC on the following set of convex bodies: $K = [-1, 1]^n$ (referred to as the "Box") and $K = [-1, 1]^n \cap \left(\frac{\sqrt{n}}{2}\mathbb{B}^n\right)$ (referred to as the "Box and Ball"), where $n = 10 \times k, k = 1, \ldots, 10$. The computed volume (normalized by 2^n for the "Box" and by 0.2×2^n for the "Box and Ball") as well as the clock time (in seconds) to terminate are reported in the figure above. From these experiments it seems that LMC and H&R roughly compute similar values for the volume (with H&R being slightly more accurate), and LMC is almost always a bit faster. These results are encouraging, but much more extensive experiments are needed to decide if LMC is indeed a competitor to H&R in practice.

References

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. In *ICML 2012*, 2012.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n). In *Advances in Neural Information Processing Systems 26 (NIPS)*, pages 773–781. 2013.

B. Cousins and S. Vempala. Bypassing kls: Gaussian cooling and an $o^*(n^3)$ volume algorithm. *Arxiv preprint arXiv:1409.6011*, 2014.

A. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Arxiv preprint arXiv:1412.7392*, 2014.

M. Dyer, A. Frieze, and R. Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *Journal of the ACM (JACM)*, 38(1):1–17, 1991.

R. Kannan and H. Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37:1–20, 2012.

J. Lehec. Representation formula for the entropy and functional inequalities. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(3):885–889, 2013.

David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005, 2006.

L. Lovász and S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

G. Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM J. Control and Optimization*, 24(4):655–666, 1986.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

A. Skorokhod. Stochastic equations for diffusion processes in a bounded region. *Theory of Probability & Its Applications*, 6(3):264–274, 1961.

H. Tanaka. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Mathematical Journal*, 9(1):163–177, 1979.

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

M. Welling and Y.W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML 2011*, 2011.