Theoretical guarantees for approximate sampling from smooth and log-concave densities
Author(s): Arnak S. Dalalyan
Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 79, No. 3 (JUNE 2017), pp. 651-676
Published by: Oxford University Press for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/44681805
Accessed: 03-06-2024 14:45 +00:00

J. R. Statist. Soc. B (2017)
79, *Part* 3, *pp.* 651–676

Theoretical guarantees for approximate sampling from smooth and log-concave densities

Arnak S. Dalalyan

ParisTech, Paris, France

[Received January 2015. Final revision February 2016]

Summary. Sampling from various kinds of distribution is an issue of paramount importance in statistics since it is often the key ingredient for constructing estimators, test procedures or confidence intervals. In many situations, exact sampling from a given distribution is impossible or computationally expensive and, therefore, one needs to resort to approximate sampling strategies. However, there is no well-developed theory providing meaningful non-asymptotic guarantees for the approximate sampling procedures, especially in high dimensional problems. The paper makes some progress in this direction by considering the problem of sampling from a distribution having a smooth and log-concave density defined on $\mathbb{R}^p$, for some integer $p > 0$. We establish non-asymptotic bounds for the error of approximating the target distribution by the distribution obtained by the Langevin Monte Carlo method and its variants. We illustrate the effectiveness of the established guarantees with various experiments. Underlying our analysis are insights from the theory of continuous time diffusion processes, which may be of interest beyond the framework of log-concave densities that are considered in the present work.

Keywords: Approximate sampling; Langevin algorithm; Markov chain Monte Carlo methods; Rates of convergence

1. Introduction

Let p be a positive integer and $f : \mathbb{R}^p \to \mathbb{R}$ be a measurable function such that the integral $\int_{\mathbb{R}^p} \exp\{-f(\boldsymbol{\theta})\}\,\mathrm{d}\boldsymbol{\theta}$ is finite. If we think of f as the negative log-likelihood or the negative log-posterior of a statistical model, then the maximum likelihood and the Bayesian estimators, which are perhaps the most popular in statistics, are respectively defined as

$$\boldsymbol{\theta}^{\mathrm{ML}} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} f(\boldsymbol{\theta}),$$

$$\boldsymbol{\theta}^{\mathrm{B}} = \frac{1}{\int_{\mathbb{R}^p} \exp\{-f(\mathbf{u})\}\,\mathrm{d}\mathbf{u}} \int_{\mathbb{R}^p} \boldsymbol{\theta} \exp\{-f(\boldsymbol{\theta})\}\,\mathrm{d}\boldsymbol{\theta}.$$

These estimators are rarely available in closed form. Therefore, optimization techniques are used for computing the maximum likelihood estimator whereas computation of the Bayes estimator often requires sampling from a density proportional to $\exp\{-f(\boldsymbol{\theta})\}$. In most situations, exact computation of these two estimators is impossible and we must resort to approximations provided by iterative algorithms. 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. However, a striking fact is that the convergence properties of optimization algorithms are much better understood than those of the approximate sampling algorithms.

Address for correspondence: Arnak S. Dalalyan, École Nationale de la Statistique et de l'Administration Économique, ParisTech, 3 Avenue Pierre Larousse, 92240 Malakoff, France.
E-mail: arnak.dalalyan@ensae.fr

 1369–7412/17/79651

The goal of the present work is partially to fill this gap by establishing easy-to-apply theoretical guarantees for some approximate sampling algorithms.

To be more precise, consider the case of a strongly convex function f having a Lipschitz continuous gradient, i.e. there are two positive constants m and M such that

$$\begin{aligned} f(\boldsymbol{\theta}) - f(\bar{\boldsymbol{\theta}}) - \nabla f(\bar{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) &\geqslant \frac{m}{2}\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2, \\ \|\nabla f(\boldsymbol{\theta}) - \nabla f(\bar{\boldsymbol{\theta}})\|_2 \leqslant M\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2, \qquad & \forall \boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^p, \end{aligned} \tag{1}$$

where ∇f stands for the gradient of f and $\|\cdot\|_2$ is the Euclidean norm. There is a simple result characterizing the convergence of the well-known gradient descent algorithm under assumption (1).

Theorem 1 (equation (9.18) in Boyd and Vandenberghe (2004)). If $f : \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable and fulfils assumption (1), then the gradient descent algorithm defined recursively by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - (2M)^{-1}\nabla f(\boldsymbol{\theta}^{(k)}), \qquad k = 0, 1, 2, \ldots, \tag{2}$$

satisfies

$$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{\mathrm{ML}}\|_2^2 \leqslant \frac{2\{f(\boldsymbol{\theta}^{(0)}) - f(\boldsymbol{\theta}^{\mathrm{ML}})\}}{m}\left(1 - \frac{m}{2M}\right)^k, \qquad \forall k \in \mathbb{N}. \tag{3}$$

Theorem 1 implies that the convergence of the gradient descent is exponential in k. More precisely, it results from equation (3) that to achieve an approximation error that is upper bounded by $\epsilon > 0$ in the Euclidean norm it suffices to perform

$$k_\epsilon = \frac{\log[2m^{-1}\{f(\boldsymbol{\theta}^{(0)}) - f(\boldsymbol{\theta}^{\mathrm{ML}})\}] + 2\log(1/\epsilon)}{\log\{2M/(2M-m)\}} \tag{4}$$

evaluations of the gradient of f. An important feature of this result is the logarithmic dependence of k_ϵ on ϵ but also its independence of the dimension p. Note also that, even though the right-hand side of equation (4) is a somewhat conservative bound on the number of iterations, all the quantities that are involved in that expression are easily computable and lead to a simple stopping rule for the iterative algorithm.

The situation for approximate computation of $\boldsymbol{\theta}^{\mathrm{B}}$ or for approximate sampling from the density proportional to $\exp\{-f(\boldsymbol{\theta})\}$ is much more imbalanced. Although there are almost as many algorithms for performing these tasks as for optimization, the convergence properties of most of them have been studied only empirically and, therefore, provide little theoretically grounded guidance for the choice of different tuning parameters or of the stopping rule. Furthermore, it is not clear how the rate of convergence of these algorithms scales with growing dimension. Although it is intuitively understandable that the problem of sampling from a distribution is more difficult than that of maximizing its density, this does not necessarily justify the huge gap between the precision of theoretical guarantees that are available for the solutions of these two problems. This gap is even more surprising in light of the numerous similarities between the optimization and approximate sampling algorithms.

Let us describe a particular example of approximate sampling algorithm, the Langevin Monte Carlo (LMC) algorithm, that will be studied throughout this work. Its definition is similar to the gradient descent algorithm for optimization but involves an additional step of random perturbation. Starting from an initial point $\boldsymbol{\vartheta}^{(0)} \in \mathbb{R}^p$ that may be deterministic or random, the subsequent steps of the algorithm are defined by the update rule

Table 1. Summary of the main findings of this work†

Algorithm	*Number of iterates, Gaussian start*	*Number of iterates, warm start*	*Complexity of one iteration*
LMC	$O^*(p^3\epsilon^{-2})$, theorem 2	$O^*(p\epsilon^{-2})$, Section 4.1	$O(p)$
LMCO	$O^*(p^{5/2}\epsilon^{-1})$, theorem 3	$O^*(p\epsilon^{-1})$, Section 5	$O(p^3)$

†The second and third columns provide the order of magnitude of the number of iterates to perform to make the error of approximation smaller than ϵ. The last column contains the worst-case complexity of one iteration. Note that in many practical situations the real complexity might be much smaller than the worst-case complexity.

$$\vartheta^{(k+1,h)} = \vartheta^{(k,h)} - h\nabla f(\vartheta^{(k,h)}) + \sqrt{(2h)}\,\xi^{(k+1)}, \qquad k = 0, 1, 2, \ldots, \tag{5}$$

where $h > 0$ is a tuning parameter, which is often referred to as the step size, and $\xi^{(1)}, \ldots, \xi^{(k)}, \ldots$ is a sequence of independent centred Gaussian vectors with covariance matrix equal to the identity matrix and independent of $\vartheta^{(0)}$. It is well known that, under some assumptions on f, when h is small and k is large (so that the product kh is large), the distribution of $\vartheta^{(k,h)}$ is close in total variation to the distribution with density proportional to $\exp\{-f(\theta)\}$, which is hereafter referred to as the target distribution. The goal of the present work is to establish a non-asymptotic upper bound, involving only explicit and computable quantities, on the total variation distance between the target distribution and its approximation by the distribution of $\vartheta^{(k,h)}$. We shall also analyse a variant of the LMC algorithm, termed the LMCO algorithm, which makes use of the Hessian of f.

To give the reader a foretaste of the main contributions of the present work, we summarize in Table 1 some guarantees established and described in detail in the next sections. To keep things simple, we translated all the non-asymptotic results into asymptotic results for large dimension p and small precision level ϵ (the O^*-notation ignores the dependence on constant and logarithmic factors). The complexity of one iteration of the LMC algorithm indicated in Table 1 corresponds to the computation of the gradient and generation of a Gaussian p-vector, whereas the complexity of one iteration of the LMCO algorithm is the cost of performing a singular values decomposition on the Hessian matrix of f, which is of size $p \times p$.

1.1. Notation

For any $p \in \mathbb{N}$ we write $\mathscr{B}(\mathbb{R}^p)$ for the σ-algebra of Borel sets of $\mathbb{R}^p$. The Euclidean norm of $\mathbb{R}^p$ is denoted by $\|\cdot\|_2$ whereas $\|\nu\|_{\mathrm{TV}}$ stands for the total variation norm of a signed measure ν: $\|\nu\|_{\mathrm{TV}} = \sup_{A \in \mathscr{B}(\mathbb{R}^p)} |\nu(A)|$. For two probability measures ν and $\bar{\nu}$ defined on a space $\mathcal{X}$ and such that ν is absolutely continuous with respect to $\bar{\nu}$, the Kullback–Leibler and χ^2-divergences between ν and $\bar{\nu}$ are respectively defined by

$$\mathrm{KL}(\nu\|\bar{\nu}) = \int_{\mathcal{X}} \log\left\{\frac{\mathrm{d}\nu}{\mathrm{d}\bar{\nu}}(\mathbf{x})\right\}\nu(\mathrm{d}\mathbf{x}),$$

$$\chi^2(\nu\|\bar{\nu}) = \int_{\mathcal{X}} \left\{\frac{\mathrm{d}\nu}{\mathrm{d}\bar{\nu}}(\mathbf{x}) - 1\right\}^2 \bar{\nu}(\mathrm{d}\mathbf{x}).$$

All the probability densities on $\mathbb{R}^p$ are with respect to the Lebesgue measure, unless otherwise specified. We denote by π the probability density function proportional to $\exp\{-f(\theta)\}$, by $\mathbf{P}_\pi$

the corresponding probability distribution and by $\mathbf{E}_\pi$ the expectation with respect to $\mathbf{P}_\pi$. For a probability density ν and a Markov kernel $\mathbf{Q}$, we denote by $\nu\mathbf{Q}$ the probability distribution $\{(\nu\mathbf{Q})(A) = \int_{\mathbb{R}^p} \nu(\mathbf{x})\mathbf{Q}(\mathbf{x}, A)\,\mathrm{d}\mathbf{x} : A \in \mathscr{B}(\mathbb{R}^p)\}$. We say that the density $\pi(\boldsymbol{\theta}) \propto \exp\{-f(\boldsymbol{\theta})\}$ is log-concave or strongly log-concave if the function f satisfies the first inequality of assumption (1) with $m = 0$ or $m > 0$ respectively. We refer the interested reader to Saumard and Wellner (2014) for a comprehensive survey on log-concave densities.

The programs that were used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Background on the Langevin Monte Carlo algorithm

The rationale behind the LMC algorithm (5) is simple: the Markov chain $\{\boldsymbol{\vartheta}^{(k,h)}\}_{k\in\mathbb{N}}$ is the Euler discretization of a continuous time diffusion process $\{\mathbf{L}_t : t \in \mathbb{R}_+\}$, known as Langevin diffusion, that has π as invariant density. The Langevin diffusion is defined by the stochastic differential equation

$$\mathrm{d}\mathbf{L}_t = -\nabla f(\mathbf{L}_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{W}_t, \qquad t \geqslant 0, \tag{6}$$

where $\{\mathbf{W}_t : t \geqslant 0\}$ is p-dimensional Brownian motion. When f satisfies condition (1), equation (6) has a unique strong solution which is a Markov process. In what follows, the transition kernel of this process is denoted by $\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, \cdot)$, i.e. $\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) = \mathbf{P}(\mathbf{L}_t \in A | \mathbf{L}_0 = \mathbf{x})$ for all Borel sets $A \subset \mathbb{R}^p$ and any initial condition $\mathbf{x} \in \mathbb{R}^p$. Furthermore, assumption (1) yields the spectral gap property of the semigroup $\{\mathbf{P}_{\mathbf{L}}^t : t \in \mathbb{R}_+\}$, which in turn implies that the process $\mathbf{L}_t$ is geometrically ergodic in the following sense.

Lemma 1. Under assumption (1), for any probability density ν,

$$\|\nu\mathbf{P}_{\mathbf{L}}^t - \pi\|_{\mathrm{TV}} \leqslant \tfrac{1}{2}\chi^2(\nu\|\pi)^{1/2}\exp(-tm/2), \qquad \forall\, t \geqslant 0. \tag{7}$$

The proof of lemma 1, which is postponed to Appendix A, is based on the bounds on the spectral gap established in Chen and Wang (1997), remark 4.14; see also Bakry *et al.* (2014), corollary 4.8.2. In simple words, inequality (7) shows that, for large values of t, the distribution of $\mathbf{L}_t$ approaches exponentially fast the target distribution, and the idea behind the LMC algorithm is to approximate $\mathbf{L}_t$ by $\boldsymbol{\vartheta}^{(k,h)}$ for $t = kh$. Inequalities of type (7) can be obtained under conditions (such as the curvature dimension condition; see Bakry *et al.* (2014), definition 1.16.1 and theorem 4.8.4) that are weaker than the strong log-concavity that is required in the present work. However, we decided to restrict ourselves to the strong log-concavity condition since it is easy to check and is commonly used in machine learning and optimization.

The first and probably the most influential work providing probabilistic analysis of asymptotic properties of the LMC algorithm is Roberts and Tweedie (1996). However, one of the recommendations made by them is to avoid using the Langevin algorithm as it is defined in expression (5), or to use it very cautiously, since the ergodicity of the corresponding Markov chain is very sensitive to the choice of the parameter h. Even in cases where the Langevin diffusion is geometrically ergodic, an inappropriate choice of h may result in the transience of the Markov chain $\{\boldsymbol{\vartheta}^{(k,h)}\}$. These findings have very strongly influenced subsequent studies since all the ensuing research focused essentially on the Metropolis-adjusted version of the LMC algorithm, known as the Metropolis-adjusted Langevin algorithm, and its modifications (Roberts and Rosenthal, 1998; Stramer and Tweedie, 1999a, b; Jarner and Hansen, 2000; Roberts and Stramer, 2002; Pillai *et al.*, 2012; Xifara *et al.*, 2014).

In contrast with this, we show here that under the strong convexity assumption that is imposed on f (or, equivalently, on $-\log(\pi)$) coupled with Lipschitz continuity of the gradient of f, we can ensure the non-transience of the Markov chain $\boldsymbol{\vartheta}^{(k,h)}$ by simply choosing $h \leqslant 1/M$. In fact, the non-explosion of this chain follows from the following proposition, the proof of which is very strongly inspired by that of theorem 1.

Proposition 1. Let the function f be continuously differentiable on $\mathbb{R}^p$ and satisfy assumption (1) with $f^* = \inf_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$. Then, for every $h \leqslant 1/M$, we have

$$\mathbf{E}[f(\boldsymbol{\vartheta}^{(k,h)}) - f^*] \leqslant (1-mh)^k \, \mathbf{E}[f(\boldsymbol{\vartheta}^{(0)}) - f^*] + \frac{Mp}{m}. \tag{8}$$

Under the condition $h \leqslant 1/M$, the quantity $1-mh$ is always non-negative. Indeed, it follows (see lemma 4 in Appendix A) from Taylor series expansion and Lipschitz continuity of the gradient ∇f that $f(\boldsymbol{\theta}) - f(\bar{\boldsymbol{\theta}}) - \nabla f(\bar{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \leqslant (M/2)\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2$ for every $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^p$, which—in view of assumption (1)—entails that $m \leqslant M$ and, therefore, $1 - mh \geqslant 1 - Mh \geqslant 0$. However, in view of the strong convexity of f, inequality (8) implies that

$$\mathbf{E}[\|\boldsymbol{\vartheta}^{(k,h)} - \boldsymbol{\theta}^*\|_2^2] \leqslant \frac{M}{m}\mathbf{E}[\|\boldsymbol{\vartheta}^{(0)} - \boldsymbol{\theta}^*\|_2^2] + \frac{2Mp}{m^2}, \tag{9}$$

where $\boldsymbol{\theta}^*$ stands for the point of (global) minimum of f. As a consequence, the sequence $\boldsymbol{\vartheta}^{(k,h)}$ that is produced by the LMC algorithm is bounded in L^2 provided that $h \leqslant 1/M$.

A crucial step in analysing the long-time behaviour of the LMC algorithm is the assessment of the distance between the distribution of the random variable $\mathbf{L}_{Kh}$ and that of $\boldsymbol{\vartheta}^{(K,h)}$. It is intuitively clear that for a fixed K this distance should tend to 0 when $h \to 0$. However, to obtain informative bounds we need to quantify the rate of this convergence. For this, we follow the ideas that were presented in Dalalyan and Tsybakov (2009, 2012) which consist in performing the following two steps. First, a continuous time Markov process $\{\mathbf{D}_t : t \geqslant 0\}$ is introduced such that the distribution of the random vectors $(\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(1,h)}, \ldots, \boldsymbol{\vartheta}^{(K,h)})$ and $(\mathbf{D}_0, \mathbf{D}_h, \ldots, \mathbf{D}_{Kh})$ coincide. Second, the distance between the distributions of the variables $\mathbf{D}_{Kh}$ and $\mathbf{L}_{Kh}$ is bounded from above by the distance between the distributions of the continuous time processes $\{\mathbf{D}_t : t \in [0, Kh]\}$ and $\{\mathbf{L}_t : t \in [0, Kh]\}$.

To be more precise, we introduce a diffusion-type continuous time process $\mathbf{D}$ obeying the stochastic differential equation

$$\mathrm{d}\mathbf{D}_t = \mathbf{b}_t(\mathbf{D})\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}\mathbf{W}_t, \qquad t \geqslant 0, \quad \mathbf{D}_0 = \boldsymbol{\vartheta}^{(0)}, \tag{10}$$

with the (non-anticipative) drift

$$\mathbf{b}_t(\mathbf{D}) = -\sum_{k=0}^{\infty} \nabla f(\mathbf{D}_{kh})\, \mathbb{1}_{[kh,(k+1)h]}(t).$$

By integrating equation (10) on the interval $[kh, (k+1)h]$, we check that the increments of this process satisfy $\mathbf{D}_{(k+1)h} - \mathbf{D}_{kh} = -h\nabla f(\mathbf{D}_{kh}) + \sqrt{(2h)}\boldsymbol{\zeta}^{(k+1)}$, where $\boldsymbol{\zeta}^{(k+1)} = (\mathbf{W}_{(k+1)h} - \mathbf{W}_{kh})/\sqrt{h}$. Since Brownian motion is a Gaussian process with independent increments, we conclude that $\{\boldsymbol{\zeta}^{(k)} : k = 1, \ldots, K\}$ is a sequence of independently and identically distributed standard Gaussian random vectors. This readily implies equality of the distributions of the random vectors $(\boldsymbol{\vartheta}^{(0)}, \boldsymbol{\vartheta}^{(1,h)}, \ldots, \boldsymbol{\vartheta}^{(K,h)})$ and $(\mathbf{D}_0, \mathbf{D}_h, \ldots, \mathbf{D}_{Kh})$.

The specific form of the drift $\mathbf{b}$ that is used in the LMC algorithm has the advantage of meeting the following two conditions. First, $\mathbf{b}_t(\mathbf{L})$ is close to $-\nabla f(\mathbf{L}_t)$, the drift of the Langevin diffusion. Second, it is possible to sample from the distribution $\mathbf{P}_{\mathrm{D}}^h(\mathbf{x}, \cdot)$, where h is the step of discretization that is used in the LMC algorithm. Any non-anticipative drift function satisfying

these two conditions may be used for defining a version of the LMC algorithm. Such an example, the LMC algorithm with Ozaki discretization, is considered in Section 5.

To close this section, we state an inequality that will be repeatedly used in this work and the proof of which—based on the Girsanov formula—can be found, for instance in Dalalyan and Tsybakov (2012). If for some $B>0$ the non-anticipative drift function $\mathbf{b}: C(\mathbb{R}_+, \mathbb{R}^p) \times \mathbb{R}_+ \to \mathbb{R}^p$ satisfies the inequality $\|\mathbf{b}(\mathbf{D}, t)\|_2 \leqslant B(1+\|\mathbf{D}\|_\infty)$ for every $t \in [0, Kh]$ and every $\mathbf{D} \in C(\mathbb{R}_+, \mathbb{R}^p)$, then the Kullback–Leibler divergence between $\mathbb{P}_{\mathbf{L}}^{\mathbf{x}, Kh}$ and $\mathbb{P}_{\mathbf{D}}^{\mathbf{x}, Kh}$, the distributions of the processes $\{\mathbf{L}: t \in [0, Kh]\}$ and $\{\mathbf{D}: t \in [0, Kh]\}$ with the initial value $\mathbf{L}_0 = \mathbf{D}_0 = \mathbf{x}$, is given by

$$\mathrm{KL}(\mathbb{P}_{\mathbf{L}}^{\mathbf{x}, Kh} \| \mathbb{P}_{\mathbf{D}}^{\mathbf{x}, Kh}) = \tfrac{1}{4} \int_0^{Kh} \mathbf{E}[\|\nabla f(\mathbf{D}_t) + \mathbf{b}_t(\mathbf{D})\|_2^2] \mathrm{d}t. \tag{11}$$

It is worth emphasizing that the last equality remains valid when the initial values of the processes $\mathbf{D}$ and $\mathbf{L}$ are random but have the same distribution.

The idea of discretizing the diffusion process to sample approximately from its invariant density is not new. It can be traced back at least to Lamberton and Pagès (2002); see also Lemaire (2005) for an overview. The results therein are stated for more general discretization with variable step sizes but are of asymptotic nature. This point of view has been adopted and extended to the non-asymptotic case in the recent work of Durmus and Moulines (2015).

3. Non-asymptotic bounds on the error of the Langevin Monte Carlo algorithm

We can now establish a non-asymptotic bound with explicit constants on the distance between the target distribution $\mathbf{P}_\pi$ and the distribution that is produced by the LMC algorithm. As explained earlier, the bound is obtained by controlling two types of error: the error of approximating $\mathbf{P}_\pi$ by the distribution of the Langevin diffusion $\mathbf{L}_{Kh}$ (6) and the error of approximating the Langevin diffusion by its discretized version $\mathbf{D}$ given by expression (10). The first error is a decreasing function of $T = Kh$: to make this error small it is necessary to choose a large T. Quite a precise quantitative assessment of this error is given by lemma 1 in the previous section. The second error vanishes when the step size h goes to 0, provided that $T = Kh$ is fixed. Thus, it is in our interest to choose a small h. However, our goal is not only to minimize the error, but also to reduce, as much as possible, the computational cost of the algorithm. For a fixed T, if we choose a small value of h then a large number of steps K is necessary for getting close to the target distribution. Therefore, the computational complexity is a decreasing function of h. To find a value of h leading to a reasonable trade-off between the computational complexity and the approximation error, we need to complement lemma 1 with a precise bound on the second approximation error. This is done in the following lemma.

Lemma 2. Let $f: \mathbb{R}^p \to \mathbb{R}$ be a function satisfying the second inequality in assumption (1) and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be a stationary point (i.e. $\nabla f(\boldsymbol{\theta}^*) = 0$). For any $T > 0$, let $\mathbb{P}_L^{\mathbf{x}, T}$ and $\mathbb{P}_D^{\mathbf{x}, T}$ be respectively the distributions of the Langevin diffusion (6) and its approximation (10) on the space of all continuous paths on $[0, T]$ with values in $\mathbb{R}^p$, with a fixed initial value $\mathbf{x}$. Then, if $h \leqslant 1/(\alpha M)$ with $\alpha \geqslant 1$, it holds that

$$\mathrm{KL}(\mathbb{P}_{\mathbf{L}}^{\mathbf{x}, Kh} \| \mathbb{P}_{\mathbf{D}}^{\mathbf{x}, Kh}) \leqslant \frac{M^3 h^2 \alpha}{12(2\alpha - 1)} (\|\mathbf{x} - \boldsymbol{\theta}^*\|_2^2 + 2Khp) + \frac{pKM^2h^2}{4}. \tag{12}$$

Set $T = Kh$. Since it simplifies the mathematical formulae and is possible to achieve in practice in view of theorem 1, we assume in what follows that the initial value of the LMC algorithm is

drawn at random from the Gaussian distribution with mean $\boldsymbol{\theta}^*$, a stationary point of f, and covariance matrix $M^{-1}\mathbf{I}_p$. Then, in view of inequality (12) and the convexity of the Kullback–Leibler divergence, we obtain (for $\nu = \mathcal{N}_p(\boldsymbol{\theta}^*, M^{-1}\mathbf{I}_p)$)

$$\begin{aligned}\mathrm{KL}(\nu\mathbb{P}_L^T \| \nu\mathbb{P}_D^T) &\leqslant \frac{pM^2h^2\alpha}{12(2\alpha-1)} + \frac{pM^3Th^2\alpha}{6(2\alpha-1)} + \frac{pM^2Th}{4} \\ &= \frac{pM^2Th}{4}\left\{\frac{\alpha}{3K(2\alpha-1)} + \frac{2Mh\alpha}{3(2\alpha-1)} + 1\right\} \leqslant \frac{pM^2Th\alpha}{2(2\alpha-1)}, \end{aligned} \tag{13}$$

for every $K \geqslant \alpha$ and $h \leqslant 1/(\alpha M)$. We can now state the main result of this section, the proof of which is postponed to Appendix A.

Theorem 2. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function satisfying assumption (1) and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be its global minimum point. Assume that, for some $\alpha \geqslant 1$, we have $h \leqslant 1/(\alpha M)$ and $K \geqslant \alpha$. Then, for any time horizon $T = Kh$, the total variation distance between the target distribution $\mathbf{P}_\pi$ and the approximation $\nu\mathbf{P}_\vartheta^K$ furnished by the LMC algorithm with the initial distribution $\nu = \mathcal{N}_p(\boldsymbol{\theta}^*, M^{-1}\mathbf{I}_p)$ satisfies

$$\|\nu\mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{M}{m}\right) - \frac{Tm}{2}\right\} + \left\{\frac{pM^2Th\alpha}{4(2\alpha-1)}\right\}^{1/2}. \tag{14}$$

Remark 1. The second term on the right-hand side of inequality (14) tends to ∞ when the time horizon $T \to \infty$ while the step size h remains fixed. Since the total variation is always bounded by 1, the bound obtained is not sharp for large values of T. The main reason for this is the fact that the total variation distance is upper bounded by the Kullback–Leibler divergence. Improving this argument to obtain a tighter upper bound is a challenging open problem.

We provide here a simple consequence of theorem 2 that furnishes easy-to-apply rules for choosing the time horizon T and the step size h.

Corollary 1. Let $p \geqslant 2$, f satisfy assumption (1) and $\epsilon \in (0, \frac{1}{2})$ be a target precision level. Let the time horizon T and the step size h be defined by

$$\begin{aligned} T &= \frac{4\log(1/\epsilon) + p\log(M/m)}{2m}, \\ h &= \frac{\epsilon^2(2\alpha-1)}{M^2Tp\alpha}, \end{aligned} \tag{15}$$

where $\alpha = (1 + MpT\epsilon^{-2})/2$. Then the output of the K-step LMC algorithm, with $K = \lceil T/h \rceil$, satisfies $\|\nu\mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \epsilon$.

Proof. The choice of T and h implies that the two summands on the right-hand side of inequality (14) are bounded by $\epsilon/2$. Furthermore, one can easily check that $\alpha = (1 + MpT\epsilon^{-2})/2$ is larger than 1 and satisfies $h \leqslant 1/(\alpha M)$. In addition, $K \geqslant T/h \geqslant \alpha MT \geqslant 2\alpha(M/m)\log(1/\epsilon) \geqslant \alpha\log(4)$, which ensures the applicability of theorem 2. □

Let us first remark that the claim of corollary 1 can be simplified by taking $\alpha = 1$. However, for this value of α the factor $(2\alpha - 1)/\alpha$ equals 1, whereas, for the slightly more complicated choice that is recommended by corollary 1, this factor is close to 2. In practice, increasing h by a factor 2 results in halving the running time, which represents a non-negligible gain.

Besides providing concrete and easily applicable guidance for choosing the step of discretization and the stopping rule for the LMC algorithm to achieve a prescribed error rate, corollary 1

tells us that, to obtain an error that is smaller than ϵ, it is enough to perform $K = O(T^2 p/\epsilon^2) = O[\epsilon^{-2}\{p^3 + p\log^2(1/\epsilon)\}]$ evaluations of the gradient of f. To the best of our knowledge, this is the first result that establishes polynomial in p guarantees for sampling from a log-concave density by using the LMC algorithm. We discuss the relationship of this and subsequent results to earlier work in Section 7.

4. Possible extensions

In this section, we state some extensions of the previous results that do not require any major change in the proofs but might lead to improved computational complexity or be valid under relaxed assumptions in some particular cases.

4.1. Improved bounds for a 'warm start'

The choice of the distribution ν of the initial value $\theta^{(0)}$ has a significant influence on the convergence of the LMC algorithm. If ν is close to π, a smaller number of iterations might be enough for making the total variation error smaller than ϵ. The goal of this section is to present quantitative bounds characterizing the influence of ν on the convergence and, as a consequence, on the computational complexity of the LMC algorithm.

The first observation that can be readily deduced from inequality (12) is that, for any $h \leqslant 1/(2M)$,

$$\mathrm{KL}(\nu\mathbb{P}_L^T \| \nu\mathbb{P}_D^T) \leqslant \frac{M^3 h^2\, \mathbf{E}_{\vartheta\sim\nu}[\|\vartheta - \theta^*\|_2^2]}{18} + \frac{pM^2Th}{3}. \tag{16}$$

Combining this bound with inequality (38) in Appendix A.2, lemma 1 and inequality (40) in Appendix A.2 we obtain

$$\|\nu\mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\exp\left[\frac{\log\{\chi^2(\nu\|\pi)\} - Tm}{2}\right] + \left(\frac{M^3h^2\,\mathbf{E}_\nu[\|\vartheta - \theta^*\|_2^2] + 6pM^2Th}{18}\right)^{1/2}.$$

Elaborating on this inequality, we obtain the following result.

Proposition 2. Let ν be a probability density on $\mathbb{R}^p$ such that the second-order moment $\mu_2 = (M/p)\mathbf{E}_{\vartheta\sim\nu}[\|\vartheta - \theta^*\|_2^2]$ and the divergence $\chi^2(\nu\|\pi)$ are finite. Then, the LMC algorithm having ν as initial distribution and using the time horizon T and step size h defined by

$$T = \frac{2\log(1/\epsilon) + \log\{\chi^2(\nu\|\pi)\}}{m}, \qquad h = \frac{9\epsilon^2}{TM^2p(6+\mu_2)} \tag{17}$$

satisfies, for $K = [T/h] \geqslant 2$, the inequality $\|\nu\mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \epsilon$.

The proof of proposition 2 is immediate and, therefore, is left to the reader. What we infer from this result is that the choice of the initial distribution ν has a strong influence on the convergence of the LMC algorithm. For instance, if for some specific π we can sample from a density ν satisfying, for some $\rho > 0$, the relationship $\chi^2(\nu\|\pi) = O(p^\rho)$ as $p \to \infty$, then the time horizon T for approximating the target density π within ϵ is $O\{\log(p \vee \epsilon^{-1})\}$ and the step size satisfies $h^{-1} = O\{\epsilon^{-2}p\log(p \vee \epsilon^{-1})\}$. Thus, in such a situation, we need to perform $[T/h] = O\{\epsilon^{-2}p\log^2(p \vee \epsilon^{-1})\}$ evaluations of the gradient of f to obtain a sampling density

within a distance of ϵ of the target, which is substantially smaller than $O[\epsilon^{-2}\{p^3 + p\log^2(1/\epsilon)\}]$ obtained in the previous section in the general case.

4.2. Preconditioning

As is frequently done in optimization, we may introduce a preconditioner in the LMC algorithm to accelerate its convergence. To some extent, it amounts to choosing a positive definite $p \times p$ matrix $\mathbf{A}$, called the preconditioner, and applying the LMC algorithm to the function $g(\mathbf{y}) = f(\mathbf{A}\mathbf{y})$. Let $\{\boldsymbol{\eta}^{(k,h)} : k \in \mathbb{N}\}$ be the sequence that is obtained by the LMC algorithm applied to the function g, i.e. the density of $\boldsymbol{\eta}^{(k,h)}$ is close to $\pi_g(\mathbf{y}) \propto \exp\{-g(\mathbf{y})\}$ when k is large and h is small. Then, the sequence $\boldsymbol{\vartheta}^{(k,h)} = \mathbf{A}\boldsymbol{\eta}^{(k,h)}$ is approximately sampled from the density $\pi_f(\mathbf{x}) \propto \exp\{-f(\mathbf{x})\}$. This follows from the fact that if $\boldsymbol{\eta} \sim \pi_g$ then $\mathbf{A}\boldsymbol{\eta} \sim \pi_f$. Furthermore, it holds that

$$\|\mathbf{P}_{\boldsymbol{\vartheta}}^k - \mathbf{P}_{\pi_f}\|_{\mathrm{TV}} = \|\mathbf{P}_{\boldsymbol{\eta}}^k - \mathbf{P}_{\pi_g}\|_{\mathrm{TV}},$$

i.e. the approximation error of the LMC algorithm with a preconditioner $\mathbf{A}$ is characterized by corollary 1. This means that, if the function g satisfies condition (1) with constants $(m_{\mathbf{A}}, M_{\mathbf{A}})$, then the number of steps K after which the preconditioned LMC algorithm has an error bounded by ϵ is given by

$$K = (M_{\mathbf{A}}/m_{\mathbf{A}})^2 p\epsilon^{-2}\{2\log(1/\epsilon) + (p/2)\log(M_{\mathbf{A}}/m_{\mathbf{A}})\}^2.$$

Hence, the preconditioner $\mathbf{A}$ yielding the best guaranteed computational complexity for the LMC algorithm is the matrix $\mathbf{A}$ minimizing the ratio $M_{\mathbf{A}}/m_{\mathbf{A}}$.

The effect of preconditioning can be measured, for instance, in the case of multi-dimensional logistic regression considered in Section 6 below. In this case, the ratio $M_{\mathbf{A}}/m_{\mathbf{A}}$ is up to some constant factor equal to the condition number of the matrix $\mathbf{A}\Sigma_{\mathbf{X}}\mathbf{A}$, where $\Sigma_{\mathbf{X}}$ is the Gram matrix of the covariates.

4.3. Non-strongly log-concave densities

Theoretical guarantees that were developed in previous sections assume that the logarithm of the target density is strongly concave; see assumption (1). However, they can also be used for approximate sampling from a density which is log-concave but not necessarily strongly log-concave; we call these densities non-strongly log-concave. The idea is then to approximate the target density by a strongly log-concave density and to apply the LMC algorithm to the latter instead of to the former.

More precisely, assume that we wish to sample approximately from a multivariate target density $\pi(\mathbf{x}) \propto \exp\{-f(\mathbf{x})\}$, where the function $f : \mathbb{R}^p \to \mathbb{R}$ is twice differentiable with Lipschitz continuous gradient (i.e. f satisfies the second inequality in assumption (1)). Assume, in addition, that for every $R \in [0, \infty]$ there exists $m_R \geqslant 0$ such that $\nabla^2 f(\mathbf{x}) \succeq m_R \mathbf{I}_p$ for every $\mathbf{x} \in B = B_R(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x} - \mathbf{x}_0\|_2 \leqslant R\}$. Here, $\mathbf{X}_0$ is an arbitrarily fixed point in $\mathbb{R}^p$. If $m_\infty > 0$, then this assumption implies the first inequality in assumption (1) with $m = m_\infty$. The purpose of this subsection is to deal with the case where $m_\infty = 0$ or is very small. Let $\gamma > 0$ be a tuning parameter; we introduce the approximate log-density

$$\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\gamma}{2}(\|\mathbf{x} - \mathbf{x}_0\|_2 - R)^2 \mathbb{1}_{B^c}(\mathbf{x}). \tag{18}$$

This function satisfies both inequalities in assumption (1) with $\bar{m} = m_{2R} \wedge (m_\infty + 0.5\gamma)$ and $\bar{M} = M + \gamma$. Denote by $\bar{\pi}$ the density that is defined by $\bar{\pi}(\mathbf{x}) \propto \exp\{-\bar{f}(\mathbf{x})\}$ and by $\mathbf{P}_{\bar{\pi}}$ the corresponding probability distribution on $\mathbb{R}^p$. Heuristically, it is natural to expect that under some mild assumptions the distribution $\mathbf{P}_{\bar{\pi}}$ is close to the target $\mathbf{P}_\pi$ when R is large and γ is small.

This claim is made rigorous thanks to the following result, which is stated in broad generality to be applicable to approximations $\bar{f}$ that are not necessarily of the form (18).

Lemma 3. Let f and $\bar{f}$ be two functions such that $f(\mathbf{x}) \leqslant \bar{f}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$ and both $\exp(-f)$ and $\exp(-\bar{f})$ are integrable. Then the Kullback–Leibler divergence between the distribution $\mathbf{P}_{\bar{\pi}}$ defined by the density $\bar{\pi}(\mathbf{x}) \propto \exp\{-\bar{f}(\mathbf{x})\}$ and the target distribution $\mathbf{P}_\pi$ can be bounded as

$$\mathrm{KL}(\mathbf{P}_\pi \| \mathbf{P}_{\bar{\pi}}) \leqslant \tfrac{1}{2} \int_{\mathbb{R}^p} \{\bar{f}(\mathbf{x}) - f(\mathbf{x})\}^2 \pi(\mathbf{x})\, \mathrm{d}\mathbf{x}. \tag{19}$$

As a consequence, $\|\mathbf{P}_{\bar{\pi}} - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2} \|\bar{f} - f\|_{L^2(\pi)}$.

Proof. Using the formula for the Kullback–Leibler divergence, we obtain

$$\mathrm{KL}(\mathbf{P}_\pi \| \mathbf{P}_{\bar{\pi}}) = \int_{\mathbb{R}^p} \{\bar{f}(\mathbf{x}) - f(\mathbf{x})\} \pi(\mathbf{x})\, \mathrm{d}x + \log\left[\int_{\mathbb{R}^p} \exp\{f(\mathbf{x}) - \bar{f}(\mathbf{x})\} \pi(\mathbf{x})\, \mathrm{d}\mathbf{x}\right]. \tag{20}$$

Applying successively the inequalities $\log(u) \leqslant u - 1$ and $\exp(-u) \leqslant 1 - u + \frac{1}{2}u^2$ for every $u \geqslant 0$, the second term on the right-hand side of equation (20) is upper bounded as follows:

$$\log\left[\int_{\mathbb{R}^p} \exp\{f(\mathbf{x}) - \bar{f}(\mathbf{x})\} \pi(\mathbf{x})\, \mathrm{d}\mathbf{x}\right] \leqslant \int_{\mathbb{R}^p} \exp(f - \bar{f})\pi - 1 \leqslant -\int_{\mathbb{R}^p} (\bar{f} - f)\pi + \tfrac{1}{2} \int_{\mathbb{R}^p} (\bar{f} - f)^2 \pi.$$

Combining this inequality with equation (20), we obtain the first claim. The last claim of lemma 3 follows from the Pinsker inequality. □

For $\bar{f}$ given by equation (18), we obtain

$$\|\mathbf{P}_{\bar{\pi}} - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{\gamma}{4} \left\{ \int_{B^{\mathrm{c}}} (\|\mathbf{x} - \mathbf{x}_0\|_2 - R)^4 \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \right\}^{1/2}.$$

Choosing the parameter γ sufficiently small and the parameter R sufficiently large to ensure that $\|\mathbf{P}_{\bar{\pi}} - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \epsilon/2$ and assuming that π has bounded fourth-order moment, we derive from this inequality and corollary 1 the following convergence result for the approximate LMC algorithm.

Corollary 2. Let f be a twice differentiable function satisfying $m_R \mathbf{I}_p \preceq \nabla^2 f(\mathbf{x}) \preceq M \mathbf{I}_p$ for every $\mathbf{x} \in B_R(\mathbf{x}_0)$ and for every $R \in [0, \infty]$. Let $\epsilon \in (0, \frac{1}{2})$ be a target precision level. Assume that for some known value μ_R we have $\int_{B_R(\mathbf{x}_0)^{\mathrm{c}}} (\|\mathbf{x} - \mathbf{x}_0\|_2 - R)^4 \pi(\mathbf{x})\, \mathrm{d}\mathbf{x} \leqslant p^2 \mu_R^2$ and define $\bar{m} = m_{2R} \wedge (m_\infty + 0.5\gamma)$ and $\bar{M} = M + \gamma$ for some $\gamma \leqslant 2\epsilon/(p\mu_R)$. Set the time horizon T and the step size h as follows:

$$T = \frac{4\log(2/\epsilon) + p\log(\bar{M}/\bar{m})}{2\bar{m}}, \qquad h = \frac{\epsilon^2}{4\bar{M}^2 T p}. \tag{21}$$

Then the output of the K-step LMC algorithm (5) applied to the approximation $\bar{f}$ provided by equation (18), with $K = \lceil T/h \rceil$, satisfies $\|\nu \mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \epsilon$.

Let us comment on this result in the case $R=0$ which concerns non-strongly log-concave densities. Then the previous result implies that $K=O\{p^5\epsilon^{-4}\log^2(p\vee\epsilon^{-1})\}$. Clearly, the dependence of K both on the dimension p and on the acceptable error level ϵ grows substantially worse compared with the strongly log-concave case. Some improvements are possible in specific cases. First, we can improve the dependence of K on p if we can simulate from a distribution ν that is not too far from $\bar{\pi}$ in the sense of χ^2-divergence. More precisely, repeating the arguments of Section 4.1 we obtain the following result: if the initial distribution of the LMC algorithm satisfies $\chi^2(\nu\|\bar{\pi})=O\{(p/\gamma)^\rho\}$ for some $\rho>0$ then we need at most $K=O\{p^3\epsilon^{-4}\log^2(p/\epsilon)\}$ steps of the LMC algorithm for obtaining an error bounded by ϵ. Second, in some cases the dependence of K on p can be further improved by using a preconditioner and/or by replacing the penalty $\|\mathbf{x}\|_2^2$ in equation (18) by $\|\mathbf{M}\mathbf{x}\|_2^2$, where $\mathbf{M}$ is a properly chosen $p\times p$ matrix.

This being said, our intuition is that corollary 2 is more helpful in the case of convex functions f that are strongly convex in a neighbourhood of their minimum point $\boldsymbol{\theta}^*$. In such a situation, our recommendation is to set $\mathbf{x}_0=\boldsymbol{\theta}^*$ and to choose R by maximizing the quantity $\bar{m}=m_{2R}\wedge\{m_\infty+\epsilon/(p\mu_R)\}$. We showcase this approach in Section 6 in the example of logistic regression.

The convergence of Markov chain Monte Carlo (MCMC) methods for sampling from log-concave densities was also studied in Brooks (1998), where a strategy for defining the stopping rule was proposed. However, as the computational complexity of that strategy increases exponentially fast in the dimension p, its scope of applicability is limited.

5. Ozaki discretization and guarantees for smooth Hessian matrices

For convex log-densities f which are not only continuously differentiable but also have a smooth Hessian matrix $\nabla^2 f$, it is possible to take advantage of the Ozaki discretization (Ozaki, 1992) of the Langevin diffusion which is more accurate than the Euler discretization that was analysed in the foregoing sections. It consists in considering the diffusion process $\mathbf{D}^{\mathrm{O}}$ defined by expression (10) with the drift function

$$\mathbf{b}_t(\mathbf{D}^{\mathrm{O}})=-\sum_{k=0}^{K-1}\{\nabla f(\mathbf{D}^{\mathrm{O}}_{kh})+\nabla^2 f(\mathbf{D}^{\mathrm{O}}_{kh})(\mathbf{D}^{\mathrm{O}}_t-\mathbf{D}^{\mathrm{O}}_{kh})\}\mathbb{1}_{[kh,(k+1)h[}(t), \tag{22}$$

where, as previously, h is the step size and K is the number of iterations to attain the desired time horizon $T=Kh$. This expression leads to a diffusion process having linear drift function on each interval $[kh,(k+1)h[$. Such a diffusion admits a closed form formula. The resulting MCMC algorithm (Stramer and Tweedie, 1999b), hereafter referred to as the LMCO algorithm (for Langevin Monte Carlo with Ozaki discretization), is defined by an initial value $\bar{\boldsymbol{\vartheta}}^{(0)}$ and the following update rule. For every $k\geqslant 0$, we set $\mathbf{H}_k=\nabla^2 f(\bar{\boldsymbol{\vartheta}}^{(k,h)})$, which is an invertible $p\times p$ matrix since f is strongly convex, and define

$$\begin{aligned}\mathbf{M}_k&=\{\mathbf{I}_p-\exp(-h\mathbf{H}_k)\}\mathbf{H}_k^{-1},\\ \boldsymbol{\Sigma}_k&=\{\mathbf{I}_p-\exp(-2h\mathbf{H}_k)\}\mathbf{H}_k^{-1},\end{aligned} \tag{23}$$

$$\bar{\boldsymbol{\vartheta}}^{(k+1,h)}=\bar{\boldsymbol{\vartheta}}^{(k,h)}-\mathbf{M}_k\nabla f(\bar{\boldsymbol{\vartheta}}^{(k,h)})+\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}^{(k+1)}, \tag{24}$$

where $\{\boldsymbol{\xi}^{(k)}:k\in\mathbb{N}\}$ is a sequence of independent random vectors distributed according to the $\mathcal{N}_p(0,\mathbf{I}_p)$ distribution. In what follows, for any matrix $\mathbf{M}$, $\|\mathbf{M}\|$ stands for the spectral norm, i.e. $\|\mathbf{M}\|=\sup_{\|\mathbf{v}\|_2=1}\|\mathbf{M}\mathbf{v}\|_2$.

Theorem 3. Assume that $p \geqslant 2$, the function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies assumption (1) and, in addition, the Hessian matrix of f is Lipschitz continuous with some constant L_f: $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}')\| \leqslant L_f \|\mathbf{x} - \mathbf{x}'\|_2$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$. Let $\boldsymbol{\theta}^*$ be the global minimum point of f and ν be the Gaussian distribution $\mathcal{N}_p(\boldsymbol{\theta}^*, M^{-1}\mathbf{I}_p)$. Then, for any step size $h \leqslant 1/(8M)$ and any time horizon $T = Kh \geqslant 4/(3M)$, the total variation distance between the target distribution $\mathbf{P}_\pi$ and the approximation furnished by the LMCO algorithm $\nu\mathbf{P}_{\bar{\boldsymbol{\vartheta}}}^K$ with $\bar{\boldsymbol{\vartheta}}^{(0)}$ drawn at random from ν satisfies

$$\|\nu\mathbf{P}_{\bar{\boldsymbol{\vartheta}}}^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{M}{m}\right) - \frac{Tm}{2}\right\} + \{L_f^2 T h^2 p^2 (0.267 M^2 h T + 0.375)\}^{1/2}.$$

The proof of theorem 3 is deferred to Appendix A. Let us state now a direct consequence of theorem 3, which provides sufficient conditions on the number of steps for the LMCO algorithm to achieve a prescribed precision level ϵ. The proof of the corollary is trivial and, therefore, has been omitted.

Corollary 3. Let f satisfy assumption (1) with a Hessian that is Lipschitz continuous with constant L_f. For every $\epsilon \in (0, \frac{1}{2})$, if the time horizon T and the step size h are chosen so that

$$T \geqslant \frac{4\log(1/\epsilon) + p\log(M/m)}{2m},$$
$$h^{-1} \geqslant (6L_f M T p \epsilon^{-1})^{2/3} \bigvee (1.25\sqrt{T L_f p \epsilon^{-1}}) \bigvee (8M),$$

then the distribution of the outcome of the LMCO algorithm with $K = [T/h]$ steps fulfils $\|\nu\mathbf{P}_{\bar{\boldsymbol{\vartheta}}}^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \varepsilon$.

This corollary provides a simple recommendation for the choice of the parameters h and T in the LMCO algorithm. It also ensures that, for the recommended choice of the parameters, it is sufficient to perform $K = O[\{p + \log(1/\epsilon)\}^{3/2} p\epsilon^{-1}]$ steps of the LMCO algorithm to reach the precision level ϵ desired. This number is much smaller than that provided earlier by corollary 1, which was of order $O[\{p + \log(1/\epsilon)\}^2 p\epsilon^{-2}]$. However, one should pay attention to the fact that each iteration of the LMCO algorithm requires computing the exponential of the Hessian of f at the current state and, therefore, the computational complexity of each iteration is usually much larger for the LMCO as compared with the LMC algorithm ($O(p^3)$ *versus* $O(p)$). This implies that the LMCO algorithm would most probably be preferable to the LMC algorithm only in situations where p is not too large, and the precision level ϵ required is very small. For instance, the arguments of this paragraph advocate using the LMCO instead of the LMC algorithm when $\epsilon = o(p^{-3/2})$.

This being said, it is worth noting that for some functions f the cost of performing a singular value decomposition on the Hessian of f, which is the typical way to compute the matrix exponential, might be much smaller than the aforementioned worst-case complexity $O(p^3)$. This is, in particular, the case for the first example that is considered in the next section. One can also approximate the matrix exponentials by matrix polynomials. For second-order polynomials, this amounts to replacing the updates (24) by

$$\bar{\boldsymbol{\vartheta}}^{(k+1,h)} = \bar{\boldsymbol{\vartheta}}^{(k,h)} - h(\mathbf{I}_p - \tfrac{1}{2}h\mathbf{H}_k)\nabla f(\bar{\boldsymbol{\vartheta}}^{(k,h)}) + \sqrt{(2h)}(\mathbf{I}_p - \tfrac{1}{2}h\mathbf{H}_k)\boldsymbol{\xi}^{(k+1)}. \tag{25}$$

Establishing guarantees for such a modified LMCO algorithm is out of the scope of the present work. We shall limit ourselves to an empirical assessment of the quality of this approximation on the example of logistic regression that is considered in Section 6.

To close this section, we remark that, in the case that a warm start is available (see Section 4.1), the number of iterations for the LMCO algorithm to reach precision ϵ may be reduced to $O^*(p\epsilon^{-1})$. Indeed, if the χ^2-divergence between the initial distribution and the target is bounded by a quantity that is independent of p, or increasing not faster than a polynomial in p, then the time horizon can be chosen as $O^*(1)$ and the choice of h that is provided by corollary 3 leads to a number of iterations K satisfying $K = O^*(p\epsilon^{-1})$.

6. Numerical experiments

To illustrate the results that were established in the previous sections, we carried out some experiments on synthetic data. The experiments were conducted on a Hewlett–Packard Elitebook personal computer with the following configuration: Intel (R) Core^{TM} i7-3687U processor with a 2.6-GHz central processor unit and 16 Gbytes of random-access memory. The code, written in MATLAB, does not use parallelization. We considered two examples; both satisfy all the assumptions that were required in previous sections. This implies that corollaries 1 and 3 apply and guarantee that the choices of h and T that are suggested by these corollaries allow us to generate random vectors having a distribution which is within a prescribed distance ϵ, in total variation, of the target distribution.

6.1. Example 1: Gaussian mixture

The goal of this first experiment is merely to show on a simple example the validity of our theoretical findings, i.e. we check below that the LMC and the LMCO algorithms with the values of time horizon T and step size h recommended by corollaries 1 and 3 produce samples distributed approximately as the target distribution within a reasonable running time. For this, we consider the simple task of sampling from the density π defined by

$$\pi(\mathbf{x}) = \frac{1}{2(2\pi)^{p/2}}\{\exp(-\|\mathbf{x}-\mathbf{a}\|_2^2/2) + \exp(-\|\mathbf{x}+\mathbf{a}\|_2^2/2)\}, \qquad \mathbf{x} \in \mathbb{R}^p, \tag{26}$$

where $\mathbf{a} \in \mathbb{R}^p$ is a given vector. This density π represents the mixture with equal weights of two Gaussian densities $\mathcal{N}(\mathbf{a}, \mathbf{I}_p)$ and $\mathcal{N}(-\mathbf{a}, \mathbf{I}_p)$. The function f, its gradient and its Hessian are given by

$$\begin{aligned} f(\mathbf{x}) &= \tfrac{1}{2}\|\mathbf{x}-\mathbf{a}\|_2^2 - \log\{1+\exp(-2\mathbf{x}^{\mathrm{T}}\mathbf{a})\}, \\ \nabla f(\mathbf{x}) &= \mathbf{x} - \mathbf{a} + 2\mathbf{a}\{1+\exp(2\mathbf{x}^{\mathrm{T}}\mathbf{a})\}^{-1}, \\ \nabla^2 f(\mathbf{x}) &= \mathbf{I}_p - 4\mathbf{a}\mathbf{a}^{\mathrm{T}}\exp(2\mathbf{x}^{\mathrm{T}}\mathbf{a})\{1+\exp(2\mathbf{x}^{\mathrm{T}}\mathbf{a})\}^{-2}. \end{aligned}$$

Using the fact that $0 \leqslant 4\exp(2\mathbf{x}^{\mathrm{T}}\mathbf{a})\{1+\exp(2\mathbf{x}^{\mathrm{T}}\mathbf{a})\}^{-2} \leqslant 1$, we infer that, for $\|\mathbf{a}\|_2 < 1$, the function f is strongly convex and satisfies assumption (1) with $m = 1 - \|\mathbf{a}\|_2^2$ and $M = 1$. Furthermore, the Hessian matrix is Lipschitz continuous with the constant $L_f = \frac{1}{2}\|\mathbf{a}\|_2^3$. Hence, both algorithms that were explored in the previous sections, LMC and LMCO, can be used for sampling from the density π defined by expression (26). Note also that we can sample directly from π by drawing independently at random a Bernoulli($\frac{1}{2}$) random variable Y and a standard Gaussian vector $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_p)$ and by computing $\mathbf{X} = Y(\mathbf{Z}-\mathbf{a}) + (1-Y)(\mathbf{Z}+\mathbf{a})$. The density of the random vector $\mathbf{X}$ defined in such a way coincides with π. One can check that the unique minimum of f is achieved at $\boldsymbol{\theta}^* = c^*\mathbf{a}$, where c^* is the unique solution of the equation $c = 1 - 2\{1+\exp(2c\|\mathbf{a}\|_2^2)\}^{-1}$. Choosing $\mathbf{a}$ so that $\|\mathbf{a}\|_2^2 = \frac{1}{2}$, we obtain $\boldsymbol{\theta}^* = 0$.

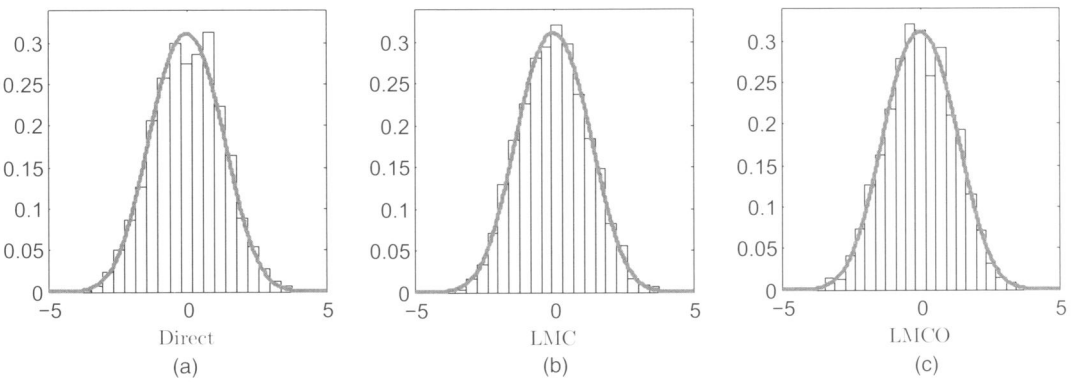

Fig. 1. Histograms of the one-dimensional projections of the samples computed by using (a) the direct, (b) LMC and (c) LMCO algorithms in the example of a Gaussian mixture (26): the dimension is $p = 8$, the target precision is $\epsilon = 0.1$ and 2500 independent samples were drawn according to each of the three methods; the results show that both the LMC and the LMCO algorithms are very accurate (nearly as accurate as the direct method)

Table 2. Number of iterations and running times in example 1

Algorithm	*Results for the following values of p:*							
	$p=4$	$p=8$	$p=12$	$p=16$	$p=20$	$p=30$	$p=40$	$p=60$
Approximate number of iterates, K (to be multiplied by 10^3)								
LMC	28	87	184	329	532	1350	2728	7741
LMCO	1	3	5.4	9	13.6	30	54.9	133
Running times (s) for $N = 10^3$ samples								
LMC	3.44	16.6	54.1	123	238	876	2488	9789
LMCO	0.18	0.70	1.78	3.5	6.4	20.4	53.9	189.1

In the experiment that is depicted in Fig. 1 (see also Table 2), we chose $\epsilon = 0.1$ and, for dimensions $p \in \{4, 8, 12, 16, 20, 30, 40, 60\}$, generated vectors by using the direct method, the LMC algorithm and the LMCO algorithm. Let $\boldsymbol{\vartheta}^{\mathrm{direct},i}$, $\boldsymbol{\vartheta}^{\mathrm{LMC},i}$ and $\boldsymbol{\vartheta}^{\mathrm{LMCO},i}$, $i = 1, \ldots, N$, be the vectors that are obtained after N repetitions of this experiment. In Fig. 1, we plot the histograms of the one-dimensional projections $\mathbf{v}^{\mathrm{T}}\boldsymbol{\vartheta}^{\mathrm{direct},i}$, $\mathbf{v}^{\mathrm{T}}\boldsymbol{\vartheta}^{\mathrm{LMC},i}$ and $\mathbf{v}^{\mathrm{T}}\boldsymbol{\vartheta}^{\mathrm{LMCO},i}$ of the sampled vectors onto the direction $\mathbf{v} = \mathbf{a}/\|\mathbf{a}\|_2$ in $\mathbb{R}^p$ determined by the vector $\mathbf{a}$. To provide a qualitative measure of accuracy of the samples obtained, we added to each histogram the curve of the true density, which can be computed analytically and is equal to a mixture with equal weights of two one-dimensional Gaussian densities. The result shows that both the LMC algorithm and the LMCO algorithm are very accurate: nearly as accurate as the direct method.

To illustrate the dependence on the dimension p of the computational complexity of the sampling strategies proposed, we report in Table 2 the number of iterations and the overall running times for generating $N = 10^3$ independent samples by the LMC and the LMCO algorithms for the target specified by expression (26), when the dimension p varies in $\{4, 8, 12, 16, 20, 30, 40, 60\}$. One may observe that the computational time is much smaller for the LMCO than for the LMC algorithm, which is mainly explained by the fact that the singular vectors of the Hessian of the

function f, in the example under consideration, do not depend on the value $\mathbf{x}$ at which the Hessian is computed.

This example confirms our theoretical findings in that it shows that

(a) the samples that are drawn from the LMC and the LMCO algorithms with the parameters T and h suggested by theoretical considerations have distributions that are very close to the target distribution and
(b) the running times for these algorithms remain reasonable even for moderately large values of dimension p.

6.2. *Example 2: binary logistic regression*

We consider the problem of logistic regression, in which an independently and identically distributed sample $\{(\mathbf{X}_i, Y_i)\}_{i=1,\dots,n}$ is observed, with features $\mathbf{X}_i \in \mathbb{R}^p$ and binary labels $Y_i \in \{0, 1\}$. The goal is to estimate the conditional distribution of Y_1 given $\mathbf{X}_1$, which amounts to estimating the regression function $r(\mathbf{x}) = \mathbf{E}[Y_1|\mathbf{X}_1 = \mathbf{x}] = \mathbf{P}(Y_1 = 1|\mathbf{X}_1 = \mathbf{x})$. In the model of logistic regression, the regression function $r(\mathbf{x})$ is approximated by a logistic function of the form $r(\boldsymbol{\theta}, \mathbf{x}) = \exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x})/\{1 + \exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x})\}$. The Bayesian approach for estimating the parameter $\boldsymbol{\theta}$ relies on introducing a prior probability density on $\boldsymbol{\theta}$, $\pi_0(\cdot)$, and by computing the posterior density $\pi(\cdot)$. Choosing a Gaussian prior π_0 with zero mean and covariance matrix proportional to the inverse of the Gram matrix $\boldsymbol{\Sigma}_{\mathbf{X}} = (1/n)\Sigma_{i=1}^n \mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}$, the posterior density takes the form

$$\pi(\boldsymbol{\theta}) \propto \exp\Big[-\mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} - \sum_{i=1}^n \log\{1 + \exp(-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\} - \frac{\lambda}{2}\|\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}\boldsymbol{\theta}\|_2^2\Big], \tag{27}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}} \in \{0, 1\}^n$ and $\mathbf{X}$ is the $n \times p$ matrix having the feature $\mathbf{X}_i$ as ith row. The first two terms in the exponential correspond to the log-likelihood of the logistic model, whereas the last term comes from the log-density of the prior and can be seen as a penalty term. The parameter $\lambda > 0$ is usually specified by the practitioner. Many researchers have studied this model from a Bayesian perspective (see for instance Holmes and Held (2006) and Roy (2012)), and it seems that there is no compelling alternative to MCMC algorithms for computing the Bayesian estimators in this model. Furthermore, even for the MCMC approach, although geometric ergodicity under some strong assumptions is established, there is no theoretically justified rule for assessing the convergence and, especially, ensuring that the convergence is achieved in polynomial time. Such guarantees are provided by our results, when either the LMC or the LMCO algorithm is used.

If we define the function f by

$$f(\boldsymbol{\theta}) = \mathbf{Y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} + \sum_{i=1}^n \log\{1 + \exp(-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\} + \frac{\lambda}{2}\|\boldsymbol{\Sigma}_{\mathbf{X}}^{1/2}\boldsymbol{\theta}\|_2^2, \tag{28}$$

we obtain the setting that was described in Section 1. It is useful here to apply the preconditioning technique of Section 4.2 with the preconditioner $\mathbf{A} = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}$. Thus, the LMC and the LMCO algorithms can be used with the function f replaced by $g(\boldsymbol{\theta}) = f(\mathbf{A}\boldsymbol{\theta})$. We check that g and f are infinitely differentiable and

$$\nabla f(\boldsymbol{\theta}) = \mathbf{X}^{\mathrm{T}}\mathbf{Y} - \sum_{i=1}^n \frac{\mathbf{X}_i}{1 + \exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)} + \lambda\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\theta},$$

$$\nabla^2 f(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{\exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)}{\{1 + \exp(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}_i)\}^2}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}} + \lambda\boldsymbol{\Sigma}_{\mathbf{X}}.$$

For the function g, since $\nabla^2 g(\boldsymbol{\theta}) = \mathbf{A}\nabla^2 f(\mathbf{A}\boldsymbol{\theta})\mathbf{A}$, we can infer from these relationships that assumption (1) holds with $m_{\mathbf{A}} = \lambda$ and $M_{\mathbf{A}} = \lambda + 0.25n$. Note here that, if we do not use any preconditioner, the constants m and M would be given by $m = \lambda\nu_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})$ and $M = (\lambda + 0.25n)\,\nu_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})$, where $\nu_{\min}(\boldsymbol{\Sigma})$ and $\nu_{\max}(\boldsymbol{\Sigma})$ are respectively the smallest and the largest eigenvalues of $\boldsymbol{\Sigma}$. This implies that the ratio $\nu_{\max}(\boldsymbol{\Sigma}_{\mathbf{X}})/\nu_{\min}(\boldsymbol{\Sigma}_{\mathbf{X}})$ quantifies the gain in efficiency that is obtained by preconditioning. This ratio might be large especially when p is large and the covariates are strongly correlated.

Furthermore, $\nabla^2 g$ is Lipschitz with a constant L_g provided by the following formula (the proof of which is postponed to Appendix A):

$$L_g = 0.1\left\|\sum_{i=1}^{n}\|\mathbf{A}\mathbf{X}_i\|_2\mathbf{A}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbf{A}\right\| \leqslant 0.1n\max_{i=1,\dots,n}\|\boldsymbol{\Sigma}_{\mathbf{X}}^{-1/2}\mathbf{X}_i\|_2. \tag{29}$$

In our second experiment, for a set of values of p and n, we randomly drew n independently and identically distributed samples $(\mathbf{X}_i, Y_i)$ according to the following data-generating device. The features $\mathbf{X}_i$ were drawn from a Rademacher distribution (i.e. each co-ordinate takes the values ± 1 with probability $\frac{1}{2}$), and then renormalized to have a Euclidean norm equal to 1. Each label Y_i, given $\mathbf{X}_i = \mathbf{x}$, was drawn from a Bernoulli distribution with parameter $r(\boldsymbol{\theta}^{\mathrm{true}}, \mathbf{x})$. The true vector $\boldsymbol{\theta}^{\mathrm{true}}$ was set to $\mathbf{1}_p = (1, 1, \dots, 1)^{\mathrm{T}}$. For each value of p and n, we generated 100 samples $(\mathbf{X}, \mathbf{Y})$. For each sample, we computed the maximum likelihood estimate by using the gradient descent as described in theorem 1 with a level of precision $\epsilon = 10^{-6}$. Following the recommendation of Hanson *et al.* (2014), the parameter λ was set to $3p/\pi^2$. We carried out two subexperiments with well-specified distinct purposes: to assess empirically the gain that is obtained by applying the trick of strong convexification described in Section 4.3 and to evaluate the loss of accuracy that is caused by applying to the LMCO algorithm the second-order approximation (25).

In the first subexperiment, we applied the strategy that was outlined in Section 4.3 for various values of n, p and ϵ. For this, we exploited the formulae

$$m_R = \lambda + \nu_{\min}(\mathbf{B}_R), \qquad \mathbf{B}_R := \sum_{i=1}^{n}\frac{\exp(|\mathbf{X}_i^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}^*| + R\|\mathbf{A}\mathbf{X}_i\|_2)}{[1+\exp\{2|\mathbf{X}_i^{\mathrm{T}}\mathbf{A}\boldsymbol{\theta}^*| + 2R\|\mathbf{A}\mathbf{X}_i\|_2\}]^2}\mathbf{A}\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\mathbf{A},$$

$$(p\mu_R)^2 = \frac{2(M/2)^{p/2}}{(m_R R^2)^{p+4}\Gamma(p/2)}\sum_{j=0}^{4}C_4^j(-m_R R^2)^j\Gamma(p+4-j; m_R R^2),$$

where $\Gamma(p; x) = \int_x^\infty t^{p-1}\exp(-t)\mathrm{d}t$ is the upper incomplete gamma function and C_4^j stands for the binomial coefficient. The proof of the fact that the quantities m_R and μ_R that are defined by these formulae satisfy all the assumptions of Section 4.3 is provided in the on-line supplementary material. In this experiment, we used two values of ϵ (0.1 and 0.01), three values of dimension p (2, 5 and 20) and five values for the sample size n (500, 1000, 2000, 4000 and 8000). We report in Table 3 the number of iterates by using the LMC algorithm, K, and the average number of iterates of the modified LMC algorithm as described in Section 4.3, K'. Note that, in the case of the modified LMC algorithm, the number of iterates depends on the original data $(\mathbf{X}, \mathbf{Y})$. Therefore, the numbers K' that are reported in Table 3 are those obtained by averaging over 100 independent trials.

The results of Table 3 show clearly the advantage of using the strong convexification trick. For instance, when $\epsilon = 0.1$, $p = 5$ and $n = 1000$, the gain is very impressive since the number of iterations is reduced from nearly 7.5×10^5 to 2.2×10^3. This represents a reduction by a factor that is close to 340. The gain is less significant in the case when the ratio p/n is larger. Our explanation of this is that, for a small ratio p/n, the posterior density has a very strong peak at

Table 3. Example 2 (binary logistic regression): number of iterates by using the LMC algorithm, K, and its modified version as described in Section 4.3, K'

p		*Results for the following values of n:*				
		$n=500$	$n=1000$	$n=2000$	$n=4000$	$n=8000$
$\epsilon=0.1$						
2	K	0.065×10^7	0.137×10^7	0.286×10^7	0.596×10^7	1.236×10^7
	K'	2.823×10^2	0.688×10^2	0.230×10^2	0.089×10^2	0.039×10^2
5	K	0.358×10^6	0.751×10^6	1.568×10^6	3.257×10^6	6.742×10^6
	K'	4.207×10^4	0.222×10^4	0.029×10^4	0.007×10^4	0.003×10^4
20	K	0.135×10^6	0.279×10^6	0.579×10^6	1.201×10^6	2.481×10^6
	K'	0.121×10^6	0.250×10^6	0.519×10^6	1.075×10^6	2.222×10^6
$\epsilon=0.01$						
2	K	0.151×10^9	0.313×10^9	0.645×10^9	1.324×10^9	2.714×10^9
	K'	1.529×10^5	0.248×10^5	0.077×10^5	0.028×10^5	0.011×10^5
5	K	0.075×10^9	0.155×10^9	0.320×10^9	0.657×10^9	1.345×10^9
	K'	3.652×10^7	0.087×10^7	0.011×10^7	0.002×10^7	0.001×10^7
20	K	0.254×10^8	0.518×10^8	1.062×10^8	2.177×10^8	4.459×10^8
	K'	0.227×10^8	0.463×10^8	0.947×10^8	1.941×10^8	3.975×10^8

its mode. Therefore, even for a relatively large radius R the condition number M/m_R is not too large. Thus, small p/n is the typical situation in which the strong convexification trick is likely to lead to considerable savings in running time.

In the second subexperiment, we aimed at verifying the validity of the second-order approximation of the LMCO algorithm, hereafter referred to as the LMCO′ algorithm, obtained by applying the update rule (25). For this, for $\epsilon=0.1$, $p\in\{2,5,10\}$ and $n\in\{200,300,400,500\}$, we generated $N_{\mathrm{MC}}=100$ Monte Carlo samples by using the LMC algorithm and the LMCO′ algorithm. To check the closeness of the distributions of these two p-dimensional samples, we compared several aspects of them. More precisely, we compared their marginal means, marginal medians and marginal quartiles. Mathematically speaking, for each data set $\mathcal{D}_{\mathrm{data}}=(\mathbf{X},\mathbf{Y})$, we generated N_{MC} samples $\mathcal{D}_{\mathrm{MC}}=\{\boldsymbol{\theta}^1,\ldots,\boldsymbol{\theta}^{N_{\mathrm{MC}}}\}$ and $\bar{\mathcal{D}}_{\mathrm{MC}}=\{\bar{\boldsymbol{\theta}}^1,\ldots,\bar{\boldsymbol{\theta}}^{N_{\mathrm{MC}}}\}$ by using the LMC and the LMCO′ algorithms respectively. We then computed the normalized distance between their marginal means: $d_{\mathrm{mean}}=(1/p)\|\mathrm{mean}(\mathcal{D}_{\mathrm{MC}})-\mathrm{mean}(\bar{\mathcal{D}}_{\mathrm{MC}})\|_1$. We also computed the quantities d_{median}, d_{Q_1} and d_{Q_3}, which are defined analogously by replacing the mean by the coordinatewise median, first quartile and third quartile respectively. The idea for considering these quantities is that, for large N_{MC} and small ϵ, all the aforementioned distances should be close to 0.

We opted for the boxplot representation of 100 values of each of these distances obtained over 100 independent replications of the data set $\mathcal{D}_{\mathrm{data}}$. These boxplots are drawn in Fig. 2. They show that the distances are small—at most of the order of 10^{-1}—which may be considered as an argument in favour of the modification that is proposed in rule (25). Indeed, with $\epsilon=0.1$ and $N_{\mathrm{MC}}=100$, we could not expect to have an error of smaller order. This is very promising since this modified LMCO algorithm has a significantly smaller computational complexity than the original LMCO algorithm: each iteration has a worst-case accuracy $O(p^2)$ instead of $O(p^3)$, thanks to the fact that matrix exponentials as well as the inversion of the Hessian are replaced by the computation of the Hessian and its product with vectors.

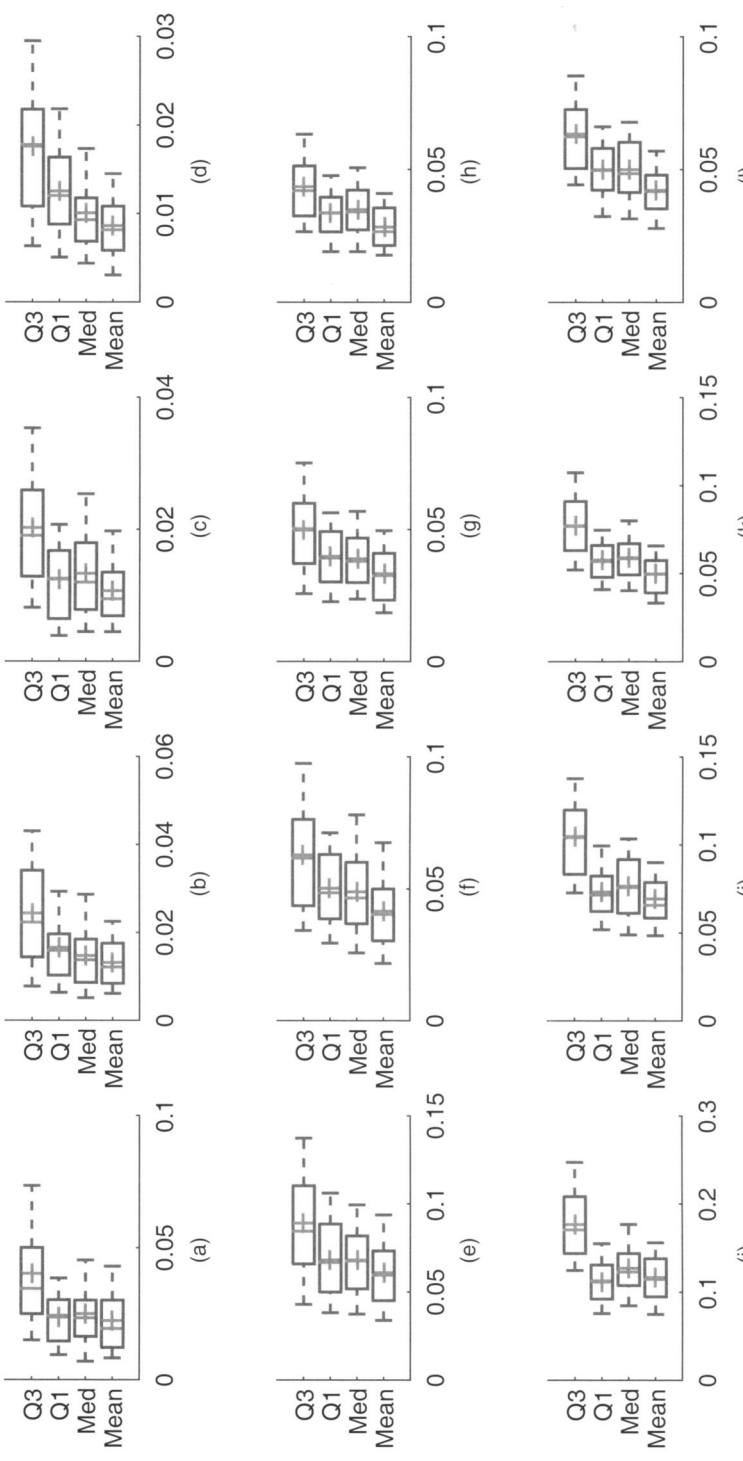

Fig. 2. Boxplots of the second subexperiment carried out within the model of logistic regression: (a) $n = 200, p = 2$; (b) $n = 300, p = 2$; (c) $n = 400, p = 2$; (d) $n = 500, p = 2$; (e) $n = 200, p = 5$; (f) $n = 300, p = 5$; (g) $n = 400, p = 5$; (h) $n = 500, p = 5$; (i) $n = 200, p = 10$; (j) $n = 300, p = 10$; (k) $n = 400, p = 10$; (l) $n = 500, p = 10$

7. Summary and conclusion

We have established easy-to-implement, non-asymptotic theoretical guarantees for approximate sampling from log-concave and strongly log-concave probability densities. For this, we have analysed the LMC algorithm and its Ozaki discretized version LMCO. These algorithms can be regarded as the natural counterparts—when the task of optimization is replaced by the task of sampling—of the gradient descent algorithm, which is widely studied in convex optimization. Despite its broad applicability in the framework of Bayesian statistics and beyond, to the best of our knowledge, there are no theoretical results in the literature proving that the computational complexity of the aforementioned algorithms scales at most polynomially in dimension and in ϵ^{-1}, the inverse of the precision level desired. The results that are proved in the present work fill this gap by showing that, to achieve a precision (in total variation) bounded from above by ϵ, the LMC algorithm needs no more than $O[\epsilon^{-2}\{p^3 + p\log(\epsilon^{-1})\}]$ evaluations of the gradient when the target density is strongly log-concave and $O\{\epsilon^{-4}p^5\log^2(p \vee \epsilon^{-1})\}$ evaluations of the gradient when the target density is non-strongly log-concave. Further improvement of the rates can be achieved if a warm start is available. More precisely, if there is an efficiently samplable distribution ν such that the χ^2-divergence between ν and the target scales polynomially in p, then the LMC algorithm with an initial value drawn from ν needs no more than $O\{\epsilon^{-2}p\log^2(p \vee \epsilon^{-1})\}$ evaluations of the gradient when the target density is strongly log-concave and $O\{\epsilon^{-4}p^3\log^2(p \vee \epsilon^{-1})\}$ gradient evaluations when the target density is non-strongly log-concave. An important advantage of our results is that all the bounds come with explicit numerical constants of reasonable magnitude.

The search for tractable theoretical guarantees for MCMC algorithms is an active topic of research not only in probability and statistics but also in theoretical computer science and in machine learning. To the best of our knowledge, first computable bounds on the constants that are involved in the geometric convergence of Markov chains were derived in Meyn and Tweedie (1994); see also the subsequent work Rosenthal (2002), Douc *et al.* (2004) and Roberts and Rosenthal (2004). However, because of the broad generality of the Markov processes considered, their results are difficult to implement for obtaining tight bounds on the constants in the context of high dimensionality. In particular, we did not succeed in deriving from their results convergence rates for the LMC algorithm (nor for its Metropolis–Hastings-adjusted version, the Metropolis-adjusted Langevin algorithm) that are polynomial in the dimension p and hold for every strongly log-concave target density. Note also that some non-asymptotic convergence results for the the Metropolis-adjusted Langevin algorithm were obtained by Bou-Rabee and Hairer (2013), where strongly log-concave four times continuously differentiable functions f were considered. Unfortunately, the constants that are involved in their bounds are not explicit and cannot be used for our purposes.

The problem of sampling from log-concave distributions is not new. It has been considered in the early references Frieze *et al.* (1994) and Frieze and Kannan (1999). Important progress in this topic was made by Lovász and Vempala (see, in particular, Lovász and Vempala (2006a,b) for the sharpest results), which are perhaps the closest to our work. They investigated the problem of sampling from a log-concave density with a compact support and derived non-asymptotic bounds on the number of steps that are sufficient for approximating the target density; the best bounds are obtained for the hit-and-run algorithm. The analysis that they carried out is very different from that presented in the present work and the constants in their results are prohibitively large (for instance, 10^{31} in Lovász and Vempala (2006b), corollary 1.2), which makes the established guarantees of little interest in practice. On the positive side, one of the most remarkable features of the results that were proved in Lovász and Vempala (2006a,b) is that the

number of steps required to achieve the level ϵ scales polylogarithmically in $1/\epsilon$. This is of course much better than the dependence on ϵ in our bounds. However, the logarithm of $1/\epsilon$ in their result is raised to power 5, which for most interesting values of ϵ behaves itself as a linear function of $1/\epsilon$. On the downside, the dependence on the dimension in the results of Lovász and Vempala (2006a,b), when no warm start is available, scales as p^4, which is worse than p^3 inferred from our analysis. A difference worth stressing between our framework and that of Lovász and Vempala (2006a,b) is that the LMC algorithm which we have analysed here is based on the evaluations of the gradient of f, whereas the algorithms that were studied in Lovász and Vempala (2006a,b) need to sample from the restriction of π_f on the lines. On a related note, building on the results by Lovázs and Vempala, Belloni and Chernozhukov (2009) provided polynomial guarantees for sampling from a distribution which converges asymptotically to a Gaussian distribution.

After the submission of the present paper, the manuscript Durmus and Moulines (2015) was posted on arXiv, which refines our results in various directions. In particular, they managed to assess more accurately the effect of the initial distribution on the final precision of the LMC algorithm and they investigated an Euler scheme with non-constant step size. Roughly speaking, they have proved that the rate which we obtained in the case of a warm start is valid for any starting point which is not too far from the mode of the density. On a related note, we focus in the present work only on the total variation distance between some MCMC algorithms and the target distribution, whereas in many applications one may be interested only in approximating integrals with respect to the target distribution. Clearly, guarantees on the total variation distance imply guarantees on the approximations of integrals, at least when the integrands are bounded functions. However, since the problem of approximating integrals is, in some sense, easier than sampling from a distribution, one could hope to obtain tighter bounds for the former problem. This and related questions are thoroughly investigated in Durmus and Moulines (2015).

Although the main contribution of the present work is of a theoretical nature, we can also draw some conclusions which might be of interest for practitioners. First, our results show that the heuristic choice of the stopping rule for the MCMC algorithms is not the only possible option: it is also possible to have theoretically grounded guidelines for choosing the stopping time. The resulting algorithm will be of polynomial complexity both in dimension and in the level of precision. Second, the results that are reported in this work show that there is no need to apply a Metropolis–Hastings correction to the Langevin algorithm and its various variants to ensure their convergence. Third, when the dimension is not very high and a high level of precision is required (i.e. when $p^{3/2}\epsilon$ is small), the LMCO algorithm is preferable to the LMC algorithm, and the modified LMCO algorithm using the update rule of equation (25) is even better. Note, however, that this last claim has been checked empirically but comes without any theoretical justification.

Finally, we mention that, in recent years, several studies making the connection between convex optimization and MCMC algorithms have been carried out. They have mainly focused on proposing new algorithms of approximate sampling (Girolami and Calderhead, 2011; Schreck *et al.*, 2013; Pereyra, 2014) inspired by the ideas coming from convex optimization. We hope that the present work will stimulate a more extensive investigation of the relationship between approximate sampling and optimization, especially in the aim of establishing easy-to-use theoretical guarantees for the MCMC algorithms.

Acknowledgements

The work of the author was partially supported by an Investissements d'Avenir grant (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

Appendix A: Postponed proofs and some technical results

A.1. Auxiliary results

Lemma 4 (lemma 1.2.3 in Nesterov (2004)). If the function f satisfies the second inequality in expression (1), then $f(\boldsymbol{\theta}) - f(\bar{\boldsymbol{\theta}}) - \nabla f(\bar{\boldsymbol{\theta}})^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \leqslant (M/2)\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2, \forall \boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^p$.

Lemma 5. Denote by $\nu_{h,\mathbf{x}}$ the conditional density of $\boldsymbol{\vartheta}^{(1,h)}$ given $\boldsymbol{\vartheta}^{(0)} = \mathbf{x}$, where the sequence $\{\boldsymbol{\vartheta}^{(k,h)}\}_{k\in\mathbb{N}}$ is defined by expression (5) with a function f satisfying condition (1). (In other terms, $\nu_{h,\mathbf{x}}$ is the density of the Gaussian distribution $\mathcal{N}\{\mathbf{x} - h\nabla f(\mathbf{x}), 2h\mathbf{I}_p\}$.) If $h \leqslant 1/(2M)$ then

$$\mathbf{E}_\pi\left[\frac{\nu_{h,\mathbf{x}}(\boldsymbol{\vartheta})^2}{\pi(\boldsymbol{\vartheta})^2}\right] \leqslant \exp\left\{\frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2 - \frac{p}{2}\log(2hm)\right\}.$$

Proof. In view of the relations

$$\begin{aligned}
\pi(\boldsymbol{\theta})^{-1} &= \exp\{f(\boldsymbol{\theta})\}\int_{\mathbb{R}^p}\exp\{-f(\bar{\boldsymbol{\theta}})\}\,\mathrm{d}\bar{\boldsymbol{\theta}} = \exp\{f(\boldsymbol{\theta}) - f(\mathbf{x})\}\int_{\mathbb{R}^p}\exp\{-f(\bar{\boldsymbol{\theta}}) + f(\mathbf{x})\}\mathrm{d}\bar{\boldsymbol{\theta}} \\
&\leqslant \exp\left\{\nabla f(\mathbf{x})^{\mathrm{T}}(\boldsymbol{\theta} - \mathbf{x}) + \frac{M}{2}\|\boldsymbol{\theta} - \mathbf{x}\|_2^2\right\}\int_{\mathbb{R}^p}\exp\left\{-\nabla f(\mathbf{x})^{\mathrm{T}}(\bar{\boldsymbol{\theta}} - \mathbf{x}) - \frac{m}{2}\|\bar{\boldsymbol{\theta}} - \mathbf{x}\|_2^2\right\}\mathrm{d}\bar{\boldsymbol{\theta}} \\
&\leqslant \left(\frac{2\pi}{m}\right)^{p/2}\exp\left\{\nabla f(\mathbf{x})^{\mathrm{T}}(\boldsymbol{\theta} - \mathbf{x}) + \frac{M}{2}\|\boldsymbol{\theta} - \mathbf{x}\|_2^2 + \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2\right\}
\end{aligned}$$

we have

$$\begin{aligned}
\mathbf{E}_\pi\left[\frac{\nu_{h,\mathbf{x}}(\boldsymbol{\vartheta})^2}{\pi(\boldsymbol{\vartheta})^2}\right] &= (4\pi h)^{-p}\int_{\mathbb{R}^p}\exp\left\{-\frac{1}{2h}\|\boldsymbol{\theta} - \mathbf{x} + h\nabla f(\mathbf{x})\|_2^2\right\}\pi(\boldsymbol{\theta})^{-1}\mathrm{d}\boldsymbol{\theta} \\
&\leqslant (4\pi h)^{-p}\left(\frac{2\pi}{m}\right)^{p/2}\exp\left\{\frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2\right\}\int_{\mathbb{R}^p}\exp\left\{-\frac{(1 - hM)\|\boldsymbol{\theta} - \mathbf{x}\|_2^2}{2h}\right\}\mathrm{d}\boldsymbol{\theta} \\
&= (4\pi h)^{-p}\left(\frac{2\pi}{m}\right)^{p/2}(2\pi h)^{p/2}(1 - hM)^{-p/2}\exp\left(\frac{1}{2m}\right)\|\nabla f(\mathbf{x})\|_2^2.
\end{aligned}$$

After a suitable rearrangement of the terms we obtain the claim of lemma 5.

A.2. Proofs of results concerning the Langevin Monte Carlo algorithm

Instead of proving proposition 1, we prove below the following stronger result.

Proposition 3. Let the function f be continuously differentiable on $\mathbb{R}^p$ and satisfy condition (1) with $f^* = \inf_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})$. Then, for every $h \leqslant 1/M$, we have

$$\mathbf{E}[f(\boldsymbol{\vartheta}^{(k,h)}) - f^*] \leqslant (1 - mh)^k\,\mathbf{E}[f(\boldsymbol{\vartheta}^{(0)}) - f^*] + \frac{Mp}{m(2 - Mh)}, \tag{30}$$

$$\mathbf{E}[\|\boldsymbol{\vartheta}^{(k,h)} - \boldsymbol{\theta}^*\|_2^2] \leqslant \frac{M\exp(-mhk)}{m}\mathbf{E}[\|\boldsymbol{\vartheta}^{(0)} - \boldsymbol{\theta}^*\|_2^2] + \frac{2Mp}{m^2(2 - Mh)}. \tag{31}$$

Proof. Throughout this proof, we use the shorthand notation $f^{(k)} = f(\boldsymbol{\vartheta}^{(k,h)})$ and $\nabla f^{(k)} = \nabla f(\boldsymbol{\vartheta}^{(k,h)})$. In view of the relation (5) and the Taylor series expansion, we have

$$\begin{aligned}
f^{(k+1)} &\leqslant f^{(k)} + (\nabla f^{(k)})^{\mathrm{T}}(\boldsymbol{\vartheta}^{(k+1,h)} - \boldsymbol{\vartheta}^{(k,h)}) + \frac{M}{2}\|\boldsymbol{\vartheta}^{(k+1,h)} - \boldsymbol{\vartheta}^{(k,h)}\|_2^2 \\
&= f^{(k)} - h\|\nabla f^{(k)}\|_2^2 + \sqrt{(2h)}\,(\nabla f^{(k)})^{\mathrm{T}}\boldsymbol{\xi}^{(k+1)} + \frac{M}{2}\|h\nabla f^{(k)} - \sqrt{(2h)}\boldsymbol{\xi}^{(k+1)}\|_2^2.
\end{aligned}$$

Taking the expectations of both sides, we obtain

$$\begin{aligned}\mathbf{E}[f^{(k+1)}] &\leqslant \mathbf{E}[f^{(k)}] - h\mathbf{E}[\|\nabla f^{(k)}\|_2^2] + \frac{M}{2}h^2\mathbf{E}[\|\nabla f^{(k)}\|_2^2] + Mhp \\ &= \mathbf{E}[f^{(k)}] - \tfrac{1}{2}h(2-Mh)\mathbf{E}[\|\nabla f^{(k)}\|_2^2] + Mhp. \end{aligned} \tag{32}$$

It is well known (see, for instance, Boyd and Vandenberghe (2004)) that, for the global minimum f^* of f over $\mathbb{R}^p$, we have

$$\|\nabla f(\mathbf{x})\|_2^2 \geqslant 2m\{f(\mathbf{x}) - f^*\}, \qquad \forall \mathbf{x} \in \mathbb{R}^p.$$

Applying this inequality to $\mathbf{x} = \boldsymbol{\vartheta}^{(k,h)}$ and combining it with equation (32), whenever $h < 2/M$ we obtain

$$\mathbf{E}[f^{(k+1)}] \leqslant \mathbf{E}[f^{(k)}] - mh(2-Mh)\,\mathbf{E}[f^{(k)} - f^*] + Mhp. \tag{33}$$

Set $\gamma = mh(2-Mh) \in (0,1)$ for any $h \in (0, 2/M)$. Subtracting f^* from both sides of inequality (33) we arrive at

$$\mathbf{E}[f^{(k+1)} - f^*] \leqslant (1-\gamma)\mathbf{E}[f^{(k)} - f^*] + Mhp. \tag{34}$$

This implies that

$$\begin{aligned}\mathbf{E}[f^{(k+1)} - f^*] &\leqslant (1-\gamma)^{k+1}\,\mathbf{E}[f(\boldsymbol{\vartheta}^{(0)}) - f^*] + Mhp\{1 + \ldots + (1-\gamma)^k\} \\ &\leqslant (1-\gamma)^{k+1}\,\mathbf{E}[f(\boldsymbol{\vartheta}^{(0)}) - f^*] + Mhp\gamma^{-1}. \end{aligned} \tag{35}$$

Inequality (30) follows by replacing γ by $mh(2-Mh)$. To prove inequality (31), it suffices to combine inequality (30) with the first inequality in expression (1), lemma 4 and the inequality $(1-mh)^k \leqslant \exp(-mhk)$. □

Corollary 4. Let $h \leqslant 1/(\alpha M)$ with $\alpha \geqslant 1$ and $K \geqslant 1$ be an integer. Under the conditions of proposition 1, it holds that

$$h\sum_{k=0}^{K-1}\mathbf{E}[\|\nabla f(\boldsymbol{\vartheta}^{(k,h)})\|_2^2] \leqslant \frac{M\alpha}{2\alpha-1}\mathbf{E}[\|\boldsymbol{\vartheta}^{(0)} - \boldsymbol{\theta}^*\|_2^2] + \frac{2\alpha MKhp}{2\alpha-1}.$$

Proof. Using inequality (32) and the fact that $2 - Mh \geqslant (2\alpha-1)/\alpha$, we obtain

$$\frac{h(2\alpha-1)}{2\alpha}\mathbf{E}[\|\nabla f^{(k)}\|_2^2] \leqslant \mathbf{E}[f^{(k)} - f^{(k+1)}] + Mhp, \qquad \forall k \in \mathbb{N}.$$

Summing these inequalities for $k = 0, \ldots, K-1$ and using the obvious bound $f^{(K)} \geqslant f^*$, we obtain

$$h\sum_{k=0}^{K-1}\mathbf{E}[\|\nabla f^{(k)}\|_2^2] \leqslant \frac{2\alpha}{2\alpha-1}\mathbf{E}[f^{(0)} - f^*] + \frac{2\alpha MKhp}{2\alpha-1}.$$

To complete the proof, it suffices to remark that, in view of lemma 4, it holds that $2\mathbf{E}[f^{(0)} - f^*] \leqslant M\mathbf{E}[\|\boldsymbol{\vartheta}^{(0)} - \boldsymbol{\theta}^*\|_2^2]$.

A.2.1. Proof of lemma 1

The first inequality in expression (1) yields $(-\nabla f(\boldsymbol{\theta}) + \nabla f(\bar{\boldsymbol{\theta}}))^{\mathrm{T}}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \leqslant -(m/2)\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_2^2$ for every $\boldsymbol{\theta}, \bar{\boldsymbol{\theta}} \in \mathbb{R}^p$. Therefore, according to Chen and Wang (1997), remark 4.14, and Bakry *et al.* (2014), corollary 4.8.2, the process $\mathbf{L}_t$ is geometrically ergodic in $L^2(\mathbb{R}^p, \pi)$, i.e.

$$\int_{\mathbb{R}^p}\{\mathbf{E}[\varphi(\mathbf{L}_t)|\mathbf{L}_0 = \mathbf{x}] - \mathbf{E}_\pi[\varphi(\boldsymbol{\vartheta})]\}^2\,\pi(\mathbf{x})\mathrm{d}\mathbf{x} \leqslant \exp(-tm)\mathbf{E}_\pi[\varphi^2(\boldsymbol{\vartheta})] \tag{36}$$

for every $t > 0$ and every $\varphi \in L^2(\mathbb{R}^p; \pi)$. The claim of lemma 1 follows from this inequality by simple application of the Cauchy–Schwarz inequality. Indeed, by definition of the total variation and in view of the fact that π is the invariant density of the semigroup $\mathbf{P}_{\mathbf{L}}^t$, we have

$$\begin{aligned}\|\nu\mathbf{P}_{\mathbf{L}}^t - \pi\|_{\mathrm{TV}} &= \sup_{A\in\mathscr{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} \mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A)\nu(\mathbf{x})\mathrm{d}\mathbf{x} - \pi(A)\right| \\ &= \sup_{A\in\mathscr{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} \{\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) - \pi(A)\}\,\nu(\mathbf{x})\mathrm{d}\mathbf{x}\right| \\ &= \sup_{A\in\mathscr{B}(\mathbb{R}^p)} \left| \int_{\mathbb{R}^p} \{\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) - \pi(A)\}\,\{\nu(\mathbf{x}) - \pi(\mathbf{x})\}\,\mathrm{d}\mathbf{x}\right| \\ &\leqslant \sup_{A\in\mathscr{B}(\mathbb{R}^p)} \int_{\mathbb{R}^p} |\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) - \pi(A)|\left|\frac{\nu(\mathbf{x})}{\pi(\mathbf{x})} - 1\right|\pi(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned}$$

Using the Cauchy–Schwarz inequality, we obtain

$$\|\nu\mathbf{P}_{\mathbf{L}}^t - \pi\|_{\mathrm{TV}} \leqslant \sup_{A\in\mathscr{B}(\mathbb{R}^p)} \left\{\int_{\mathbb{R}^p} |\mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) - \pi(A)|^2\pi(\mathbf{x})\mathrm{d}\mathbf{x}\right\}^{1/2} \chi^2(\nu\|\pi)^{1/2}.$$

For every fixed Borel set A, if we set $\varphi(\mathbf{x}) = \mathbb{1}_A(\mathbf{x}) - \pi(A)$ and use inequality (36), we obtain that

$$\begin{aligned}\int_{\mathbb{R}^p} \mathbf{P}_{\mathbf{L}}^t(\mathbf{x}, A) - \pi(A)|^2\pi(\mathbf{x})\mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^p} \{\mathbf{E}[\varphi(\mathbf{L}_t)|\mathbf{L}_0 = \mathbf{x}] - \mathbf{E}_\pi[\varphi(\vartheta)]\}^2\pi(\mathbf{x})\mathrm{d}\mathbf{x} \\ &\leqslant \exp(-tm)\mathbf{E}_\pi[\varphi(\vartheta)^2] \\ &= \exp(-tm)\pi(A)\{1 - \pi(A)\} \leqslant \tfrac{1}{4}\exp(-tm).\end{aligned}$$

This completes the proof of lemma 1.

A.2.2. Proof of lemma 2

Setting $T = Kh$ and using expression (11), we obtain

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathbf{L}}^{\mathbf{x},T}\|\mathbb{P}_{\mathbf{D}}^{\mathbf{x},T}) &= \tfrac{1}{4}\int_0^T \mathbf{E}[\|\nabla f(\mathbf{D}_t) + \mathbf{b}_t(\mathbf{D})\|_2^2]\,\mathrm{d}t \\ &= \tfrac{1}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbf{E}[\|\nabla f(\mathbf{D}_t) - \nabla f(\mathbf{D}_{kh})\|_2^2]\,\mathrm{d}t.\end{aligned}$$

Since ∇f is Lipschitz continuous with Lipschitz constant M, we have

$$\mathrm{KL}(\mathbb{P}_{\mathbf{L}}^{\mathbf{x},T}\|\mathbb{P}_{\mathbf{D}}^{\mathbf{x},T}) \leqslant \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \mathbf{E}[\|\mathbf{D}_t - \mathbf{D}_{kh}\|_2^2]\,\mathrm{d}t.$$

In view of expression (10) we obtain

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathbf{L}}^{\mathbf{x},T}\|\mathbb{P}_{\mathbf{D}}^{\mathbf{x},T}) &\leqslant \frac{M^2}{4}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h} \{\mathbf{E}[\|\nabla f(\mathbf{D}_{kh})\|_2^2(t-kh)^2] + 2p(t-kh)\}\,\mathrm{d}t \\ &= \frac{M^2h^3}{12}\sum_{k=0}^{K-1}\mathbf{E}[\|\nabla f(\vartheta^{(k,h)})\|_2^2] + \frac{pKM^2h^2}{4}. \qquad (37)\end{aligned}$$

Applying corollary 4, the desired inequality follows.

A.2.3. Proof of theorem 2

In view of the triangle inequality, we have

$$\|\nu\mathbf{P}_\vartheta^K - \mathbf{P}_\pi\|_{\mathrm{TV}} = \|\nu\mathbf{P}_{\mathbf{D}}^{Kh} - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \|\nu\mathbf{P}_{\mathbf{L}}^T - \mathbf{P}_\pi\|_{\mathrm{TV}} + \|\nu\mathbf{P}_{\mathbf{D}}^T - \nu\mathbf{P}_{\mathbf{L}}^T\|_{\mathrm{TV}}. \qquad (38)$$

The first term on the right-hand side is what we call the first-type error. The source of this error is the finiteness of time, since it would be equal to 0 if we could choose $T = Kh = \infty$. The second term on the right-hand side of inequality (38) is the second-type error, which is caused by the practical impossibility of taking the step size h equal to 0. These two errors can be evaluated as follows.

For the first-type error, apply lemma 1 to obtain $\|\nu\mathbf{P}_{\mathbf{L}}^T - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\chi^2(\nu\|\pi)^{1/2}\exp(-Tm/2)$. Since ν is a Gaussian distribution, the expectation in the above formula is not difficult to evaluate. The corresponding result, which is provided by lemma 5, yields

$$\|\nu\mathbf{P}_{\mathbf{L}}^T - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{M}{m}\right) - \frac{Tm}{2}\right\}. \tag{39}$$

To evaluate the second-type error, we use the Pinsker inequality:

$$\|\nu\mathbf{P}_{\mathbf{D}}^T - \nu\mathbf{P}_{\mathbf{L}}^T\|_{\mathrm{TV}} \leqslant \|\nu\mathbb{P}_{\mathbf{D}}^T - \nu\mathbb{P}_{\mathbf{L}}^T\|_{\mathrm{TV}} \leqslant \{\tfrac{1}{2}\mathbf{KL}(\nu\mathbb{P}_{\mathbf{L}}^T\|\nu\mathbb{P}_{\mathbf{D}}^T)\}^{1/2}. \tag{40}$$

Combining this inequality with expression (13), we obtain the desired result.

A.3. Proofs of results concerning the LMCO algorithm

A.3.1. Proof of theorem 3

Using the same arguments as those of the proof of theorem 2, this leads to the inequality

$$\|\nu\mathbf{P}_{\vartheta}^K - \mathbf{P}_\pi\|_{\mathrm{TV}} \leqslant \frac{1}{2}\exp\left\{\frac{p}{4}\log\left(\frac{2M}{m}\right) - \frac{Tm}{2}\right\} + \left\{\frac{1}{2}\mathbf{KL}(\nu\mathbb{P}_{\mathbf{L}}^T\|\nu\mathbb{P}_{\mathbf{D}^{\mathrm{O}}}^T)\right\}^{1/2}, \tag{41}$$

where $\mathbb{P}_{\mathbf{D}^{\mathrm{O}}}^T$ is the probability distribution induced by the diffusion process $\mathbf{D}^{\mathrm{O}}$ corresponding to the Ozaki discretization (in fact, it is a piecewise Ornstein–Uhlenbeck process). Relation (11) implies that

$$\mathbf{KL}(\nu\mathbb{P}_{\mathbf{L}}^T\|\nu\mathbb{P}_{\mathbf{D}^{\mathrm{O}}}^T) = \tfrac{1}{4}\int_0^T \mathbf{E}[\|\nabla f(\mathbf{D}_t^{\mathrm{O}}) + b_t(\mathbf{D}^{\mathrm{O}})\|_2^2]\,\mathrm{d}t. \tag{42}$$

Since on each interval $[kh, (k+1)h[$ the function $t \mapsto b_t$ is linear, for every $t \in [kh, (k+1)h[$, we obtain $\|\nabla f(\mathbf{D}_t^{\mathrm{O}}) + b_t(\mathbf{D}^{\mathrm{O}})\|_2^2 = \|\nabla f(\mathbf{D}_t^{\mathrm{O}}) - \nabla f(\mathbf{D}_{kh}^{\mathrm{O}}) - \nabla^2 f(\mathbf{D}_{kh}^{\mathrm{O}})(\mathbf{D}_t^{\mathrm{O}} - \mathbf{D}_{kh}^{\mathrm{O}})\|_2^2$. Using the mean value theorem and the Lipschitz continuity of the Hessian of f, we derive from the above relation that

$$\|\nabla f(\mathbf{D}_t^{\mathrm{O}}) + b_t(\mathbf{D}^{\mathrm{O}})\|_2^2 \leqslant \tfrac{1}{4}L_f^2\|\mathbf{D}_t^{\mathrm{O}} - \mathbf{D}_{kh}^{\mathrm{O}}\|_2^4, \tag{43}$$

for every $t \in [kh, (k+1)h[$. Note now that equation (24) provides the conditional distribution of $\mathbf{D}_{(k+1)h}^{\mathrm{O}}$ given $\mathbf{D}_{kh}^{\mathrm{O}}$. An analogous formula holds for the conditional distribution of $\mathbf{D}_t^{\mathrm{O}} - \mathbf{D}_{kh}^{\mathrm{O}}$ given $\mathbf{D}_{kh}^{\mathrm{O}}$, which is multivariate Gaussian with mean $[\mathbf{I}_p - \exp\{-(t-kh)\mathbf{H}_k\}]\mathbf{H}_k^{-1}\nabla f(\mathbf{D}_{kh}^{\mathrm{O}})$ and covariance matrix $\boldsymbol{\Sigma}_k = [\mathbf{I}_p - \exp\{-2(t-hk)\mathbf{H}_k\}]\mathbf{H}_k^{-1}$, where $\mathbf{H}_k = \nabla^2 f(\mathbf{D}_{kh}^{\mathrm{O}})$. Under the convexity condition on f, we have $\|\{\mathbf{I}_p - \exp(-s\mathbf{H}_k)\}\mathbf{H}_k^{-1}\| \leqslant s$ for every $s > 0$. Therefore, conditioning with respect to $\mathbf{D}_{kh}^{\mathrm{O}}$ and using the inequality $(a+b)^4 \leqslant 8(a^4+b^4)$, for every $t \in [kh, (k+1)h[$ we obtain

$$\begin{aligned}
\tfrac{1}{4}\mathbf{E}[\|\mathbf{D}_t^{\mathrm{O}} - \mathbf{D}_{kh}^{\mathrm{O}}\|_2^4|\mathbf{D}_{kh}^{\mathrm{O}}] &\leqslant \|[\mathbf{I}_p - \exp\{-(t-kh)\mathbf{H}_k\}]\mathbf{H}_k^{-1}\nabla f(\mathbf{D}_{kh}^{\mathrm{O}})\|_2^4 + \mathbf{E}[\|\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\xi}^{(k+1)}\|_2^4|\mathbf{D}_{kh}^{\mathrm{O}}] \\
&\leqslant (t-hk)^4\|\nabla f(\mathbf{D}_{kh}^{\mathrm{O}})\|_2^4 + (p+1)^2\|[\mathbf{I}_p - \exp\{-2(t-hk)\mathbf{H}_k\}]\mathbf{H}_k^{-1}\|^2 \\
&\leqslant (t-hk)^4\|\nabla f(\mathbf{D}_{kh}^{\mathrm{O}})\|_2^4 + 4(t-hk)^2(p+1)^2.
\end{aligned}$$

This inequality, in conjunction with expressions (42) and (43), yields

$$\begin{aligned}
\mathbf{KL}(\nu\mathbb{P}_{\mathbf{L}}^T\|\nu\mathbb{P}_{\mathbf{D}^{\mathrm{O}}}^T) &\leqslant \frac{L_f^2}{16}\sum_{k=0}^{K-1}\int_{kh}^{(k+1)h}\mathbf{E}[\mathbf{E}[\|\mathbf{D}_t^{\mathrm{O}} - \mathbf{D}_{kh}^{\mathrm{O}}\|_2^4|\mathbf{D}_{kh}^{\mathrm{O}}]]\,\mathrm{d}t \\
&\leqslant \frac{L_f^2h^5}{20}\sum_{k=0}^{K-1}\mathbf{E}[\|\nabla f(\mathbf{D}_{kh}^{\mathrm{O}})\|_2^4] + \frac{1}{3}L_f^2Kh^3(p+1)^2.
\end{aligned} \tag{44}$$

To bound the last expectation, we use the fact that $\mathbf{D}_{kh}^{\mathrm{O}}$ equals $\bar{\boldsymbol{\vartheta}}^{(k,h)}$ in distribution, and the next lemma (the proof of which is provided in the on-line supplementary material).

Lemma 6. If $p \geqslant 2$, $T \geqslant 4/(3M)$ and $h \leqslant 1/(8M)$, then the iterates of the LMCO algorithm satisfy

$$\mathbf{E}\left[\left\{\sum_{k=0}^{K-1}\|\nabla f(\bar{\boldsymbol{\vartheta}}^{(k,h)})\|_2^2\right\}^2\right] \leqslant \frac{32}{3}\left(\frac{TMp}{h}\right)^2.$$

Combining lemma 6 and inequality (44), the Kullback–Leibler divergence is upper bounded as follows: $\mathbf{KL}(\nu\mathbb{P}_{\mathbf{L}}^T\|\nu\mathbb{P}_{\mathbf{D}^{\mathrm{O}}}^T) \leqslant 0.534h^3(L_fTMp)^2 + 0.75T(L_fhp)^2$, which completes the proof.

References

Atchadé, Y., Fort, G., Moulines, E. and Priouret, P. (2011) Adaptive Markov chain Monte Carlo: theory and methods. In *Bayesian Time Series Models*, pp. 32–51. Cambridge: Cambridge University Press.

Bakry, D., Gentil, I. and Ledoux, M. (2014) *Analysis and Geometry of Markov Diffusion Operators*. Cham: Springer.

Belloni A. and Chernozhukov, V. (2009) On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.*, **37**, 2011–2055.

Bou-Rabee, N. and Hairer, M. (2013) Nonasymptotic mixing of the MALA algorithm. *IMA J. Numer. Anal.*, **33**, 80–110.

Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization*. Cambridge: Cambridge University Press.

Brooks, S. P. (1998) MCMC convergence diagnosis via multivariate bounds on log-concave densities. *Ann. Statist.*, **26**, 398–433.

Chen, M.-F. and Wang, F.-Y. (1997) Estimation of spectral gap for elliptic operators. *Trans. Am. Math. Soc.*, **349**, 1239–1267.

Dalalyan, A. S. and Tsybakov, A. B. (2009) Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proc. 22nd Conf. Learning Theory, Montreal, June 18th–21st*, pp. 1–10. (Available from `http://colt2009.cs.mcgill.ca/proceedings.html`.)

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

Douc, R., Moulines, E. and Rosenthal, J. S. (2004) Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.*, **14**, 1643–1665.

Durmus, A. and Moulines, E. (2015) Non-asymptotic convergence analysis for the unadjusted langevin algorithm. *Preprint arXiv:1507.05021*. Telecom ParisTech, Paris.

Frieze, A. and Kannan, R. (1999) Log-Sobolev inequalities and sampling from log-concave distributions. *Ann. Appl. Probab.*, **9**, 14–26.

Frieze, A., Kannan, R. and Polson, N. (1994) Sampling from log-concave distributions. *Ann. Appl. Probab.*, **4**, 812–837.

Girolami, M. and Calderhead, B. (2011) Riemann manifold Langevin and Hamiltonian Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **73**, 123–214.

Hanson, T., Branscum, A. and Johnson, W. (2014) Informative g-priors for logistic regression. *Baysn Anal.*, **9**, 597–611.

Holmes, C. and Held, L. (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Baysn Anal.*, **1**, 145–168.

Jarner, S. F. and Hansen, E. (2000) Geometric ergodicity of Metropolis algorithms. *Stoch. Processes Appl.*, **85**, 341–361.

Lamberton, D. and Pagès, G. (2002) Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, **8**, 367–405.

Lemaire, V. (2005) Estimation numérique de la mesure invariante d'un processus de diffusion. *PhD Thesis*. Université de Marne-la-Vallée, Marne-la-Vallée.

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

Lovász, L. and Vempala, S. (2006b) Fast algorithms for logconcave functions: sampling, rounding, integration and optimization. In *Proc. 47th A. Symp. Foundations of Computer Science Berkeley, Oct. 21st–24th*, pp. 57–68.

Meyn, S. P. and Tweedie, R. L. (1994) Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, **4**, 981–1011.

Nesterov, Yu. (2004) *Introductory Lectures on Convex Optimization*. Boston: Kluwer Academic.

Ozaki, T. (1992) A bridge between nonlinear time series models and nonlinear stochastic dynamical systems: a local linearization approach. *Statist. Sin.*, **2**, 113–135.

Pereyra, M. (2014) Proximal markov chain monte carlo algorithms. *Preprint arXiv:1306.0187*. University of Bristol, Bristol.

Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2012) Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *Ann. Appl. Probab.*, **22**, 2320–2356.

Roberts, G. O. and Rosenthal, J. S. (1998) Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B, **60**, 255–268.

Roberts, G. O. and Rosenthal, J. S. (2004) General state space markov chains and mcmc algorithms. *Probab. Surv.*, **1**, 20–71.

Roberts, G. O. and Stramer, O. (2002) Langevin diffusions and Metropolis-Hastings algorithms. *Methodol. Comput. Appl. Probab.*, **4**, 337–357.

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

Rosenthal, J. S. (2002) Quantitative convergence rates of Markov chains: a simple account. *Electron. Communs Probab.*, **7**, 123–128.

Roy, V. (2012) Convergence rates for MCMC algorithms for a robust Bayesian binary regression model. *Electron. J. Statist*, **6**, 2463–2485.

Saumard, A. and Wellner, J. A. (2014) Log-concavity and strong log-concavity: a review. *Statist. Surv.*, **8**, 45–114.
Schreck, A., Fort, G., Le Corff, S. and Moulines, E. (2013) A shrinkage-thresholding metropolis adjusted langevin algorithm for bayesian variable selection. *Preprint arXiv:1312.5658*. Telecom ParisTech, Paris.
Stramer, O. and Tweedie, R. L. (1999a) Langevin-type models: I, Diffusions with given stationary distributions and their discretizations. *Methodol. Comput. Appl. Probab.*, **1**, 283–306.
Stramer, O. and Tweedie, R. L. (1999b) Langevin-type models: II, Self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.*, **1**, 307–328.
Xifara, T., Sherlock, C., Livingstone, S., Byrne, S. and Girolami, M. (2014) Langevin diffusions and the Metropolis-adjusted Langevin algorithm. *Statist. Probab. Lett.*, **91**, 14–19.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to the manuscript "Theoretical guarantees for approximate sampling from smooth and log-concave densities"'.