CREST
CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

Série des Documents de Travail

n° 2017-21

Optimal Kullback-Leibler Aggregation in Mixture Estimation by Maximum Likelihood

A. DALALYAN[1]
M. SEBBAR[2]

[1] CREST ; ENSAE
[2] CREST ; ENSAE

Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent

Arnak S. Dalalyan ARNAK.DALALYAN@ENSAE.FR
ENSAE/CREST/Université Paris Saclay

Abstract

In this paper[1], we revisit the recently established theoretical guarantees for the convergence of the Langevin Monte Carlo algorithm of sampling from a smooth and (strongly) log-concave density. We improve the existing results when the convergence is measured in the Wasserstein distance and provide further insights on the very tight relations between, on the one hand, the Langevin Monte Carlo for sampling and, on the other hand, the gradient descent for optimization. Finally, we also establish guarantees for the convergence of a version of the Langevin Monte Carlo algorithm that is based on noisy evaluations of the gradient.

Keywords: Markov Chain Monte Carlo, Approximate sampling, Rates of convergence, Langevin algorithm, Gradient descent

1. Introduction

Let p be a positive integer and $f : \mathbb{R}^p \to \mathbb{R}$ be a measurable function such that the integral $\int_{\mathbb{R}^p} \exp\{-f(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}$ is finite. In various applications, one is faced with the problems of finding the minimum point of f or computing the average with respect to the probability density

$$\pi(\boldsymbol{\theta}) = \frac{e^{-f(\boldsymbol{\theta})}}{\int_{\mathbb{R}^p} e^{-f(\boldsymbol{u})}\, d\boldsymbol{u}}.$$

In other words, one often looks for approximating the values $\boldsymbol{\theta}^*$ and $\bar{\boldsymbol{\theta}}$ defined as

$$\bar{\boldsymbol{\theta}} = \int_{\mathbb{R}^p} \boldsymbol{\theta}\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad \boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta}).$$

In most situations, the approximations of these values are computed using iterative algorithms which share many common features. There is a vast variety of such algorithms for solving both tasks, see for example (Boyd and Vandenberghe, 2004) for optimization and (Atchadé et al., 2011) for approximate sampling. The similarities between the task of optimization and that of averaging have been recently exploited in the papers (Dalalyan, 2014; Durmus and Moulines, 2016; Durmus et al., 2016) in order to establish fast and accurate theoretical guarantees for sampling from and averaging with respect to the density π using the Langevin Monte Carlo algorithm. The goal of the present work is to push further this study both by improving the existing bounds and by extending them in some directions.

1. This paper has been published in proceedings of COLT 2017. However, this version is more recent. We have corrected some typos ($2/(m+M)$ instead of $1/(m+M)$ on pages 3-4) and slightly improved the upper bound of Theorem 3.

We will focus on strongly convex functions f having a Lipschitz continuous gradient. That is, we assume that there exist two positive constants m and M such that

$$\begin{cases} f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}') - \nabla f(\boldsymbol{\theta}')^\top (\boldsymbol{\theta} - \boldsymbol{\theta}') \geq (m/2)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\boldsymbol{\theta}')\|_2 \leq M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{cases} \qquad \forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^p, \tag{1}$$

where ∇f stands for the gradient of f and $\|\cdot\|_2$ is the Euclidean norm. We say that the density $\pi(\boldsymbol{\theta}) \propto e^{-f(\boldsymbol{\theta})}$ is log-concave (resp. strongly log-concave) if the function f satisfies the first inequality of (1) with $m = 0$ (resp. $m > 0$).

The Langevin Monte Carlo (LMC) algorithm studied throughout this work is the analogue of the gradient descent algorithm for optimization. Starting from an initial point $\boldsymbol{\vartheta}^{(0)} \in \mathbb{R}^p$ that may be deterministic or random, the iterations of the algorithm are defined by the update rule

$$\boldsymbol{\vartheta}^{(k+1,h)} = \boldsymbol{\vartheta}^{(k,h)} - h\nabla f(\boldsymbol{\vartheta}^{(k,h)}) + \sqrt{2h}\,\boldsymbol{\xi}^{(k+1)}; \qquad k = 0, 1, 2, \ldots \tag{2}$$

where $h > 0$ is a tuning parameter, referred to as the step-size, and $\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(k)}, \ldots$ is a sequence of mutually independent, and independent of $\boldsymbol{\vartheta}^{(0)}$, centered Gaussian vectors with covariance matrices equal to identity. Under the assumptions imposed on f, when h is small and k is large (so that the product kh is large), the distribution of $\boldsymbol{\vartheta}^{(k,h)}$ is close in various metrics to the distribution with density $\pi(\boldsymbol{\theta})$, hereafter referred to as the target distribution. An important question is to quantify this closeness; this might be particularly useful for deriving a stopping rule for the LMC algorithm.

The measure of approximation used in this paper is the Wasserstein-Monge-Kantorovich distance W_2. For two measures μ and ν defined on $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$, W_2 is defined by

$$W_2(\mu, \nu) = \Big(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2^2 \, d\gamma(\boldsymbol{\theta}, \boldsymbol{\theta}') \Big)^{1/2},$$

where the $\inf$ is with respect to all joint distributions γ having μ and ν as marginal distributions. This distance is perhaps more suitable for quantifying the quality of approximate sampling schemes than other metrics such as the total variation. Indeed, on the one hand, bounds on the Wasserstein distance—unlike the bounds on the total-variation distance—directly provide the level of approximating the first order moment. For instance, if μ and ν are two Dirac measures at the points $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, respectively, then the total-variation distance $D_{\mathrm{TV}}(\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{\theta}'})$ equals one whenever $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$, whereas $W_2(\delta_{\boldsymbol{\theta}}, \delta_{\boldsymbol{\theta}'}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$ is a smoothly increasing function of the Euclidean distance between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This seems to better correspond to the intuition on the closeness of two distributions.

2. Improved guarantees for the Wasserstein distance

The rationale behind the LMC algorithm (2) is simple: the Markov chain $\{\boldsymbol{\vartheta}^{(k,h)}\}_{k\in\mathbb{N}}$ is the Euler discretization of a continuous-time diffusion process $\{\boldsymbol{L}_t : t \in \mathbb{R}_+\}$, known as Langevin diffusion, that has π as invariant density (Bhattacharya, 1978, Thm. 3.5). The Langevin diffusion is defined by the stochastic differential equation

$$d\boldsymbol{L}_t = -\nabla f(\boldsymbol{L}_t)\, dt + \sqrt{2}\, d\boldsymbol{W}_t, \qquad t \geq 0, \tag{3}$$

where $\{\boldsymbol{W}_t : t \geq 0\}$ is a p-dimensional Brownian motion. When f satisfies condition (1), equation (3) has a unique strong solution which is a Markov process. Let ν_k be the distribution of the k-th iterate of the LMC algorithm, that is $\vartheta^{(k,h)} \sim \nu_k$.

Theorem 1 *Assume that $h \in (0, 2/M)$. The following claims hold:*

(a) *If $h \leq 2/(m+M)$ then* $W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + 1.82(M/m)(hp)^{1/2}$.

(b) *If $h \geq 2/(m+M)$ then* $W_2(\nu_K, \pi) \leq (Mh - 1)^K W_2(\nu_0, \pi) + 1.82\dfrac{Mh}{2 - Mh}(hp)^{1/2}$.

The proof of this theorem is postponed to Section 6. We content ourselves here by discussing the relation of this result to previous work. Note that if the initial value $\boldsymbol{\vartheta}^{(0)} = \boldsymbol{\theta}^{(0)}$ is deterministic then, according to (Durmus and Moulines, 2016, Theorem 1), we have

$$\begin{aligned} W_2(\nu_0, \pi)^2 &= \int_{\mathbb{R}^p} \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &= \|\boldsymbol{\theta}^{(0)} - \bar{\boldsymbol{\theta}}\|_2^2 + \int_{\mathbb{R}^p} \|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &\leq \|\boldsymbol{\theta}^{(0)} - \bar{\boldsymbol{\theta}}\|_2^2 + p/m. \end{aligned} \tag{4}$$

First of all, let us remark that if we choose h and K so that

$$h \leq 2/(m+M), \qquad e^{-mhK} W_2(\nu_0, \pi) \leq \varepsilon/2, \quad 1.82(M/m)(hp)^{1/2} \leq \varepsilon/2, \tag{5}$$

then we have $W_2(\nu_K, \pi) \leq \varepsilon$. In other words, conditions (5) are sufficient for the density of the output of the LMC algorithm with K iterations to be within the precision ε of the target density when the precision is measured using the Wasserstein distance. This readily yields

$$h \leq \frac{m^2\varepsilon^2}{14M^2p} \wedge \frac{2}{m+M} \quad \text{and} \quad hK \geq \frac{1}{m} \log\Big(\frac{2(\|\boldsymbol{\theta}^{(0)} - \bar{\boldsymbol{\theta}}\|_2^2 + p/m)^{1/2}}{\varepsilon}\Big)$$

Assuming m, M and $\|\boldsymbol{\theta}^{(0)} - \bar{\boldsymbol{\theta}}\|_2^2/p$ to be constants, we can deduce from the last display that it suffices $K = Cp\varepsilon^{-2}\log(p/\varepsilon)$ number of iterations in order to reach the precision level ε. This fact has been first established in (Dalalyan, 2014) for the LMC algorithm with a warm start and the total-variation distance. It was later improved by Durmus and Moulines (2016), who showed that the same result holds for any starting point and established similar bounds for the Wasserstein distance.

In order to make the comparison easier, let us recall below the corresponding result from[2] (Durmus and Moulines, 2016). It asserts that under condition (1), if $h \leq 2/(m+M)$ then

$$W_2^2(\nu_K, \pi) \leq 2\Big(1 - \frac{mMh}{m+M}\Big)^K W_2^2(\nu, \pi) + \frac{Mhp}{m}(m+M)\Big(h + \frac{m+M}{2mM}\Big)\Big(2 + \frac{M^2h}{m} + \frac{M^2h^2}{6}\Big). \tag{6}$$

When we compare this inequality with the claims of Theorem 1, we see that

i) Theorem 1 holds under weaker conditions: $h \leq 2/M$ instead of $h \leq 2/(m+M)$.

ii) The analytical expressions of the upper bounds on the Wasserstein distance in Theorem 1 are not as involved as those of (6).

2. We slightly adapt the original result taking into account the fact that we are dealing with the LMC algorithm with a constant step.

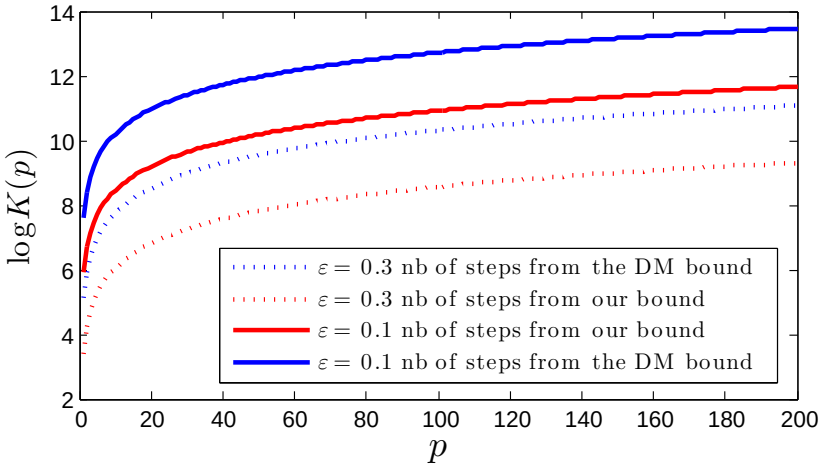

Figure 1: The curves of the functions $p \mapsto \log K(p)$, where $K(p)$ is the number of steps— derived either from our bound or from the bound (6) of (Durmus and Moulines, 2016)—sufficing for reaching the precision level ε (for $\varepsilon = 0.1$ and $\varepsilon = 0.3$).

iii) If we take a closer look, we can check that when $h \leq 2/(m+M)$, the upper bound in part (a) of Theorem 1 is sharper than that of (6).

In order to better illustrate the claim in iii) above, we consider a numerical example in which $m = 4$, $M = 5$ and $\|\boldsymbol{\theta}^{(0)} - \bar{\boldsymbol{\theta}}\|_2^2 = p$. Let $F_{\rm our}(h, K, p)$ and $F_{\rm DM}(h, K, p)$ be the upper bounds on $W_2(\nu_K, \pi)$ provided by Theorem 1 and (6). For different values of p, we compute

$$K_{\rm our}(p) = \min\big\{K : \text{ there exists } h \leq 2/(m+M) \text{ such that } F_{\rm our}(h, K, p) \leq \varepsilon\big\},$$
$$K_{\rm DM}(p) = \min\big\{K : \text{ there exists } h \leq 2/(m+M) \text{ such that } F_{\rm DM}(h, K, p) \leq \varepsilon\big\}.$$

The curves of the functions $p \mapsto \log K_{\rm our}(p)$ and $p \mapsto \log K_{\rm DM}(p)$, for $\varepsilon = 0.1$ and $\varepsilon = 0.3$ are plotted in Figure 1. We can deduce from these plots that the number of iterations yielded by our bound is more than 5 times smaller than the number of iterations recommended by bound (6) of Durmus and Moulines (2016).

Remark 2 *Although the upper bound on $W_2(\nu_0, \pi)$ provided by (4) is relevant for understanding the order of magnitude of $W_2(\nu_0, \pi)$, it has limited applicability since the distance $\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|$ might*

be hard to evaluate. An attractive alternative to that bound is the following[3]*:*

$$\begin{aligned} W_2(\nu_0,\pi)^2 &= \int_{\mathbb{R}^p} \|\boldsymbol{\theta}^{(0)}-\boldsymbol{\theta}\|_2^2 \pi(d\boldsymbol{\theta}) \\ &\le \frac{2}{m}\int_{\mathbb{R}^p} \Big(f(\boldsymbol{\theta}_0)-f(\boldsymbol{\theta})-\nabla f(\boldsymbol{\theta})^\top(\boldsymbol{\theta}_0-\boldsymbol{\theta})\Big)\pi(d\boldsymbol{\theta}) \\ &= \frac{2}{m}\Big(f(\boldsymbol{\theta}_0)-\int_{\mathbb{R}^p} f(\boldsymbol{\theta})\,\pi(d\boldsymbol{\theta})+p\Big). \end{aligned}$$

If f is lower bounded by some known constant, for instance if $f \ge 0$, the last inequality provides the computable upper bound $W_2(\nu_0,\pi)^2 \le \frac{2}{m}\big(f(\boldsymbol{\theta}_0)+p\big)$.

3. Relation with optimization

We have already mentioned that the LMC algorithm is very close to the gradient descent algorithm for computing the minimum $\boldsymbol{\theta}^*$ of the function f. However, when we compare the guarantees of Theorem 1 with those available for the optimization problem, we remark the following striking difference. The approximate computation of $\boldsymbol{\theta}^*$ requires a number of steps of the order of $\log(1/\varepsilon)$ to reach the precision ε, whereas, for reaching the same precision in sampling from π, the LMC algorithm needs a number of iterations proportional to $(p/\varepsilon^2)\log(p/\varepsilon)$. The goal of this section is to explain that this, at first sight very disappointing behavior of the LMC algorithm is, in fact, continuously connected to the exponential convergence of the gradient descent.

The main ingredient for the explanation is that the function $f(\boldsymbol{\theta})$ and the function $f_\tau(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/\tau$ have the same point of minimum $\boldsymbol{\theta}^*$, whatever the real number $\tau > 0$. In addition, if we define the density function $\pi_\tau(\boldsymbol{\theta}) \propto \exp\big(-f_\tau(\boldsymbol{\theta})\big)$, then the average value

$$\bar{\boldsymbol{\theta}}_\tau = \int_{\mathbb{R}^p} \boldsymbol{\theta}\,\pi_\tau(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

tends to the minimum point $\boldsymbol{\theta}^*$ when τ goes to zero. Furthermore, the distribution $\pi_\tau(d\boldsymbol{\theta})$ tends to the Dirac measure at $\boldsymbol{\theta}^*$. Clearly, f_τ satisfies (1) with the constants $m_\tau = m/\tau$ and $M_\tau = M/\tau$. Therefore, on the one hand, we can apply to π_τ claim (a) of Theorem 1, which tells us that if we choose $h = 1/M_\tau = \tau/M$, then

$$W_2(\nu_K,\pi_\tau) \le \Big(1-\frac{m}{M}\Big)^K W_2(\delta_{\boldsymbol{\theta}^{(0)}},\pi_\tau) + 2\Big(\frac{M}{m}\Big)\Big(\frac{p\tau}{M}\Big)^{1/2}. \tag{7}$$

On the other hand, the LMC algorithm with the step-size $h = \tau/M$ applied to f_τ reads as

$$\boldsymbol{\vartheta}^{(k+1,h)} = \boldsymbol{\vartheta}^{(k,h)} - \frac{1}{M}\nabla f(\boldsymbol{\vartheta}^{(k,h)}) + \sqrt{\frac{2\tau}{M}}\,\boldsymbol{\xi}^{(k+1)}; \qquad k = 0,1,2,\dots \tag{8}$$

When the parameter τ goes to zero, the LMC sequence (8) tends to the gradient descent sequence $\boldsymbol{\theta}^{(k)}$. Therefore, the limiting case of (7) corresponding to $\tau \to 0$ writes as

$$\|\boldsymbol{\theta}^{(K)}-\boldsymbol{\theta}^*\|_2 \le \Big(1-\frac{m}{M}\Big)^K \|\boldsymbol{\theta}^{(0)}-\boldsymbol{\theta}^*\|_2,$$

which is a well-known result in Optimization. This clearly shows that Theorem 1 is a natural extension of the results of convergence from optimization to sampling.

3. The second line follows from strong convexity whereas the third line is a consequence of the two identities $\int_{\mathbb{R}^p}\nabla f(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) = 0$ and $\int_{\mathbb{R}^p}\boldsymbol{\theta}^\top\nabla f(\boldsymbol{\theta})\pi(d\boldsymbol{\theta}) = p$. These identities follow from the fundamental theorem of calculus and the integration by parts formula, respectively.

4. Guarantees for the noisy gradient version

In some situations, the precise evaluation of the gradient $\nabla f(\boldsymbol{\theta})$ is computationally expensive or practically impossible, but it is possible to obtain noisy evaluations of ∇f at any point. This is the setting considered in the present section. More precisely, we assume that at any point $\boldsymbol{\vartheta}^{(k,h)} \in \mathbb{R}^p$ of the LMC algorithm, we can observe the value

$$\boldsymbol{Y}^{(k,h)} = \nabla f(\boldsymbol{\vartheta}^{(k,h)}) + \sigma\, \boldsymbol{\zeta}^{(k)},$$

where $\{\boldsymbol{\zeta}^{(k)} : k = 0, 1, \ldots\}$ is a sequence of independent zero mean random vectors such that $\mathbf{E}[\|\boldsymbol{\zeta}^{(k)}\|_2^2] \le p$ and $\sigma > 0$ is a deterministic noise level. Furthermore, the noise vector $\boldsymbol{\zeta}^{(k)}$ is independent of the past states $\boldsymbol{\vartheta}^{(1,h)}, \ldots, \boldsymbol{\vartheta}^{(k,h)}$. The noisy LMC (nLMC) algorithm is then defined as

$$\boldsymbol{\vartheta}^{(k+1,h)} = \boldsymbol{\vartheta}^{(k,h)} - h\boldsymbol{Y}^{(k,h)} + \sqrt{2h}\, \boldsymbol{\xi}^{(k+1)}; \qquad k = 0, 1, 2, \ldots \tag{9}$$

where $h > 0$ and $\boldsymbol{\xi}^{(k+1)}$ are as in (2). The next theorem extends the guarantees of Theorem 1 to the noisy-gradient setting and to the nLMC algorithm.

Theorem 3 *Let $\boldsymbol{\vartheta}^{(K,h)}$ be the K-th iterate of the nLMC algorithm (9) and ν_K be its distribution. If the function f satisfies condition (1) and $h \le 2/M$ then the following claims hold:*

(a) *If $h \le 2/(m+M)$ then*

$$W_2(\nu_K, \pi) \le \Big(1 - \frac{mh}{2}\Big)^K W_2(\nu_0, \pi) + \Big(\frac{2hp}{m}\Big)^{1/2} \Big\{\sigma^2 + \frac{3.3M^2}{m}\Big\}^{1/2}. \tag{10}$$

(b) *If $h \ge 2/(m+M)$ then*

$$W_2(\nu_K, \pi) \le \Big(\frac{Mh}{2}\Big)^K W_2(\nu_0, \pi) + \Big(\frac{2h^2p}{2 - Mh}\Big)^{1/2} \Big\{\sigma^2 + \frac{6.6M}{2 - Mh}\Big\}^{1/2}.$$

To understand the potential scope of applicability of this result, let us consider a typical statistical problem in which $f(\boldsymbol{\theta})$ is the negative log-likelihood of n independent random variables $X_1, \ldots, X_n$. Then, if $\ell(\boldsymbol{\theta}, x)$ is the log-likelihood of one variable, we have

$$f(\boldsymbol{\theta}) = \sum_{i=1}^n \ell(\boldsymbol{\theta}, X_i).$$

In such a situation, if the Fisher information is not degenerated, both m and M are proportional to the sample size n. When the gradient of $\ell(\boldsymbol{\theta}, X_i)$ with respect to parameter $\boldsymbol{\theta}$ is hard to compute, one can replace the evaluation of $\nabla f(\boldsymbol{\vartheta}^{(k,h)})$ at each step k by that of $Y_k = n\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\vartheta}^{(k,h)}, X_k)$. Under suitable assumptions, this random vector satisfies the conditions of Theorem 3 with a σ^2 proportional to n. Therefore, if we analyze the expression between curly brackets in (10), we see that the additional term, σ^2, due to the subsampling is of the same order of magnitude as the term $3.3M^2/m$. Thus, using the subsampled gradient in the LMC algorithm does not cause a significant deterioration of the precision while reducing considerably the computational burden.

5. Discussion and outlook

We have established simple guarantees for the convergence of the Langevin Monte Carlo algorithm under the Wasserstein metric. These guarantees are valid under strong convexity and Lipschitz-gradient assumptions on the log-density function, for a step-size smaller than $2/M$, where M is the constant in the Lipschitz condition. These guarantees are sharper than previously established analogous results and in perfect agreement with the analogous results in Optimization. Furthermore, we have shown that similar results can be obtained in the case where only noisy evaluations of the gradient are possible.

There are a number of interesting directions in which this work can be extended. One relevant and closely related problem is the approximate computation of the volume of a convex body, or, the problem of sampling from the uniform distribution on a convex body. This problem has been analyzed by other Monte Carlo methods such as "Hit and Run" in a series of papers by Lovász and Vempala (2006b,a), see also the more recent paper (Bubeck et al., 2015). Numerical experiments reported in (Bubeck et al., 2015) suggest that the LMC algorithm might perform better in practice than "Hit and Run". It would be interesting to have a theoretical result corroborating this observation.

Other interesting avenues for future research include the possible adaptation of the Nesterov acceleration to the problem of sampling, extensions to second-order methods as well as the alleviation of the strong-convexity assumptions. We also plan to investigate in more depth the applications is high-dimensional statistics (see, for instance, Dalalyan and Tsybakov (2012)). Some results in these directions are already obtained in (Dalalyan, 2014; Durmus and Moulines, 2016; Durmus et al., 2016). It is a stimulating question whether we can combine ideas of the present work and the aforementioned earlier results to get improved guarantees.

6. Proofs

The first part of the proofs of Theorem 1 and Theorem 3 is the same. We start this section by this common part and then we proceed with the proofs of the two theorems separately.

Let $\boldsymbol{W}$ be a p-dimensional Brownian Motion such that $\boldsymbol{W}_{(k+1)h} - \boldsymbol{W}_{kh} = \sqrt{h}\,\boldsymbol{\xi}^{(k+1)}$. We define the stochastic process $\boldsymbol{L}$ so that $\boldsymbol{L}_0 \sim \pi$ and

$$\boldsymbol{L}_t = \boldsymbol{L}_0 - \int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,\boldsymbol{W}_t, \qquad \forall\, t > 0. \tag{11}$$

It is clear that this equation implies that

$$\begin{aligned}\boldsymbol{L}_{(k+1)h} &= \boldsymbol{L}_{kh} - \int_{kh}^{(k+1)h} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\,(\boldsymbol{W}_{(k+1)h} - \boldsymbol{W}_{kh}) \\ &= \boldsymbol{L}_{kh} - \int_{kh}^{(k+1)h} \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2h}\,\boldsymbol{\xi}^{(k+1)}.\end{aligned}$$

Furthermore, $\{\boldsymbol{L}_t : t \geq 0\}$ is a diffusion process having π as the stationary distribution. Since the initial value $\boldsymbol{L}_0$ is drawn from π, we have $\boldsymbol{L}_t \sim \pi$ for every $t \geq 0$.

Let us denote $\boldsymbol{\Delta}_k = \boldsymbol{L}_{kh} - \boldsymbol{\vartheta}^{(k,h)}$ and $I_k = (kh, (k+1)h]$. We have

$$\begin{aligned}\boldsymbol{\Delta}_{k+1} &= \boldsymbol{\Delta}_k + h\boldsymbol{Y}^{(k,h)} - \int_{I_k} \nabla f(\boldsymbol{L}_t)\, dt \\ &= \boldsymbol{\Delta}_k - h\big(\underbrace{\nabla f(\boldsymbol{\vartheta}^{(k,h)} + \boldsymbol{\Delta}_k) - \nabla f(\boldsymbol{\vartheta}^{(k,h)})}_{:=\boldsymbol{U}_k}\big) + \sigma h \boldsymbol{\zeta}^{(k)} - \underbrace{\int_{I_k} \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_{kh})\big)\, dt}_{:=\boldsymbol{V}_k}.\end{aligned}$$

In view of the triangle inequality, we get

$$\|\boldsymbol{\Delta}_{k+1}\|_2 \le \|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k + \sigma h \boldsymbol{\zeta}^{(k)}\|_2 + \|\boldsymbol{V}_k\|_2. \tag{12}$$

For the first norm in the right hand side, we can use the following inequalities:

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k + \sigma h \boldsymbol{\zeta}^{(k)}\|_2^2] &= \mathbf{E}[\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k\|_2^2] + \mathbf{E}[\|\sigma h \boldsymbol{\zeta}^{(k)}\|_2^2] \\ &= \mathbf{E}[\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k\|_2^2] + \sigma^2 h^2 p. \end{aligned}\tag{13}$$

We need now three technical lemmas the proofs of which are postponed to Section 6.3.

Lemma 1 *Let us introduce the constant γ that equals $|1 - mh|$ if $h \le 2/(m+M)$ and $|1 - Mh|$ if $h \ge 2/(m+M)$. (Since $h \in (0, 2/M)$, this value γ satisfies $0 < \gamma < 1$). It holds that*

$$\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k\|_2 \le \gamma \|\boldsymbol{\Delta}_k\|_2. \tag{14}$$

Lemma 2 *If the function f is continuously differentiable and the gradient of f is Lipschitz with constant M, then*

$$\int_{\mathbb{R}^p} \|\nabla f(\boldsymbol{x})\|_2^2\, \pi(\boldsymbol{x})\, d\boldsymbol{x} \le Mp.$$

Lemma 3 *If the function f has a Lipschitz-continuous gradient with the Lipschitz constant M, $\boldsymbol{L}$ is the Langevin diffusion (11) and $\boldsymbol{V}(a) = \int_a^{a+h} \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_a)\big)\, dt$ for some $a \ge 0$, then*

$$\big(\mathbf{E}[\|\boldsymbol{V}(a)\|_2^2]\big)^{1/2} \le \left(\frac{1}{3} h^4 M^3 p\right)^{1/2} + (h^3 p)^{1/2} M.$$

This completes the common part of the proof. We present below the proofs of the theorems.

6.1. Proof of Theorem 1

Using (12) with $\sigma = 0$ and Lemma 1, we get

$$\|\boldsymbol{\Delta}_{k+1}\|_2 \le \gamma \|\boldsymbol{\Delta}_k\|_2 + \|\boldsymbol{V}_k\|_2, \qquad \forall k \in \mathbb{N}.$$

In view of the Minkowski inequality and Lemma 3, this yields

$$\begin{aligned}(\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2])^{1/2} &\le \gamma (\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^2])^{1/2} + (\mathbf{E}[\|\boldsymbol{V}_k\|_2^2])^{1/2} \\ &\le \gamma (\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^2])^{1/2} + 1.82 (h^3 M^2 p)^{1/2},\end{aligned}$$

where we have used the fact that $h \leq 2/M$. Using this inequality iteratively with $k-1, \dots, 0$ instead of k, we get

$$\begin{aligned}(\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2])^{1/2} &\leq \gamma^{k+1}(\mathbf{E}[\|\boldsymbol{\Delta}_0\|_2^2])^{1/2} + 1.82(h^3M^2p)^{1/2}\sum_{j=0}^{k}\gamma^j \\ &\leq \gamma^{k+1}(\mathbf{E}[\|\boldsymbol{\Delta}_0\|_2^2])^{1/2} + 1.82(h^3M^2p)^{1/2}(1-\gamma)^{-1}. \end{aligned} \tag{15}$$

Since $\boldsymbol{\Delta}_{k+1} = \boldsymbol{L}_{(k+1)h} - \boldsymbol{\vartheta}^{(k+1,h)}$ and $\boldsymbol{L}_{(k+1)h} \sim \pi$, we readily get the inequality $W_2(\nu_{k+1}, \pi) \leq \big(\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2]\big)^{1/2}$. In addition, one can choose $\boldsymbol{L}_0$ so that $W_2(\nu_0, \pi) = \big(\mathbf{E}[\|\boldsymbol{\Delta}_0\|_2^2]\big)^{1/2}$. Using these relations and substituting γ by its expression in (15), we get the two claims of the theorem.

6.2. Proof of Theorem 3

Using (12), (13) and Lemma 1, we get (for every $t > 0$)

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2] &= \mathbf{E}[\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k + \boldsymbol{V}_k\|_2^2] + \mathbf{E}[\|\sigma h\boldsymbol{\zeta}^{(k)}\|_2^2] \\ &\leq (1+t)\mathbf{E}[\|\boldsymbol{\Delta}_k - h\boldsymbol{U}_k\|_2^2] + (1+t^{-1})\mathbf{E}[\|\boldsymbol{V}_k\|_2^2] + \sigma^2h^2p \\ &\leq (1+t)\gamma^2\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^2] + (1+t^{-1})\mathbf{E}[\|\boldsymbol{V}_k\|_2^2] + \sigma^2h^2p.\end{aligned}$$

Since $h \leq 2/M$, Lemma 3 implies that

$$\mathbf{E}[\|\boldsymbol{\Delta}_{k+1}\|_2^2] \leq (1+t)\gamma^2\mathbf{E}[\|\boldsymbol{\Delta}_k\|_2^2] + (1+t^{-1})(1.82)^2h^3M^2p + \sigma^2h^2p$$

for every $t > 0$. Let us choose $t = (\frac{1+\gamma}{2\gamma})^2 - 1$ so that $(1+t)\gamma^2 = (\frac{1+\gamma}{2})^2$. By recursion, this leads to

$$W_2^2(\nu_{k+1}, \pi) \leq \Big(\frac{1+\gamma}{2}\Big)^{2(k+1)}W_2^2(\nu_0, \pi) + \Big(\frac{2}{1-\gamma}\Big)\Big\{\sigma^2h^2p + (1+t^{-1})(1.82)^2h^3M^2p\Big\}.$$

In the case $h \leq 2/(m+M)$, $\gamma = 1 - mh$ and we get $\frac{1+\gamma}{2} = 1 - \frac{1}{2}mh$. Furthermore,

$$(1+t^{-1})h^3M^2p = \frac{(1+\gamma)^2h^3M^2p}{(1-\gamma)(1+3\gamma)} \leq \frac{h^2M^2p}{m}.$$

This readily yields

$$W_2(\nu_{k+1}, \pi) \leq \Big(1 - \frac{mh}{2}\Big)^{k+1}W_2(\nu_0, \pi) + \Big(\frac{2hp}{m}\Big)^{1/2}\Big\{\sigma^2 + \frac{3.3M^2}{m}\Big\}^{1/2}.$$

Similarly, in the case $h \geq 2/(m+M)$, $\gamma = Mh - 1$ and we get $\frac{1+\gamma}{2} = \frac{1}{2}Mh$. Furthermore,

$$(1+t^{-1})h^3M^2p = \frac{(1+\gamma)^2h^3M^2p}{(1-\gamma)(1+3\gamma)} \leq \frac{h^3M^2p}{2-Mh} \leq \frac{2h^2Mp}{2-Mh}.$$

This implies the inequality

$$W_2(\nu_{k+1}, \pi) \leq \Big(\frac{Mh}{2}\Big)^{k+1}W_2(\nu_0, \pi) + \Big(\frac{2h^2p}{2-Mh}\Big)^{1/2}\Big\{\sigma^2 + \frac{6.6M}{2-Mh}\Big\}^{1/2},$$

which completes the proof.

6.3. Proofs of lemmas

Proof [Proof of Lemma 1] Since f is m-strongly convex, it satisfies the inequality

$$\boldsymbol{\Delta}^\top\big(\nabla f(\boldsymbol{\vartheta}+\boldsymbol{\Delta})-\nabla f(\boldsymbol{\vartheta})\big)\geq\frac{mM}{m+M}\|\boldsymbol{\Delta}\|_2^2+\frac{1}{m+M}\|\nabla f(\boldsymbol{\vartheta}+\boldsymbol{\Delta})-\nabla f(\boldsymbol{\vartheta})\|_2^2,$$

for all $\boldsymbol{\Delta},\boldsymbol{\vartheta}\in\mathbb{R}^p$. Therefore, simple algebra yields

$$\begin{aligned}\|\boldsymbol{\Delta}_k-h\boldsymbol{U}_k\|_2^2&=\|\boldsymbol{\Delta}_k\|_2^2-2h\boldsymbol{\Delta}_k^\top\boldsymbol{U}_k+h^2\|\boldsymbol{U}_k\|_2^2\\&=\|\boldsymbol{\Delta}_k\|_2^2-2h\boldsymbol{\Delta}_k^\top\big(\nabla f(\boldsymbol{\vartheta}^{(k,h)}+\boldsymbol{\Delta}_k)-\nabla f(\boldsymbol{\vartheta}^{(k,h)})\big)+h^2\|\boldsymbol{U}_k\|_2^2\\&\leq\|\boldsymbol{\Delta}_k\|_2^2-\frac{2hmM}{m+M}\|\boldsymbol{\Delta}_k\|_2^2-\frac{2h}{m+M}\|\boldsymbol{U}_k\|_2^2+h^2\|\boldsymbol{U}_k\|_2^2\\&=\Big(1-\frac{2hmM}{m+M}\Big)\|\boldsymbol{\Delta}_k\|_2^2+h\Big(h-\frac{2}{m+M}\Big)\|\boldsymbol{U}_k\|_2^2.\end{aligned}\tag{16}$$

Note that, thanks to the strong convexity of f, the inequality $\|\boldsymbol{U}_k\|_2=\|\nabla f(\boldsymbol{\vartheta}^{(k,h)}+\boldsymbol{\Delta}_k)-\nabla f(\boldsymbol{\vartheta}^{(k,h)})\|_2\geq m\|\boldsymbol{\Delta}_k\|_2$ is true. If $h\leq 2/(m+M)$, this inequality can be combined with (16) to obtain

$$\|\boldsymbol{\Delta}_k-h\boldsymbol{U}_k\|_2^2\leq(1-hm)^2\|\boldsymbol{\Delta}_k\|_2^2.$$

Similarly, when $h\geq 2/(m+M)$, we can use the Lipschitz property of ∇f to infer that $\|\boldsymbol{U}_k\|_2\leq M\|\boldsymbol{\Delta}_k\|_2$. Combining with (16), this yields

$$\|\boldsymbol{\Delta}_k-h\boldsymbol{U}_k\|_2^2\leq(hM-1)^2\|\boldsymbol{\Delta}_k\|_2^2,\qquad\text{if}\qquad h\geq 2/(m+M).$$

Thus, we have checked that (14) is true for every $h\in(0,2/M)$. ■

Proof [Proof of Lemma 2] To simplify notations, we prove the lemma for $p=1$. The function $x\mapsto f'(x)$ being Lipschitz continuous is almost surely differentiable. Furthermore, it is clear that $|f''(\boldsymbol{x})|\leq M$ for every $\boldsymbol{x}$ for which this second derivative exists. The result of (Rudin, 1987, Theorem 7.20) implies that

$$f'(x)-f'(0)=\int_0^x f''(y)\,dy.$$

Therefore, using $f'(x)\,\pi(x)=-\pi'(x)$, we get

$$\begin{aligned}\int_{\mathbb{R}}f'(x)^2\,\pi(x)\,dx&=f'(0)\int_{\mathbb{R}}f'(x)\,\pi(x)\,dx+\int_{\mathbb{R}}\Big(\int_0^x f''(y)\,dy\Big)f'(x)\,\pi(x)\,dx\\&=-f'(0)\int_{\mathbb{R}}\pi'(x)\,dx-\int_{\mathbb{R}}\Big(\int_0^x f''(y)\,dy\Big)\pi'(x)\,dx\\&=-\int_0^\infty\int_0^x f''(y)\,\pi'(x)\,dy\,dx+\int_{-\infty}^0\int_x^0 f''(y)\,\pi'(x)\,dy\,dx.\end{aligned}$$

In view of Fubini's theorem, we arrive at

$$\int_{\mathbb{R}}f'(x)^2\,\pi(x)\,dx=\int_0^\infty f''(y)\,\pi(y)\,dy+\int_{-\infty}^0 f''(y)\,\pi(y)\,dy\leq M.$$

This completes the proof. ■

Proof [Proof of Lemma 3] Since the process $\boldsymbol{L}$ is stationary, $V(a)$ has the same distribution as $V(0)$. For this reason, it suffices to prove the claim of the lemma for $a = 0$ only. Using the Lipschitz continuity of f, we get

$$\begin{aligned}\mathbf{E}[\|\boldsymbol{V}(0)\|_2^2] &= \mathbf{E}\Big[\Big\| \int_0^h \big(\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big)\, dt\Big\|_2^2\Big] \\ &\le h \int_0^h \mathbf{E}\big[\big\|\nabla f(\boldsymbol{L}_t) - \nabla f(\boldsymbol{L}_0)\big\|_2^2\big]\, dt \\ &\le hM^2 \int_0^h \mathbf{E}\big[\big\|\boldsymbol{L}_t - \boldsymbol{L}_0\big\|_2^2\big]\, dt.\end{aligned}$$

Combining this inequality with the stationarity of $\boldsymbol{L}_t$, we arrive at

$$\begin{aligned}\Big(\mathbf{E}[\|\boldsymbol{V}(0)\|_2^2]\Big)^{1/2} &\le \left(hM^2 \int_0^h \mathbf{E}\big[\big\| - \int_0^t \nabla f(\boldsymbol{L}_s)\, ds + \sqrt{2}\, \boldsymbol{W}_t\big\|_2^2\big]\, dt\right)^{1/2} \\ &\le \left(hM^2 \int_0^h \mathbf{E}\big[\big\| \int_0^t \nabla f(\boldsymbol{L}_s)\, ds\big\|_2^2\big]\, dt\right)^{1/2} + \left(2hpM^2 \int_0^h t\, dt\right)^{1/2} \\ &\le \left(hM^2 \mathbf{E}\big[\big\|\nabla f(\boldsymbol{L}_0)\big\|_2^2\big] \int_0^h t^2\, dt\right)^{1/2} + \left(2hpM^2 \int_0^h t\, dt\right)^{1/2} \\ &= \left(\frac{1}{3} h^4 M^2 \mathbf{E}\big[\big\|\nabla f(\boldsymbol{L}_0)\big\|_2^2\big]\right)^{1/2} + \big(h^3 M^2 p\big)^{1/2}.\end{aligned}$$

To complete the proof, it suffices to apply Lemma 2. ■

Acknowledgments

The work of the author was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047). The author would like to thank Nicolas Brosse, who suggested an improvement in Theorem 3.

References

Y. Atchadé, G. Fort, E. Moulines, and P. Priouret. Adaptive Markov chain Monte Carlo: theory and methods. In *Bayesian time series models*, pages 32–51. Cambridge Univ. Press, Cambridge, 2011.

R. N. Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.*, 6(4):541–553, 08 1978.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with Projected Langevin Monte Carlo. *ArXiv e-prints*, July 2015.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *ArXiv e-prints*, December 2014.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *ArXiv e-prints*, May 2016.

Alain Durmus, Eric Moulines, and Marcelo Pereyra. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin. February 2016. URL https://hal.archives-ouvertes.fr/hal-01267115.

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

L. Lovász and S. Vempala. Fast algorithms for logconcave functions: Sampling, rounding, integration and optimization. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 57–68, 2006b.

Walter Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.