arXiv:1811.10072v1 [stat.ML] 25 Nov 2018

The promises and pitfalls of Stochastic Gradient Langevin Dynamics

Nicolas Brosse, Éric Moulines
Centre de Mathématiques Appliquées, UMR 7641,
Ecole Polytechnique, Palaiseau, France.
nicolas.brosse@polytechnique.edu, eric.moulines@polytechnique.edu

Alain Durmus
Ecole Normale Supérieure CMLA,
61 Av. du Président Wilson 94235 Cachan Cedex, France.
alain.durmus@cmla.ens-cachan.fr

Abstract

Stochastic Gradient Langevin Dynamics (SGLD) has emerged as a key MCMC algorithm for Bayesian learning from large scale datasets. While SGLD with decreasing step sizes converges weakly to the posterior distribution, the algorithm is often used with a constant step size in practice and has demonstrated successes in machine learning tasks. The current practice is to set the step size inversely proportional to N where N is the number of training samples. As N becomes large, we show that the SGLD algorithm has an invariant probability measure which significantly departs from the target posterior and behaves like Stochastic Gradient Descent (SGD). This difference is inherently due to the high variance of the stochastic gradients. Several strategies have been suggested to reduce this effect; among them, SGLD Fixed Point (SGLDFP) uses carefully designed control variates to reduce the variance of the stochastic gradients. We show that SGLDFP gives approximate samples from the posterior distribution, with an accuracy comparable to the Langevin Monte Carlo (LMC) algorithm for a computational cost sublinear in the number of data points. We provide a detailed analysis of the Wasserstein distances between LMC, SGLD, SGLDFP and SGD and explicit expressions of the means and covariance matrices of their invariant distributions. Our findings are supported by limited numerical experiments.

1 Introduction

Most MCMC algorithms have not been designed to process huge sample sizes, a typical setting in machine learning. As a result, many classical MCMC methods fail in this context, because the mixing time becomes prohibitively long and the cost per iteration increases proportionally to the number of training samples N. The computational cost in standard Metropolis-Hastings algorithm comes from 1) the computation of the proposals, 2) the acceptance/rejection step. Several approaches to solve these issues have been recently proposed in machine learning and computational statistics.

Among them, the stochastic gradient langevin dynamics (SGLD) algorithm, introduced in [37], is a popular choice. This method is based on the Langevin Monte Carlo (LMC) algorithm proposed in [19, 20]. Standard versions of LMC require to compute the gradient of the log-posterior at the current fit of the parameter, but avoid the accept/reject step. The LMC algorithm is a discretization of a continuous-time process, the overdamped Langevin diffusion, which leaves invariant the target distribution π. To further reduce the computational cost, SGLD uses unbiased estimators of the

Preprint. Work in progress.

gradient of the log-posterior based on subsampling. This method has triggered a huge number of works among others [1, 24, 2, 7, 9, 14, 27, 15, 4] and have been successfully applied to a range of state of the art machine learning problems [30, 26].

The properties of SGLD with decreasing step sizes have been studied in [34]. The two key findings in this work are that 1) the SGLD algorithm converges weakly to the target distribution π, 2) the optimal rate of convergence to equilibrium scales as $n^{-1/3}$ where n is the number of iterations, see [34, Section 5]. However, in most of the applications, constant rather than decreasing step sizes are used, see [1, 9, 21, 25, 33, 36]. A natural question for the practical design of SGLD is the choice of the minibatch size. This size controls on the one hand the computational complexity of the algorithm per iteration and on the other hand the variance of the gradient estimator. Non-asymptotic bounds in Wasserstein distance between the marginal distribution of the SGLD iterates and the target distribution π have been established in [11, 12]. These results highlight the cost of using stochastic gradients and show that, for a given precision ϵ in Wasserstein distance, the computational cost of the plain SGLD algorithm does not improve over the LMC algorithm; Nagapetyan et al. [28] reports also similar results on the mean square error.

It has been suggested to use control variates to reduce the high variance of the stochastic gradients. For strongly log-concave models, Nagapetyan et al. [28], Baker et al. [3] use the mode of the posterior distribution as a reference point and introduce the SGLDFP (Stochastic Gradient Langevin Dynamics Fixed Point) algorithm. Nagapetyan et al. [28], Baker et al. [3] provide upper bounds on the mean square error and the Wasserstein distance between the marginal distribution of the iterates of SGLDFP and the posterior distribution. In addition, Nagapetyan et al. [28], Baker et al. [3] show that the overall cost remains sublinear in the number of individual data points, up to a preprocessing step. Other control variates methodologies are provided for non-concave models in the form of SAGA-Langevin Dynamics and SVRG-Langevin Dynamics [15, 8], albeit a detailed analysis in Wasserstein distance of these algorithms is only available for strongly log-concave models [6].

In this paper, we provide further insights on the links between SGLD, SGLDFP, LMC and SGD (Stochastic Gradient Descent). In our analysis, the algorithms are used with a constant step size and the parameters are set to the standard values used in practice [1, 9, 21, 25, 33, 36]. The LMC, SGLD and SGLDFP algorithms define homogeneous Markov chains, each of which admits a unique stationary distribution used as a hopefully close proxy of π. The main contribution of this paper is to show that, while the invariant distributions of LMC and SGLDFP become closer to π as the number of data points increases, on the opposite, the invariant measure of SGLD never comes close to the target distribution π and is in fact very similar to the invariant measure of SGD.

In Section 3.1, we give an upper bound in Wasserstein distance of order 2 between the marginal distribution of the iterates of LMC and the Langevin diffusion, SGLDFP and LMC, and SGLD and SGD. We provide a lower bound on the Wasserstein distance between the marginal distribution of the iterates of SGLDFP and SGLD. In Section 3.2, we give a comparison of the means and covariance matrices of the invariant distributions of LMC, SGLDFP and SGLD with those of the target distribution π. Our claims are supported by numerical experiments in Section 4.

2 Preliminaries

Denote by $\mathbf{z} = \{z_i\}_{i=1}^N$ the observations. We are interested in situations where the target distribution π arises as the posterior in a Bayesian inference problem with prior density $\pi_0(\theta)$ and a large number $N \gg 1$ of i.i.d. observations z_i with likelihoods $p(z_i|\theta)$. In this case, $\pi(\theta) = \pi_0(\theta) \prod_{i=1}^N p(z_i|\theta)$. We denote $U_i(\theta) = -\log(p(z_i|\theta))$ for $i \in \{1, \ldots, N\}$, $U_0(\theta) = -\log(\pi_0(\theta))$, $U = \sum_{i=0}^N U_i$.

Under mild conditions, π is the unique invariant probability measure of the Langevin Stochastic Differential Equation (SDE):

$$\mathrm{d}\theta_t = -\nabla U(\theta_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \ , \tag{1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Based on this observation, Langevin Monte Carlo (LMC) is an MCMC algorithm that enables to sample (approximately) from π using an Euler discretization of the Langevin SDE:

$$\theta_{k+1} = \theta_k - \gamma \nabla U(\theta_k) + \sqrt{2\gamma} Z_{k+1} \ , \tag{2}$$

where $\gamma > 0$ is a constant step size and $(Z_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian vectors. Discovered and popularised in the seminal works [19, 20, 32], LMC has recently received renewed attention [10, 17, 16, 12]. However, the cost of one iteration is Nd which is prohibitively large for massive datasets. In order to scale up to the big data setting, Welling and Teh [37] suggested to replace ∇U with an unbiased estimate $\nabla U_0 + (N/p)\sum_{i\in S}\nabla U_i$ where S is a minibatch of $\{1,\ldots,N\}$ with replacement of size p. A single update of SGLD is then given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right) + \sqrt{2\gamma}Z_{k+1}\ . \tag{3}$$

The idea of using only a fraction of data points to compute an unbiased estimate of the gradient at each iteration comes from Stochastic Gradient Descent (SGD) which is a popular algorithm to minimize the potential U. SGD is very similar to SGLD because it is characterised by the same recursion as SGLD but without Gaussian noise:

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}}\nabla U_i(\theta_k)\right)\ . \tag{4}$$

Assuming for simplicity that U has a minimizer $\theta^\star$, we can define a control variates version of SGLD, SGLDFP, see [15, 8], given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - \gamma\left(\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S_{k+1}}\{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\}\right) + \sqrt{2\gamma}Z_{k+1}\ . \tag{5}$$

It is worth mentioning that the objectives of the different algorithms presented so far are distinct. On the one hand, LMC, SGLD and SGDLFP are MCMC methods used to obtain approximate samples from the posterior distribution π. On the other hand, SGD is a stochastic optimization algorithm used to find an estimate of the mode $\theta^\star$ of the posterior distribution. In this paper, we focus on the fixed step-size SGLD algorithm and assess its ability to reliably sample from π. For that purpose and to quantify precisely the relation between LMC, SGLD, SGDFP and SGD, we make for simplicity the following assumptions on U.

H1. *For all $i \in \{0,\ldots,N\}$, U_i is four times continuously differentiable and for all $j \in \{2,3,4\}$, $\sup_{\theta\in\mathbb{R}^d}\left\|\mathrm{D}^j U_i(\theta)\right\| \leq \tilde{L}$. In particular for all $i \in \{0,\ldots,N\}$, U_i is $\tilde{L}$-gradient Lipschitz, i.e. for all $\theta_1,\theta_2 \in \mathbb{R}^d$, $\|\nabla U_i(\theta_1) - \nabla U_i(\theta_2)\| \leq \tilde{L}\,\|\theta_1 - \theta_2\|$.*

H2. *U is m-strongly convex, i.e. for all $\theta_1,\theta_2 \in \mathbb{R}^d$, $\langle \nabla U(\theta_1) - \nabla U(\theta_2), \theta_1 - \theta_2\rangle \geq m\,\|\theta_1 - \theta_2\|^2$.*

H3. *For all $i \in \{0,\ldots,N\}$, U_i is convex.*

Note that under **H** 1, U is four times continuously differentiable and for $j \in \{2,3,4\}$, $\sup_{\theta\in\mathbb{R}^d}\left\|\mathrm{D}^j U(\theta)\right\| \leq L$, with $L = (N+1)\tilde{L}$ and where $\left\|\mathrm{D}^j U(\theta)\right\| = \sup_{\|u_1\|\leq 1,\ldots,\|u_j\|\leq 1}\mathrm{D}^j U(\theta)[u_1,\ldots,u_j]$. In particular, U is L-gradient Lipschitz. Furthermore, under **H**2, U has a unique minimizer $\theta^\star$. In this paper, we focus on the asymptotic $N \to +\infty$,. We assume that $\liminf_{N\to+\infty} N^{-1}m > 0$, which is a common assumption for the analysis of SGLD and SGLDFP [3, 6]. In practice [1, 9, 21, 25, 33, 36], γ is of order $1/N$ and we adopt this convention in this article.

For a practical implementation of SGLDFP, an estimator $\hat{\theta}$ of $\theta^\star$ is necessary. The theoretical analysis and the bounds remain unchanged if, instead of considering SGLDFP centered w.r.t. $\theta^\star$, we study SGLDFP centered w.r.t. $\hat{\theta}$ satisfying $\mathbb{E}[\|\hat{\theta} - \theta^\star\|^2] = O(1/N)$. Such an estimator $\hat{\theta}$ can be computed using for example SGD with decreasing step sizes, see [29, eq.(2.8)] and [3, Section 3.4], for a computational cost linear in N.

3 Results

3.1 Analysis in Wasserstein distance

Before presenting the results, some notations and elements of Markov chain theory have to be introduced. Denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite second moment and by

$\mathcal{B}(\mathbb{R}^d)$ the Borel σ-algebra of $\mathbb{R}^d$. For $\lambda, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, define the Wasserstein distance of order 2 by

$$\mathrm{W}_2(\lambda, \nu) = \inf_{\xi \in \Pi(\lambda,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \vartheta\|^2 \, \xi(\mathrm{d}\theta, \mathrm{d}\vartheta) \right)^{1/2} ,$$

where $\Pi(\lambda, \nu)$ is the set of probability measures ξ on $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^d)$ satisfying for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\xi(\mathsf{A} \times \mathbb{R}^d)) = \lambda(\mathsf{A})$ and $\xi(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$.

A Markov kernel R on $\mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ is a mapping $R : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, 1]$ satisfying the following conditions: (i) for every $\theta \in \mathbb{R}^d$, $R(\theta, \cdot) : \mathsf{A} \mapsto R(\theta, \mathsf{A})$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ (ii) for every $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $R(\cdot, A) : \theta \mapsto R(\theta, A)$ is a measurable function. For any probability measure λ on $\mathcal{B}(\mathbb{R}^d)$, we define λR for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $\lambda R(\mathsf{A}) = \int_{\mathbb{R}^d} \lambda(\mathrm{d}\theta) R(\theta, \mathsf{A})$. For all $k \in \mathbb{N}^*$, we define the Markov kernel R^k recursively by $R^1 = R$ and for all $\theta \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $R^{k+1}(\theta, \mathsf{A}) = \int_{\mathbb{R}^d} R^k(\theta, \mathrm{d}\vartheta) R(\vartheta, \mathsf{A})$. A probability measure $\bar{\pi}$ is invariant for R if $\bar{\pi} R = \bar{\pi}$.

The LMC, SGLD, SGD and SGLDFP algorithms defined respectively by (2), (3), (4) and (5) are homogeneous Markov chains with Markov kernels denoted $R_{\mathrm{LMC}}, R_{\mathrm{SGLD}}, R_{\mathrm{SGD}}$, and R_{FP}. To avoid overloading the notations, the dependence on γ and N is implicit.

Lemma 1. *Assume **H**1, **H**2 and **H**3. For any step size $\gamma \in (0, 2/L)$, R_{SGLD} (respectively $R_{\mathrm{LMC}}, R_{\mathrm{SGD}}, R_{\mathrm{FP}}$) has a unique invariant measure $\pi_{\mathrm{SGLD}} \in \mathcal{P}_2(\mathbb{R}^d)$ (respectively $\pi_{\mathrm{LMC}}, \pi_{\mathrm{SGD}}, \pi_{\mathrm{FP}}$). In addition, for all $\gamma \in (0, 1/L]$, $\theta \in \mathbb{R}^d$ and $k \in \mathbb{N}$,*

$$\mathrm{W}_2^2(R^k_{\mathrm{SGLD}}(\theta, \cdot), \pi_{\mathrm{SGLD}}) \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\theta - \vartheta\|^2 \, \pi_{\mathrm{SGLD}}(\mathrm{d}\vartheta)$$

and the same inequality holds for LMC, SGD and SGLDFP.

Proof. The proof is postponed to Appendix A.1. □

Under **H**1, (1) has a unique strong solution $(\theta_t)_{t \geq 0}$ for every initial condition $\theta_0 \in \mathbb{R}^d$ [23, Chapter 5, Theorems 2.5 and 2.9]. Denote by $(P_t)_{t \geq 0}$ the semigroup of the Langevin diffusion defined for all $\theta_0 \in \mathbb{R}^d$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by $P_t(\theta_0, \mathsf{A}) = \mathbb{P}(\theta_t \in \mathsf{A})$.

Theorem 2. *Assume **H**1, **H**2 and **H**3. For all $\gamma \in (0, 1/L]$, $\lambda, \mu \in \mathcal{P}_2(\mathbb{R}^d)$ and $n \in \mathbb{N}$, we have the following upper-bounds in Wasserstein distance between*

i) LMC and SGLDFP,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{LMC}}, \mu R^n_{\mathrm{FP}}) \leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda, \mu) + \frac{2L^2 \gamma d}{pm^2} + \frac{L^2 \gamma^2}{p} n (1 - m\gamma)^{n-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \, \mu(\mathrm{d}\vartheta) ,$$

ii) the Langevin diffusion and LMC,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{LMC}}, \mu P_{n\gamma}) \leq 2 \left(1 - \frac{mL\gamma}{m + L}\right)^n \mathrm{W}_2^2(\lambda, \mu) + d\gamma \frac{m + L}{2m} \left(3 + \frac{L}{m}\right) \left(\frac{13}{6} + \frac{L}{m}\right) + n \mathrm{e}^{-(m/2)\gamma(n-1)} L^3 \gamma^3 \left(1 + \frac{m + L}{2m}\right) \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \, \mu(\mathrm{d}\vartheta) ,$$

iii) SGLD and SGD,

$$\mathrm{W}_2^2(\lambda R^n_{\mathrm{SGLD}}, \mu R^n_{\mathrm{SGD}}) \leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda, \mu) + (2d)/m .$$

Proof. The proof is postponed to Appendix A.2. □

Corollary 3. *Assume **H**1, **H**2 and **H**3. Set $\gamma = \eta/N$ with $\eta \in (0, 1/(2\tilde{L})]$ and assume that $\liminf_{N \to \infty} mN^{-1} > 0$. Then,*

i) for all $n \in \mathbb{N}$, *we get* $\mathrm{W}_2(R^n_{\mathrm{LMC}}(\theta^\star,\cdot), R^n_{\mathrm{FP}}(\theta^\star,\cdot)) = \sqrt{d\eta}\, O(N^{-1/2})$ *and* $\mathrm{W}_2(\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}) = \sqrt{d\eta}\, O(N^{-1/2})$, $\mathrm{W}_2(\pi_{\mathrm{LMC}}, \pi) = \sqrt{d\eta}\, O(N^{-1/2})$.

ii) for all $n \in \mathbb{N}$, *we get* $\mathrm{W}_2(R^n_{\mathrm{SGLD}}(\theta^\star,\cdot), R^n_{\mathrm{SGD}}(\theta^\star,\cdot)) = \sqrt{d}\, O(N^{-1/2})$ *and* $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGD}}) = \sqrt{d}\, O(N^{-1/2})$.

Theorem 2 implies that the number of iterations necessary to obtain a sample ε-close from π in Wasserstein distance is the same for LMC and SGLDFP. However for LMC, the cost of one iteration is Nd which is larger than pd the cost of one iteration for SGLDFP. In other words, to obtain an approximate sample from the target distribution at an accuracy $O(1/\sqrt{N})$ in 2-Wasserstein distance, LMC requires $\Theta(N)$ operations, in contrast with SGLDFP that needs only $\Theta(1)$ operations.

We show in the sequel that $\mathrm{W}_2(\pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}) = \Omega(1)$ when $N \to +\infty$ in the case of a Bayesian linear regression, where for two sequences $(u_N)_{N\geq 1}$, $(v_N)_{N\geq 1}$, $u_N = \Omega(v_N)$ if $\liminf_{N\to+\infty} u_N/v_N > 0$. The dataset is $\mathbf{z} = \{(y_i, x_i)\}_{i=1}^N$ where $y_i \in \mathbb{R}$ is the response variable and $x_i \in \mathbb{R}^d$ are the covariates. Set $\mathbf{y} = (y_1, \ldots, y_N) \in \mathbb{R}^N$ and $\mathbf{X} \in \mathbb{R}^{N\times d}$ the matrix of covariates such that the ith row of $\mathbf{X}$ is x_i. Let $\sigma_y^2, \sigma_\theta^2 > 0$. For $i \in \{1, \ldots, N\}$, the conditional distribution of y_i given x_i is Gaussian with mean $x_i^{\mathrm{T}}\theta$ and variance σ_y^2. The prior $\pi_0(\theta)$ is a normal distribution of mean 0 and variance $\sigma_\theta^2 \,\mathrm{Id}$. The posterior distribution π is then proportional to $\pi(\theta) \propto \exp\left(-(1/2)(\theta-\theta^\star)^{\mathrm{T}}\Sigma(\theta-\theta^\star)\right)$ where

$$\Sigma = \mathrm{Id}\,/\sigma_\theta^2 + \mathbf{X}^{\mathrm{T}}\mathbf{X}/\sigma_y^2 \quad \text{and} \quad \theta^\star = \Sigma^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y})/\sigma_y^2 \ .$$

We assume that $\mathbf{X}^{\mathrm{T}}\mathbf{X} \succeq m\,\mathrm{Id}$, with $\liminf_{N\to+\infty} m/N > 0$. Let S be a minibatch of $\{1, \ldots, N\}$ with replacement of size p. Define

$$\nabla U_0(\theta) + (N/p)\sum_{i\in S}\nabla U_i(\theta) = \Sigma(\theta-\theta^\star) + \rho(S)(\theta-\theta^\star) + \xi(S)$$

where

$$\rho(S) = \frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S} x_i x_i^{\mathrm{T}} - \Sigma \ , \ \xi(S) = \frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S}\left(x_i^{\mathrm{T}}\theta^\star - y_i\right)x_i \ . \tag{6}$$

$\rho(S)(\theta-\theta^\star)$ is the multiplicative part of the noise in the stochastic gradient, and $\xi(S)$ the additive part that does not depend on θ. The additive part of the stochastic gradient for SGLDFP disappears since

$$\nabla U_0(\theta) - \nabla U_0(\theta^\star) + (N/p)\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\} = \Sigma(\theta-\theta^\star) + \rho(S)(\theta-\theta^\star) \ .$$

In this setting, the following theorem shows that the Wasserstein distances between the marginal distribution of the iterates of SGLD and SGLDFP, and π_{SGLD} and π, is of order $\Omega(1)$ when $N \to +\infty$. This is in sharp contrast with the results of Corollary 3 where the Wasserstein distances tend to 0 as $N \to +\infty$ at a rate $N^{-1/2}$. For simplicity, we state the result for $d = 1$.

Theorem 4. *Consider the case of the Bayesian linear regression in dimension* 1.

i) For all $\gamma \in (0, \Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}]$ *and* $n \in \mathbb{N}^*$,

$$\left(\frac{1-\mu}{1-\mu^n}\right)^{1/2}\mathrm{W}_2(R^n_{\mathrm{SGLD}}(\theta^\star,\cdot), R^n_{\mathrm{FP}}(\theta^\star,\cdot)) \geq \left\{2\gamma + \frac{\gamma^2 N}{p}\sum_{i=1}^N\left(\frac{(x_i\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2}\right)^2\right\}^{1/2} - \sqrt{2\gamma} \ ,$$

where $\mu \in (0, 1-\gamma\Sigma]$.

ii) Set $\gamma = \eta/N$ *with* $\eta \in (0, \liminf_{N\to+\infty} N\Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}]$ *and assume that* $\liminf_{N\to+\infty} N^{-1}\sum_{i=1}^N x_i^2 > 0$. *We have* $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi) = \Omega(1)$.

Proof. The proof is postponed to Appendix A.3. □

The study in Wasserstein distance emphasizes the different behaviors of the LMC, SGLDFP, SGLD and SGD algorithms. When $N \to \infty$ and $\lim_{N\to+\infty} m/N > 0$, the marginal distributions of the kth iterates of the LMC and SGLDFP algorithm are very close to the Langevin diffusion and their invariant probability measures π_{LMC} and π_{FP} are similar to the posterior distribution of interest π. In contrast, the marginal distributions of the kth iterates of SGLD and SGD are analogous and their invariant probability measures π_{SGLD} and π_{SGD} are very different from π when $N \to +\infty$.

Note that to fix the asymptotic bias of SGLD, other strategies can be considered: choosing a step size $\gamma \propto N^{-\beta}$ where $\beta > 1$ and/or increasing the batch size $p \propto N^{\alpha}$ where $\alpha \in [0, 1]$. Using the Wasserstein (of order 2) bounds of SGLD w.r.t. the target distribution π, see e.g. [12, Theorem 3], $\alpha + \beta$ should be equal to 2 to guarantee the ε-accuracy in Wasserstein distance of SGLD for a cost proportional to N (up to logarithmic terms), independently of the choice of α and β.

3.2 Mean and covariance matrix of $\pi_{\text{LMC}}, \pi_{\text{FP}}, \pi_{\text{SGLD}}$

We now establish an expansion of the mean and second moments of $\pi_{\text{LMC}}, \pi_{\text{FP}}, \pi_{\text{SGLD}}$ and π_{SGD} as $N \to +\infty$, and compare them. We first give an expansion of the mean and second moments of π as $N \to +\infty$.

Proposition 5. *Assume **H1** and **H2** and that* $\liminf_{N\to+\infty} N^{-1}m > 0$. *Then,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi(\mathrm{d}\theta) = \nabla^2 U(\theta^\star)^{-1} + O_{N\to+\infty}(N^{-3/2}) \ ,$$

$$\int_{\mathbb{R}^d} \theta \, \pi(\mathrm{d}\theta) - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1} \, \mathrm{D}^3 \, U(\theta^\star)[\nabla^2 U(\theta^\star)^{-1}] + O_{N\to+\infty}(N^{-3/2}) \ .$$

Proof. The proof is postponed to Appendix B.1. □

Contrary to the Bayesian linear regression where the covariance matrices can be explicitly computed, see Appendix C, only approximate expressions are available in the general case. For that purpose, we consider two types of asymptotic. For LMC and SGLDFP, we assume that $\lim_{N\to+\infty} m/N > 0$, $\gamma = \eta/N$, for $\eta > 0$, and we develop an asymptotic when $N \to +\infty$. Combining Proposition 5 and Theorem 6 , we show that the biases and covariance matrices of π_{LMC} and π_{FP} are of order $\Theta(1/N)$ with remainder terms of the form $O(N^{-3/2})$, where for two sequences $(u_N)_{N\geq 1}$, $(v_N)_{N\geq 1}$, $u = \Theta(v)$ if $0 < \liminf_{N\to+\infty} u_N/v_N \leq \limsup_{N\to+\infty} u_N/v_N < +\infty$.

Regarding SGD and SGLD, we do not have such concentration properties when $N \to +\infty$ because of the high variance of the stochastic gradients. The biases and covariance matrices of SGLD and SGD are of order $\Theta(1)$ when $N \to +\infty$. To obtain approximate expressions of these quantities, we set $\gamma = \eta/N$ where $\eta > 0$ is the step size for the gradient descent over the normalized potential U/N. Assuming that m is proportional to N and $N \geq 1/\eta$, we show by combining Proposition 5 and Theorem 7 that the biases and covariance matrices of SGLD and SGD are of order $\Theta(\eta)$ with remainder terms of the form $O(\eta^{3/2})$ when $\eta \to 0$.

Before giving the results associated to $\pi_{\text{LMC}}, \pi_{\text{FP}}, \pi_{\text{SGLD}}$ and π_{SGD}, we need to introduce some notations. For any matrices $A_1, A_2 \in \mathbb{R}^{d\times d}$, we denote by $A_1 \otimes A_2$ the Kronecker product defined on $\mathbb{R}^{d\times d}$ by $A_1 \otimes A_2 : Q \mapsto A_1 Q A_2$ and $A^{\otimes 2} = A \otimes A$. Besides, for all $\theta_1 \in \mathbb{R}^d$ and $\theta_2 \in \mathbb{R}^d$, we denote by $\theta_1 \otimes \theta_2 \in \mathbb{R}^{d\times d}$ the tensor product of θ_1 and θ_2. For any matrix $A \in \mathbb{R}^{d\times d}$, $\mathrm{Tr}(A)$ is the trace of A.

Define $\mathrm{K} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ for all $A \in \mathbb{R}^{d\times d}$ by

$$\mathrm{K}(A) = \frac{N}{p} \sum_{i=1}^{N} \left(\nabla^2 U_i(\theta^\star) - \frac{1}{N} \sum_{j=1}^{N} \nabla^2 U_j(\theta^\star) \right)^{\otimes 2} A \ . \tag{7}$$

and H and $\mathrm{G} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ by

$$\mathrm{H} = \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma \nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) \ , \tag{8}$$

$$\mathrm{G} = \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma (\nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) + \mathrm{K}) \ . \tag{9}$$

K, H and G can be interpreted as perturbations of $\nabla^2 U(\theta^\star)^{\otimes 2}$ and $\nabla^2 U(\theta^\star)$, respectively, due to the noise of the stochastic gradients. It can be shown, see Appendix B.2, that for γ small enough, H and G are invertible.

Theorem 6. *Assume **H1**, **H2** and **H3**. Set $\gamma = \eta/N$ and assume that $\liminf_{N\to+\infty} N^{-1}m > 0$. There exists an (explicit) η_0 independent of N such that for all $\eta \in (0, \eta_0)$,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\text{LMC}}(\mathrm{d}\theta) = \mathrm{H}^{-1}(2\,\mathrm{Id}) + O_{N\to+\infty}(N^{-3/2}) \ , \tag{10}$$

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\text{FP}}(\mathrm{d}\theta) = \mathrm{G}^{-1}(2\,\mathrm{Id}) + O_{N\to+\infty}(N^{-3/2}) \ , \tag{11}$$

and

$$\int_{\mathbb{R}^d} \theta \pi_{\text{LMC}}(\mathrm{d}\theta) - \theta^\star = -\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3\, U(\theta^\star)[\mathrm{H}^{-1}\,\mathrm{Id}] + O_{N\to+\infty}(N^{-3/2}) \ ,$$

$$\int_{\mathbb{R}^d} \theta \pi_{\text{FP}}(\mathrm{d}\theta) - \theta^\star = -\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3\, U(\theta^\star)[\mathrm{G}^{-1}\,\mathrm{Id}] + O_{N\to+\infty}(N^{-3/2}) \ .$$

Proof. The proof is postponed to Appendix B.2.2. □

Theorem 7. *Assume **H1**, **H2** and **H3**. Set $\gamma = \eta/N$ and assume that $\liminf_{N\to+\infty} N^{-1}m > 0$. There exists an (explicit) η_0 independent of N such that for all $\eta \in (0, \eta_0)$ and $N \geq 1/\eta$,*

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\text{SGLD}}(\mathrm{d}\theta) = \mathrm{G}^{-1}\left\{2\,\mathrm{Id} + (\eta/p)\,\mathrm{M}\right\} + O_{\eta\to 0}(\eta^{3/2}) \ , \tag{12}$$

$$\int_{\mathbb{R}^d} (\theta - \theta^\star)^{\otimes 2} \pi_{\text{SGD}}(\mathrm{d}\theta) = (\eta/p)\,\mathrm{G}^{-1}\,\mathrm{M} + O_{\eta\to 0}(\eta^{3/2}) \ , \tag{13}$$

and

$$\int_{\mathbb{R}^d} \theta \pi_{\text{SGLD}}(\mathrm{d}\theta) - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3\, U(\theta^\star)[\mathrm{G}^{-1}\left\{2\,\mathrm{Id} + (\eta/p)\,\mathrm{M}\right\}] + O_{\eta\to 0}(\eta^{3/2}) \ ,$$

$$\int_{\mathbb{R}^d} \theta \pi_{\text{SGD}}(\mathrm{d}\theta) - \theta^\star = -(\eta/2p)\nabla^2 U(\theta^\star)^{-1}\, \mathrm{D}^3\, U(\theta^\star)[\mathrm{G}^{-1}\,\mathrm{M}] + O_{\eta\to 0}(\eta^{3/2}) \ ,$$

where

$$\mathrm{M} = \sum_{i=1}^{N} \left(\nabla U_i(\theta^\star) - \frac{1}{N} \sum_{j=1}^{N} \nabla U_j(\theta^\star) \right)^{\otimes 2} \ , \tag{14}$$

and G *is defined in* (9).

Proof. The proof is postponed to Appendix B.2.2. □

Note that this result implies that the mean and the covariance matrix of π_{SGLD} and π_{SGD} stay lower bounded by a positive constant for any $\eta > 0$ as $N \to +\infty$. In Appendix D, a figure illustrates the results of Theorem 6 and Theorem 7 in the asymptotic $N \to +\infty$.

4 Numerical experiments

Simulated data For illustrative purposes, we consider a Bayesian logistic regression in dimension $d = 2$. We simulate $N = 10^5$ covariates $\{x_i\}_{i=1}^N$ drawn from a standard 2-dimensional Gaussian distribution and we denote by $\mathbf{X} \in \mathbb{R}^{N\times d}$ the matrix of covariates such that the ith row of $\mathbf{X}$ is x_i. Our Bayesian regression model is specified by a Gaussian prior of mean 0 and covariance matrix the identity, and a likelihood given for $y_i \in \{0, 1\}$ by $p(y_i|x_i, \theta) = (1 + \mathrm{e}^{-x_i^{\mathrm{T}}\theta})^{-y_i}(1 + \mathrm{e}^{x_i^{\mathrm{T}}\theta})^{y_i - 1}$. We simulate N observations $\{y_i\}_{i=1}^N$ under this model. In this setting, **H**1 and **H**3 are satisfied, and **H**2 holds if the state space is compact.

To illustrate the results of Section 3.2, we consider 10 regularly spaced values of N between 10^2 and 10^5 and we truncate the dataset accordingly. We compute an estimator $\hat{\theta}$ of $\theta^\star$ using SGD [31]

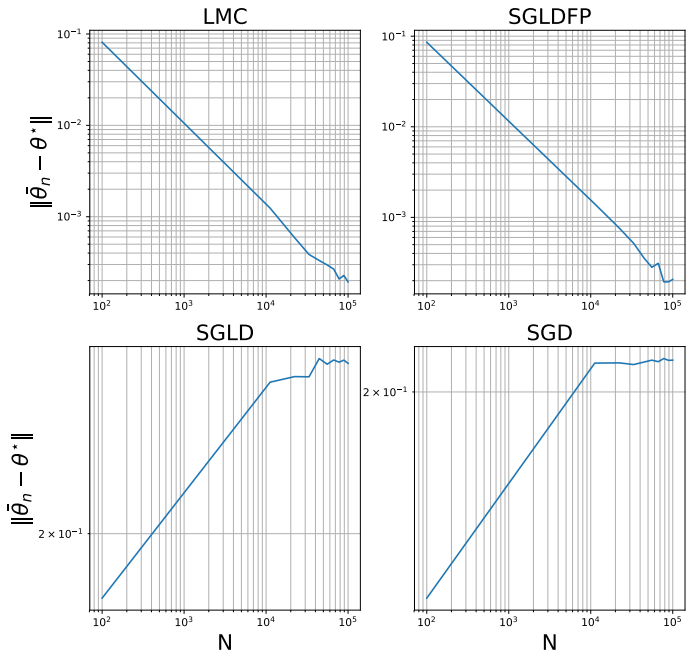

Figure 1: Distance to $\theta^\star$, $\left\|\bar{\theta}_n - \theta^\star\right\|$ for LMC, SGLDFP, SGLD and SGD, function of N, in logarithmic scale.

combined with the BFGS algorithm [22]. For the LMC, SGLDFP, SGLD and SGD algorithms, the step size γ is set equal to $(1+\delta/4)^{-1}$ where δ is the largest eigenvalue of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. We start the algorithms at $\theta_0 = \hat{\theta}$ and run $n = 1/\gamma$ iterations where the first 10% samples are discarded as a burn-in period.

We estimate the means and covariance matrices of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} by their empirical averages $\bar{\theta}_n = (1/n)\sum_{k=0}^{n-1}\theta_k$ and $\{1/(n-1)\}\sum_{k=0}^{n-1}(\theta_k - \bar{\theta}_n)^{\otimes 2}$. We plot the mean and the trace of the covariance matrices for the different algorithms, averaged over 100 independent trajectories, in Figure 1 and Figure 2 in logarithmic scale.

The slope for LMC and SGLDFP is -1 which confirms the convergence of $\left\|\bar{\theta}_n - \theta^\star\right\|$ to 0 at a rate N^{-1}. On the other hand, we can observe that $\left\|\bar{\theta}_n - \theta^\star\right\|$ converges to a constant for SGD and SGLD.

Covertype dataset We then illustrate our results on the covertype dataset[1] with a Bayesian logistic regression model. The prior is a standard multivariate Gaussian distribution. Given the size of the dataset and the dimension of the problem, LMC requires high computational resources and is not included in the simulations. We truncate the training dataset at $N \in \left\{10^3, 10^4, 10^5\right\}$. For all algorithms, the step size γ is set equal to $1/N$ and the trajectories are started at $\hat{\theta}$, an estimator of $\theta^\star$, computed using SGD combined with the BFGS algorithm.

We empirically check that the variance of the stochastic gradients scale as N^2 for SGD and SGLD, and as N for SGLDFP. We compute the empirical variance estimator of the gradients, take the mean over the dimension and display the result in a logarithmic plot in Figure 3. The slopes are 2 for SGD and SGLD, and 1 for SGLDFP.

On the test dataset, we also evaluate the negative loglikelihood of the three algorithms for different values of $N \in \left\{10^3, 10^4, 10^5\right\}$, as a function of the number of iterations. The plots are shown in Figure 4. We note that for large N, SGLD and SGD give very similar results that are below the performance of SGLDFP.

[1] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/covtype.libsvm.binary.scale.bz2`

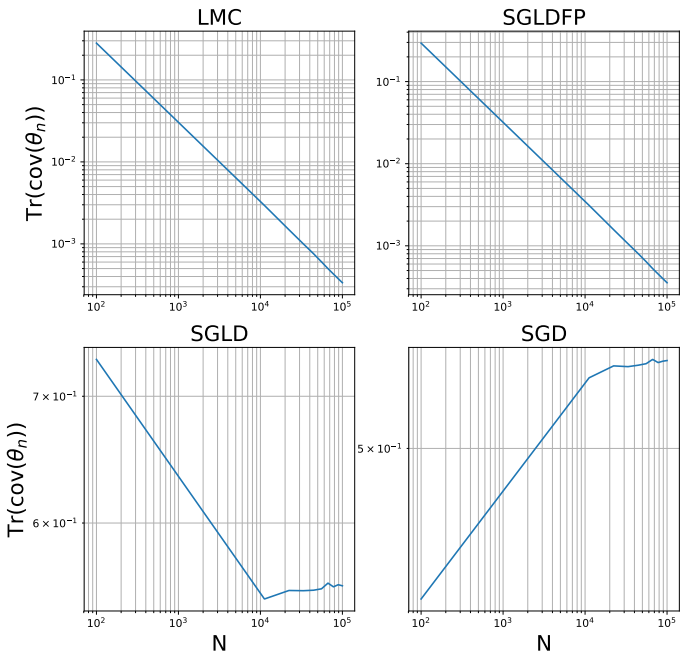

Figure 2: Trace of the covariance matrices for LMC, SGLDFP, SGLD and SGD, function of N, in logarithmic scale.

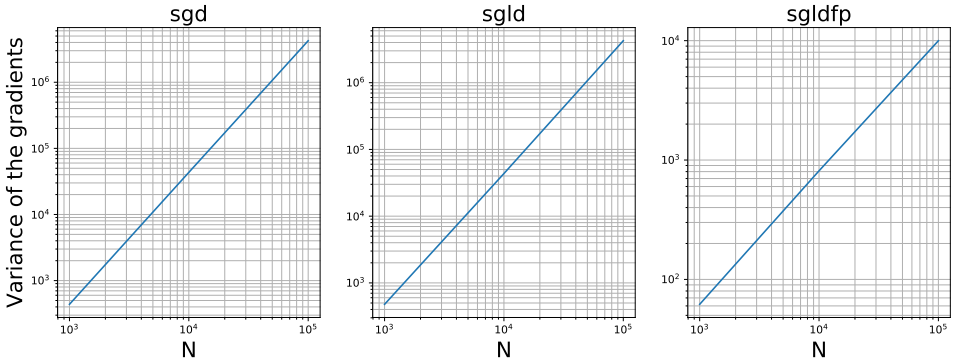

Figure 3: Variance of the stochastic gradients of SGLD, SGLDFP and SGD function of N, in logarithmic scale.

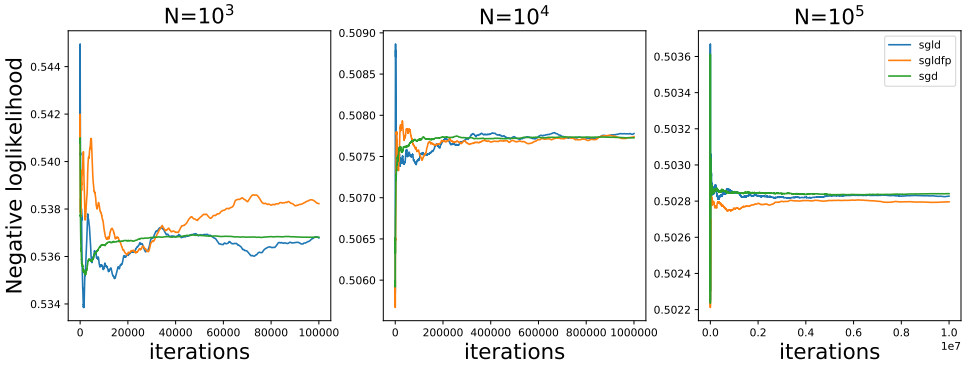

Figure 4: Negative loglikelihood on the test dataset for SGLD, SGLDFP and SGD function of the number of iterations for different values of $N \in \{10^3, 10^4, 10^5\}$.

A Proofs of Section 3.1

A.1 Proof of Lemma 1

The convergence in Wasserstein distance is classically done via a standard synchronous coupling [13, Proposition 2]. We prove the statement for SGLD; the adaptation for LMC, SGLDFP and SGD is immediate. Let $\gamma \in (0, 2/L)$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [35, Theorem 4.1], there exists a couple of random variables $(\theta_0^{(1)}, \theta_0^{(2)})$ such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\left\|\theta_0^{(1)} - \theta_0^{(2)}\right\|^2\right]$. Let $(\theta_k^{(1)}, \theta_k^{(2)})_{k\in\mathbb{N}}$ be the SGLD iterates starting from $\theta_0^{(1)}$ and $\theta_0^{(2)}$ respectively and driven by the same noise, *i.e.* for all $k \in \mathbb{N}$,

$$\begin{cases} \theta_{k+1}^{(1)} &= \theta_k^{(1)} - \gamma \left\{ \nabla U_0(\theta_k^{(1)}) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) \right\} + \sqrt{2\gamma} Z_{k+1} \ , \\ \theta_{k+1}^{(2)} &= \theta_k^{(2)} - \gamma \left\{ \nabla U_0(\theta_k^{(2)}) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\} + \sqrt{2\gamma} Z_{k+1} \ , \end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples of $\{1, \dots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k^{(1)}, \theta_k^{(2)})_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\begin{aligned} &\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 = \\ &\left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 + \gamma^2 \left\| \nabla U_0(\theta_k^{(1)}) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) - \nabla U_0(\theta_k^{(2)}) - \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\|^2 \\ &- 2\gamma \left\langle \theta_k^{(1)} - \theta_k^{(2)}, \nabla U_0(\theta_k^{(1)}) + \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(1)}) - \nabla U_0(\theta_k^{(2)}) - \frac{N}{p} \sum_{i\in S_{k+1}} \nabla U_i(\theta_k^{(2)}) \right\rangle \ . \end{aligned}$$

By **H**1 and **H**3, $\theta \mapsto \nabla U_0(\theta) + (N/p) \sum_{i\in S} \nabla U_i(\theta)$ is $\mathbb{P}$-a.s. L-co-coercive [38]. Taking the conditional expectation w.r.t. $\mathcal{F}_k$, we obtain

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 - 2\gamma \left\{1 - (\gamma L)/2\right\} \left\langle \theta_k^{(1)} - \theta_k^{(2)}, \nabla U(\theta_k^{(1)}) - \nabla U(\theta_k^{(2)}) \right\rangle \ ,$$

and by **H**2

$$\mathbb{E}\left[\left\|\theta_{k+1}^{(1)} - \theta_{k+1}^{(2)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left\{1 - 2m\gamma(1 - (\gamma L)/2)\right\} \left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2 \ .$$

Since for all $k \geq 0$, $(\theta_k^{(1)}, \theta_k^{(2)})$ belongs to $\Pi(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_2 R_{\mathrm{SGLD}}^k)$, we get by a straightforward induction

$$\mathrm{W}_2^2(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_2 R_{\mathrm{SGLD}}^k) \leq \mathbb{E}\left[\left\|\theta_k^{(1)} - \theta_k^{(2)}\right\|^2\right] \leq \left\{1 - 2m\gamma(1 - (\gamma L)/2)\right\}^k \mathrm{W}_2^2(\lambda_1, \lambda_2) \ . \quad (15)$$

By **H**1, $\lambda_1 R_{\mathrm{SGLD}} \in \mathcal{P}_2(\mathbb{R}^d)$ and taking $\lambda_2 = \lambda_1 R_{\mathrm{SGLD}}$, we get $\sum_{k=0}^{+\infty} \mathrm{W}_2^2(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_1 R_{\mathrm{SGLD}}^{k+1}) < +\infty$.By [35, Theorem 6.16], $\mathcal{P}_2(\mathbb{R}^d)$ endowed with W_2 is a Polish space. $(\lambda_1 R_{\mathrm{SGLD}}^k)_{k\geq 0}$ is a Cauchy sequence and converges to a limit $\pi_{\mathrm{SGLD}}^{\lambda_1} \in \mathcal{P}_2(\mathbb{R}^d)$. The limit $\pi_{\mathrm{SGLD}}^{\lambda_1}$ does not depend on λ_1 because, given $\lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$, by the triangle inequality

$$\mathrm{W}_2(\pi_{\mathrm{SGLD}}^{\lambda_1}, \pi_{\mathrm{SGLD}}^{\lambda_2}) \leq \mathrm{W}_2(\pi_{\mathrm{SGLD}}^{\lambda_1}, \lambda_1 R_{\mathrm{SGLD}}^k) + \mathrm{W}_2(\lambda_1 R_{\mathrm{SGLD}}^k, \lambda_2 R_{\mathrm{SGLD}}^k) + \mathrm{W}_2(\pi_{\mathrm{SGLD}}^{\lambda_2}, \lambda_2 R_{\mathrm{SGLD}}^k) \ .$$

Taking the limit $k \to +\infty$, we get $\mathrm{W}_2(\pi_{\mathrm{SGLD}}^{\lambda_1}, \pi_{\mathrm{SGLD}}^{\lambda_2}) = 0$. The limit is thus the same for all initial distributions and is denoted π_{SGLD}. π_{SGLD} is invariant for R_{SGLD} since we have for all $k \in \mathbb{N}^*$,

$$\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}) \leq \mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}^k) + \mathrm{W}_2(\pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}^k) \ .$$

Taking the limit $k \to +\infty$, we obtain $\mathrm{W}_2(\pi_{\mathrm{SGLD}}, \pi_{\mathrm{SGLD}} R_{\mathrm{SGLD}}) = 0$. Using (15), π_{SGLD} is the unique invariant probability measure for R_{SGLD}.

A.2 Proof of Theorem 2

Proof of i). Let $\gamma \in (0, 1/L]$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [35, Theorem 4.1], there exists a couple of random variables (θ_0, ϑ_0) such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\|\theta_0 - \vartheta_0\|^2\right]$. Let $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$ be the LMC and SGLDFP iterates starting from θ_0 and ϑ_0 respectively and driven by the same noise, *i.e.* for all $k \in \mathbb{N}$,

$$\begin{cases} \theta_{k+1} &= \theta_k - \gamma \nabla U(\theta_k) + \sqrt{2\gamma} Z_{k+1} \ , \\ \vartheta_{k+1} &= \vartheta_k - \gamma \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i\in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) + \sqrt{2\gamma} Z_{k+1} \ , \end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples with replacement of $\{1, \ldots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \vartheta_k\|^2 - 2\gamma \langle \theta_k - \vartheta_k, \nabla U(\theta_k) - \nabla U(\vartheta_k)\rangle + \gamma^2 A \tag{16}$$

where

$$\begin{aligned} A &= \mathbb{E}\left[\left\| \nabla U(\theta_k) - \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i\in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) \right\|^2 \middle| \mathcal{F}_k \right] \\ &= A_1 + A_2 \ , \\ A_1 &= \|\nabla U(\theta_k) - \nabla U(\vartheta_k)\|^2 \ , \\ A_2 &= \mathbb{E}\left[\left\| \nabla U(\vartheta_k) - \left(\nabla U_0(\vartheta_k) - \nabla U_0(\theta^\star) + (N/p) \sum_{i\in S_{k+1}} \{\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star)\} \right) \right\|^2 \middle| \mathcal{F}_k \right] \ . \end{aligned}$$

Denote by W the random variable equal to $\nabla U_i(\vartheta_k) - \nabla U_i(\theta^\star) - (1/N)\sum_{j=1}^N \{\nabla U_j(\vartheta_k) - \nabla U_j(\theta^\star)\}$ for $i \in \{1, \ldots, N\}$ with probability $1/N$. By **H**1 and using the fact that the subsamples $(S_k)_{k\geq 1}$ are drawn with replacement, we obtain

$$A_2 = (N^2/p)\mathbb{E}\left[\|W\|^2 |\mathcal{F}_k\right] \leq (L^2/p) \|\vartheta_k - \theta^\star\|^2 \ .$$

Combining it with (16), and using the L-co-coercivity of ∇U under **H**1 and **H**2, we get

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] \leq (1 - m\gamma) \|\theta_k - \vartheta_k\|^2 + \{(L^2\gamma^2)/p\} \|\vartheta_k - \theta^\star\|^2 \ .$$

Iterating and using Lemma 8-i), we have for $n \in \mathbb{N}$

$$\begin{aligned} &\mathrm{W}_2^2(\lambda_1 R^n_{\mathrm{LMC}}, \lambda_2 R^n_{\mathrm{FP}}) \leq \mathbb{E}\left[\|\theta_n - \vartheta_n\|^2\right] \\ &\leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{L^2\gamma^2}{p} \sum_{k=0}^{n-1} (1 - m\gamma)^{n-1-k} \mathbb{E}\left[\|\vartheta_k - \theta^\star\|^2\right] \\ &\leq (1 - m\gamma)^n \, \mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{L^2\gamma^2}{p} n (1 - m\gamma)^{n-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \, \lambda_2(\mathrm{d}\vartheta) + \frac{2L^2\gamma d}{pm^2} \ . \end{aligned}$$

Proof of ii). Denote by $\kappa = (2mL)/(m + L)$. By **H**1, **H**2 and [16, Theorem 5], we have for all $n \in \mathbb{N}$,

$$\begin{aligned} \mathrm{W}_2^2(\lambda_1 P_{n\gamma}, \lambda_2 R^n_{\mathrm{LMC}}) \leq 2\,(1 - \kappa\gamma/2)^n \, \mathrm{W}_2^2(\lambda_1, \lambda_2) + \frac{2L^2\gamma}{\kappa}(\kappa^{-1} + \gamma)\left(2d + \frac{dL^2\gamma^2}{6}\right) \\ + L^4\gamma^3(\kappa^{-1} + \gamma) \sum_{k=1}^n \delta_k \, \{1 - \kappa\gamma/2\}^{n-k} \end{aligned}$$

where for all $k \in \{1, \dots, n\}$,

$$\delta_k \leq \mathrm{e}^{-2m(k-1)\gamma} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda_1(\mathrm{d}\vartheta) + d/m \ .$$

We get the result by straightforward simplifications and using $\gamma \leq 1/L$.

Proof of iii). Let $\gamma \in (0, 1/L]$ and $\lambda_1, \lambda_2 \in \mathcal{P}_2(\mathbb{R}^d)$. By [35, Thereom 4.1], there exists a couple of random variables (θ_0, ϑ_0) such that $\mathrm{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\|\theta_0 - \vartheta_0\|^2\right]$. Let $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$ be the SGLD and SGD iterates starting from θ_0 and ϑ_0 respectively and driven by the same noise, *i.e.* for all $k \in \mathbb{N}$,

$$\begin{cases} \theta_{k+1} &= \theta_k - \gamma \left(\nabla U_0(\theta_k) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k) \right) + \sqrt{2\gamma} Z_{k+1} \ , \\ \vartheta_{k+1} &= \vartheta_k - \gamma \left(\nabla U_0(\vartheta_k) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\vartheta_k) \right) \ , \end{cases}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of standard Gaussian variables and $(S_k)_{k\geq 1}$ an i.i.d. sequence of subsamples with replacement of $\{1, \dots, N\}$ of size p. Denote by $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k, \vartheta_k)_{k\in\mathbb{N}}$. We have for $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \vartheta_k\|^2 - 2\gamma \left\langle \theta_k - \vartheta_k, \nabla U(\theta_k) - \nabla U(\vartheta_k) \right\rangle + 2\gamma d$$
$$+ \gamma^2 \mathbb{E}\left[\left\| \nabla U_0(\theta_k) + (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\theta_k) - \nabla U_0(\vartheta_k) - (N/p) \sum_{i\in S_{k+1}} \nabla U_i(\vartheta_k) \right\|^2 \middle| \mathcal{F}_k \right] \ .$$

By **H**1 and **H**3, $\theta \mapsto \nabla U_0(\theta) + (N/p) \sum_{i\in S} \nabla U_i(\theta)$ is $\mathbb{P}$-a.s. L-co-coercive and we obtain

$$\mathbb{E}\left[\|\theta_{k+1} - \vartheta_{k+1}\|^2 \middle| \mathcal{F}_k\right] \leq \{1 - 2m\gamma(1 - \gamma L/2)\} \|\theta_k - \vartheta_k\|^2 + 2\gamma d \ ,$$

which concludes the proof by a straightforward induction.

A.3 Proof of Theorem 4

Proof of i). Let $\gamma \in \left(0, \Sigma^{-1}\{1 + N/(p\sum_{i=1}^N x_i^2)\}^{-1}\right]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLD (3) started at $\theta^\star$ and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the associated filtration. For all $k \in \mathbb{N}$, $\mathbb{E}[\theta_k] = \theta^\star$. The variance of θ_k satisfies the following recursion for $k \in \mathbb{N}$

$$\begin{aligned} &\mathbb{E}\left[(\theta_{k+1} - \theta^\star)^2 \middle| \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\left\{ \theta_k - \theta^\star - \gamma \left(\Sigma(\theta_k - \theta^\star) + \rho(S_{k+1})(\theta_k - \theta^\star) + \xi(S_{k+1}) \right) + \sqrt{2\gamma} Z_{k+1} \right\}^2 \middle| \mathcal{F}_k \right] \\ &= \mu(\theta_k - \theta^\star)^2 + 2\gamma + \gamma^2 A \ , \end{aligned}$$

where

$$\mu = \mathbb{E}\left[\left\{ 1 - \gamma \left(\frac{1}{\sigma_\theta^2} + \frac{N}{\sigma_y^2 p} \sum_{i\in S} x_i^2 \right) \right\}^2 \right] \ , \quad A = \mathbb{E}\left[\left\{ \frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{\sigma_y^2 p} \sum_{i\in S} (x_i\theta^\star - y_i)\, x_i \right\}^2 \right] \ .$$

We have for μ,

$$\begin{aligned} \mu &= 1 - 2\gamma\Sigma + \gamma^2 \mathbb{E}\left[\left\{ \frac{N}{\sigma_y^2 p} \sum_{i\in S} x_i^2 - \frac{1}{\sigma_y^2} \sum_{i=1}^N x_i^2 \right\}^2 \right] + \gamma^2 \Sigma^2 \\ &= 1 - 2\gamma\Sigma + \gamma^2 \left\{ \Sigma^2 + \frac{N}{\sigma_y^4 p} \sum_{i=1}^N \left(x_i^2 - \frac{1}{N} \sum_{j=1}^N x_j^2 \right) \right\} \leq 1 - \gamma\Sigma \ , \end{aligned}$$

and for A,

$$A = \frac{N}{p} \sum_{i=1}^N \left\{ \frac{(x_i\theta^\star - y_i)\, x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2} \right\}^2 \ .$$

By a straightforward induction, we obtain that the variance of the nth iterate of SGLD started at $\theta^\star$ is for $n \in \mathbb{N}^*$

$$\int_{\mathbb{R}} (\theta - \theta^\star)^2 R^n_{\text{SGLD}}(\theta^\star, \mathrm{d}\theta) = \frac{1-\mu^n}{1-\mu} 2\gamma + \frac{1-\mu^n}{1-\mu} \frac{N\gamma^2}{p} \sum_{i=1}^{N} \left\{ \frac{(x_i\theta^\star - y_i)\, x_i}{\sigma_y^2} + \frac{\theta^\star}{N\sigma_\theta^2} \right\}^2 .$$

For SGLDFP, the additive part of the noise in the stochastic gradient disappears and we obtain similarly for $n \in \mathbb{N}^*$

$$\int_{\mathbb{R}} (\theta - \theta^\star)^2 R^n_{\text{FP}}(\theta^\star, \mathrm{d}\theta) = \frac{1-\mu^n}{1-\mu} 2\gamma .$$

To conclude, we use that for two probability measures with given mean and covariance matrices, the Wasserstein distance between the two Gaussians with these respective parameters is a lower bound for the Wasserstein distance between the two measures [18, Theorem 2.1].

The proof of ii) is straightforward.

B Proofs of Section 3.2

B.1 Proof of Proposition 5

Let θ be distributed according to π. By **H**2, for all $\vartheta \in \mathbb{R}^d$, $U(\vartheta) \geq U(\theta^\star) + (m/2)\left\|\vartheta - \theta^\star\right\|^2$ and $\mathbb{E}\left[\nabla U(\theta)\right] = 0$. By a Taylor expansion of ∇U around $\theta^\star$, we obtain

$$0 = \mathbb{E}\left[\nabla U(\theta)\right] = \nabla^2 U(\theta^\star)\left(\mathbb{E}\left[\theta\right] - \theta^\star\right) + (1/2)\,\mathrm{D}^3\, U(\theta^\star)[\mathbb{E}\left[(\theta-\theta^\star)^{\otimes 2}\right]] + \mathbb{E}\left[\mathcal{R}_1(\theta)\right] ,$$

where by **H**1, $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \left\{ \left\|\mathcal{R}_1(\vartheta)\right\| / \left\|\vartheta - \theta^\star\right\|^3 \right\} \leq L/6 . \tag{17}$$

Rearranging the terms, we get

$$\mathbb{E}\left[\theta\right] - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)^{-1}\,\mathrm{D}^3\, U(\theta^\star)[\mathbb{E}\left[(\theta-\theta^\star)^{\otimes 2}\right]] - \nabla^2 U(\theta^\star)^{-1}\mathbb{E}\left[\mathcal{R}_1(\theta)\right] .$$

To estimate the covariance matrix of π around $\theta^\star$, we start again from the Taylor expansion of ∇U around $\theta^\star$ and we obtain

$$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right] = \mathbb{E}\left[\left(\nabla^2 U(\theta^\star)(\theta-\theta^\star) + \mathcal{R}_2(\theta)\right)^{\otimes 2}\right] = \nabla^2 U(\theta^\star)^{\otimes 2}\mathbb{E}\left[(\theta-\theta^\star)^{\otimes 2}\right] + \mathbb{E}\left[\mathcal{R}_3(\theta)\right] \tag{18}$$

where by **H**1, $\mathcal{R}_2 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \left\{ \left\|\mathcal{R}_2(\vartheta)\right\| / \left\|\vartheta - \theta^\star\right\|^2 \right\} \leq L/2 , \tag{19}$$

and $\mathcal{R}_3 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is defined for all $\vartheta \in \mathbb{R}^d$ by

$$\mathcal{R}_3(\vartheta) = \nabla^2 U(\theta^\star)(\vartheta - \theta^\star) \otimes \mathcal{R}_2(\vartheta) + \mathcal{R}_2(\vartheta) \otimes \nabla^2 U(\theta^\star)(\vartheta - \theta^\star) + \mathcal{R}_2(\vartheta)^{\otimes 2} . \tag{20}$$

$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right]$ is the Fisher information matrix and by a Taylor expansion of $\nabla^2 U$ around $\theta^\star$ and an integration by parts,

$$\mathbb{E}\left[\nabla U(\theta)^{\otimes 2}\right] = \mathbb{E}\left[\nabla^2 U(\theta)\right] = \nabla^2 U(\theta^\star) + \mathbb{E}\left[\mathcal{R}_4(\theta)\right]$$

where by **H**1, $\mathcal{R}_4 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfies

$$\sup_{\vartheta \in \mathbb{R}^d} \left\{ \left\|\mathcal{R}_4(\vartheta)\right\| / \left\|\vartheta - \theta^\star\right\| \right\} \leq L . \tag{21}$$

Combining this result, (17), (18), (19), (20), (21) and $\mathbb{E}[\left\|\theta - \theta^\star\right\|^4] \leq d(d+2)/m^2$ by [5, Lemma 9] conclude the proof.

B.2 Proofs of Theorem 6 and Theorem 7

First note that under **H**1, **H**2 and **H**3, there exists $r \in [0, L/(\sqrt{p}m)]$ such that

$$\mathrm{K} \preceq r^2 (\nabla^2 U(\theta^\star))^{\otimes 2} \ , \tag{22}$$

i.e. for all $A \in \mathbb{R}^{d\times d}$,

$$\mathrm{Tr}(A^{\mathrm{T}}\,\mathrm{K}(A)) \leq r^2\,\mathrm{Tr}(A^{\mathrm{T}}(\nabla^2 U(\theta^\star))^{\otimes 2} A) \ ,$$

and where K is defined in (7). In addition, if $\liminf_{N\to+\infty} N^{-1} m > 0$, r can be chosen independently of N.

Moreover, for all $\gamma \in (0, 2/L)$, H defined in (8), is invertible and for all $\gamma \in (0, 2/\{(1+r^2)L\})$, G defined in (9), is invertible. Indeed,

$$\begin{aligned}
\mathrm{H} &= \nabla^2 U(\theta^\star) \otimes \left(\mathrm{Id} - \frac{\gamma}{2}\nabla^2 U(\theta^\star)\right) + \left(\mathrm{Id} - \frac{\gamma}{2}\nabla^2 U(\theta^\star)\right) \otimes \nabla^2 U(\theta^\star) \succ 0 \ , \\
\mathrm{G} &\succeq \nabla^2 U(\theta^\star) \otimes \mathrm{Id} + \mathrm{Id} \otimes \nabla^2 U(\theta^\star) - \gamma(1+r^2)\nabla^2 U(\theta^\star) \otimes \nabla^2 U(\theta^\star) \\
&\succeq \nabla^2 U(\theta^\star) \otimes \left(\mathrm{Id} - \frac{\gamma(1+r^2)}{2}\nabla^2 U(\theta^\star)\right) + \left(\mathrm{Id} - \frac{\gamma(1+r^2)}{2}\nabla^2 U(\theta^\star)\right) \otimes \nabla^2 U(\theta^\star) \succ 0 \ .
\end{aligned}$$

For simplicity of notation, in this Section, we use $\epsilon(\theta)$ to denote the difference between the stochastic and the exact gradients at $\theta \in \mathbb{R}^d$. More precisely, ϵ is the null function for LMC and is defined for $\theta \in \mathbb{R}^d$ by

$$\epsilon(\theta) = \frac{N}{p}\sum_{i\in S} \nabla U_i(\theta) - \sum_{j=1}^{N} \nabla U_j(\theta) \quad \text{for SGLD and SGD,} \tag{23}$$

$$\epsilon(\theta) = \nabla U_0(\theta) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\} - \nabla U(\theta) \quad \text{for SGLDFP,} \tag{24}$$

where S is a random subsample of $\{1,\dots,N\}$ with replacement of size $p \in \mathbb{N}^*$. In this setting, the update equation for LMC, SGLD and SGLDFP is given for $k \in \mathbb{N}$ by

$$\theta_{k+1} = \theta_k - (\nabla U(\theta_k) + \epsilon_{k+1}(\theta_k)) + \sqrt{2\gamma} Z_{k+1} \ , \tag{25}$$

where $(Z_k)_{k\geq 1}$ is a sequence of i.i.d. standard d-dimensional Gaussian variables and the sequence of vector fields $(\epsilon_k)_{k\geq 1}$ is associated to a sequence $(S_k)_{k\geq 1}$ of i.i.d. random subsample of $\{1,\dots,N\}$ with replacement of size $p \in \mathbb{N}^*$. We also denote by $\bar{\pi} \in \mathcal{P}_2(\mathbb{R}^d)$ the invariant probability measure of LMC, SGLDFP or SGLD.

B.2.1 Control of the moments of order 2 and 4 of LMC, SGLDFP and SGLD

Lemma 8. *Assume **H**1, **H**2 and **H**3.*

i) For all initial distribution $\lambda \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in (0, 1/L]$ and $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \,\lambda(\mathrm{d}\vartheta) + (2d)/m$$

where $(\theta_k)_{k\in\mathbb{N}}$ are the iterates of SGLDFP (5) *or LMC* (2)*.*

ii) For all initial distribution $\lambda \in \mathcal{P}_2(\mathbb{R}^d)$, $\gamma \in (0, 1/(2L)]$ and $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leq (1 - m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \,\lambda(\mathrm{d}\vartheta) + \frac{2d}{m} \\
+ \frac{2\gamma N}{mp}\sum_{i=1}^{N}\left\|\nabla U_i(\theta^\star) - \frac{1}{N}\sum_{j=1}^{N}\nabla U_j(\theta^\star)\right\|^2
\end{aligned}$$

where $(\theta_k)_{k\in\mathbb{N}}$ are the iterates of SGLD (3)*.*

Proof. i). We prove the result for SGLDFP, the case of LMC is identical. Let $\gamma \in (0, 1/L]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLDFP and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k)_{k\in\mathbb{N}}$. By (5), we have for all $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] = \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d$$
$$+ \gamma^2 \mathbb{E}\left[\left\|\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S_{k+1}} \{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\}\right\|^2 \middle| \mathcal{F}_k\right]$$

By **H**1 and **H**3, $\theta \mapsto \nabla U_0(\theta) - \nabla U_0(\theta^\star) + (N/p)\sum_{i\in S}\{\nabla U_i(\theta) - \nabla U_i(\theta^\star)\}$ is $\mathbb{P}$-a.s. L-co-coercive and we obtain

$$\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] \leq \{1 - 2m\gamma(1 - \gamma L/2)\}\|\theta_k - \theta^\star\|^2 + 2\gamma d\ .$$

A straightforward induction concludes the proof.

ii). Let $\gamma \in (0, 1/(2L)]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of SGLD and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ the filtration associated to $(\theta_k)_{k\in\mathbb{N}}$. By (3), we have for all $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] &= \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d \\
&\quad + \gamma^2 \mathbb{E}\left[\left\|\nabla U_0(\theta_k) + \frac{N}{p}\sum_{i\in S_{k+1}} \nabla U_i(\theta_k)\right\|^2 \middle| \mathcal{F}_k\right] \\
&\leq \|\theta_k - \theta^\star\|^2 - 2\gamma \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\rangle + 2\gamma d \\
&\quad + 2\gamma^2 \mathbb{E}\left[\left\|\nabla U_0(\theta_k) - \nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S_{k+1}} \{\nabla U_i(\theta_k) - \nabla U_i(\theta^\star)\}\right\|^2 \middle| \mathcal{F}_k\right] \\
&\quad + 2\gamma^2 \mathbb{E}\left[\left\|\nabla U_0(\theta^\star) + \frac{N}{p}\sum_{i\in S_{k+1}} \nabla U_i(\theta^\star)\right\|^2 \middle| \mathcal{F}_k\right]\ .
\end{aligned}$$

By **H**1 and **H**3, $\theta \mapsto \nabla U_0(\theta) + (N/p)\sum_{i\in S}\nabla U_i(\theta)$ is $\mathbb{P}$-a.s. L-co-coercive and we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{k+1} - \theta^\star\|^2 \middle| \mathcal{F}_k\right] &\leq \{1 - 2m\gamma(1 - \gamma L)\}\|\theta_k - \theta^\star\|^2 + 2\gamma d \\
&\quad + \frac{2\gamma^2 N}{p}\sum_{i=1}^N \left\|\nabla U_i(\theta^\star) - \frac{1}{N}\sum_{j=1}^N \nabla U_j(\theta^\star)\right\|^2\ .
\end{aligned}$$

A straightforward induction concludes the proof. □

Lemma 9. *Assume **H**1, **H**2 and **H**3. For all initial distribution $\lambda \in \mathcal{P}_4(\mathbb{R}^d)$, $\gamma \in (0, 1/\{12(L \vee 1)\}]$ and $k \in \mathbb{N}$,*

$$\begin{aligned}
\mathbb{E}\left[\|\theta_k - \theta^\star\|^4\right] &\leq (1 - 2m\gamma)^k \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^4 \lambda(\mathrm{d}\vartheta) \\
&\quad + \left\{12\gamma^2 \mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} k(1 - m\gamma)^{k-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda(\mathrm{d}\vartheta) \\
&\quad + \left\{\frac{2d+1}{m} + \frac{6\gamma}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}^2 \\
&\quad + \frac{2\gamma d(2+d)}{m} + \frac{4\gamma^3}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + \frac{4\gamma^2(d+2)}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\ .
\end{aligned}$$

where $(\theta_k)_{k\in\mathbb{N}}$ are the iterates of LMC (2)*, SGLD* (3) *or SGLDFP* (5)*.*

Proof. Let $\gamma \in (0, 1/\{12(L \vee 1)\}]$, $(\theta_k)_{k\in\mathbb{N}}$ be the iterates of LMC (2), SGLD (3) or SGLDFP (5) and $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the associated filtration. By developing the square, we have

$$\|\theta_1 - \theta^\star\|^4 = \Big(\|\theta_0 - \theta^\star\|^2 + 2\gamma \|Z_1\|^2 + \gamma^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2$$
$$- 2\gamma \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle + \sqrt{2\gamma} \langle \theta_0 - \theta^\star, Z_1 \rangle - (2\gamma)^{3/2} \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle \Big)^2 ,$$

and taking the conditional expectation w.r.t. $\mathcal{F}_0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0\right] = \mathbb{E}\Big[\|\theta_0 - \theta^\star\|^4 + 4\gamma^2 \|Z_1\|^4 + \gamma^4 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^4 \\
&+ 4\gamma^2 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle^2 + 2\gamma \langle \theta_0 - \theta^\star, Z_1 \rangle^2 + (2\gamma)^3 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle^2 \\
&+ 4\gamma \|Z_1\|^2 \|\theta_0 - \theta^\star\|^2 + 2\gamma^2 \|\theta_0 - \theta^\star\|^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \\
&- 4\gamma \|\theta_0 - \theta^\star\|^2 \langle \nabla U(\theta_0), \theta_0 - \theta^\star \rangle + 4\gamma^3 \|Z_1\|^2 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \\
&- 8\gamma^2 \|Z_1\|^2 \langle \nabla U(\theta_0), \theta_0 - \theta^\star \rangle - 4\gamma^3 \|\nabla U(\theta_0) + \epsilon_1(\theta_0)\|^2 \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), \theta_0 - \theta^\star \rangle \\
&- 8\gamma^2 \langle \theta_0 - \theta^\star, Z_1 \rangle \langle \nabla U(\theta_0) + \epsilon_1(\theta_0), Z_1 \rangle \,|\mathcal{F}_0\Big] .
\end{aligned}$$

By **H**1 and **H**3, $\theta \mapsto \nabla U(\theta) + \epsilon_1(\theta)$ is $\mathbb{P}$-a.s. L-co-coercive and we have for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s. ,

$$\begin{aligned}
&\|\nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star)\|^2 \le L \langle \theta - \theta^\star, \nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star) \rangle \ , \\
&\|\nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star)\|^4 \le L^2 \|\theta - \theta^\star\|^2 \langle \theta - \theta^\star, \nabla U(\theta) + \epsilon_1(\theta) - \epsilon_1(\theta^\star) \rangle \ .
\end{aligned}$$

Combining it with $\mathbb{E}\left[\|Z_1\|^4\right] = d(2+d)$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0, S_1\right] \le \|\theta_0 - \theta^\star\|^4 - 4\gamma(1 - 3\gamma L - 2\gamma^3 L^2) \|\theta_0 - \theta^\star\|^2 \\
&\times \langle \theta_0 - \theta^\star, \nabla U(\theta_0) + \epsilon_1(\theta_0) - \epsilon_1(\theta^\star) \rangle + (12\gamma^2 \|\epsilon_1(\theta^\star)\|^2 + 2\gamma(2d+1)) \|\theta_0 - \theta^\star\|^2 \\
&+ 4\gamma^2 d(2+d) + 8\gamma^4 \|\epsilon_1(\theta^\star)\|^4 + 8\gamma^3(d+2) \|\epsilon_1(\theta^\star)\|^2 \\
&- 8(d+1)\gamma^2(1 - 2\gamma L) \langle \theta_0 - \theta^\star, \nabla U(\theta_0) + \epsilon_1(\theta_0) - \epsilon_1(\theta^\star) \rangle \ .
\end{aligned}$$

By **H**2 and using $\gamma \le 1/\{12(L \vee 1)\}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_1 - \theta^\star\|^4 \middle| \mathcal{F}_0\right] \le (1 - 2m\gamma) \|\theta_0 - \theta^\star\|^4 + \left\{12\gamma^2 \mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} \|\theta_0 - \theta^\star\|^2 \\
&+ 4\gamma^2 d(2+d) + 8\gamma^4 \mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^4\right] + 8\gamma^3(d+2)\mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] \ .
\end{aligned}$$

By a straightforward induction, we have for all $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[\|\theta_n - \theta^\star\|^4\right] \le (1 - 2m\gamma)^n \mathbb{E}\left[\|\theta_0 - \theta^\star\|^4\right] \\
&+ \left\{12\gamma^2 \mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} \sum_{k=0}^{n-1} (1 - 2m\gamma)^{n-1-k} \mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \\
&+ (2m\gamma)^{-1} \left\{4\gamma^2 d(2+d) + 8\gamma^4 \mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + 8\gamma^3(d+2)\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}
\end{aligned}$$

and by Lemma 8,

$$\begin{aligned}
\mathbb{E}\left[\|\theta_n - \theta^\star\|^4\right] &\le (1 - 2m\gamma)^n \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^4 \lambda(\mathrm{d}\vartheta) \\
&+ \left\{12\gamma^2 \mathbb{E}\left[\|\epsilon_1(\theta^\star)\|^2\right] + 2\gamma(2d+1)\right\} n(1 - m\gamma)^{n-1} \int_{\mathbb{R}^d} \|\vartheta - \theta^\star\|^2 \lambda(\mathrm{d}\vartheta) \\
&+ \left\{\frac{2d+1}{m} + \frac{6\gamma}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right]\right\}^2 \\
&+ \frac{2\gamma d(2+d)}{m} + \frac{4\gamma^3}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^4\right] + \frac{4\gamma^2(d+2)}{m}\mathbb{E}\left[\|\epsilon(\theta^\star)\|^2\right] \ .
\end{aligned}$$

□

Thanks to this lemma, we obtain the following corollary. The upper bound for SGD is given by [13, Lemma 13].

Corollary 10. *Assume **H1**, **H2** and **H3**.*

i) Let $\gamma = \eta/N$ with $\eta \in (0, 1/\{24(\tilde{L} \vee 1)\}]$ and assume that $\liminf_{N\to+\infty} N^{-1}m > 0$. Then,

$$\int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{LMC}}(\mathrm{d}\theta) = d^2 O_{N\to+\infty}(N^{-2}) \ ,$$
$$\int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{FP}}(\mathrm{d}\theta) = d^2 O_{N\to+\infty}(N^{-2}) \ .$$

ii) Let $\gamma = \eta/N$ with $\eta \in (0, 1/\{24(\tilde{L} \vee 1)\}]$ and assume that $\liminf_{N\to+\infty} N^{-1}m > 0$ and that $N \geq 1/\eta$. Then,

$$\int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = d^2 O_{\eta\to 0}(\eta^2) \ , \ \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^4 \, \pi_{\mathrm{SGD}}(\mathrm{d}\theta) = d^2 O_{\eta\to 0}(\eta^2) \ .$$

B.2.2 Proofs of Theorem 6 and Theorem 7

Denote by

$$\eta_0 = \inf_{N\geq 1} \left\{ \frac{N}{12(L \vee 1)} \wedge \frac{2N}{(1+r^2)L} \right\} > 0 \ , \tag{26}$$

and set $\gamma = \eta/N$ with $\eta \in (0, \eta_0)$. Let $\delta \in \{0, 1\}$ be equal to 1 for LMC, SGLDFP and SGLD and 0 for SGD. Let θ_0 be distributed according to $\bar{\pi}$. By (25) and using a Taylor expansion around $\theta^\star$ for ∇U, we obtain

$$\theta_1 - \theta^\star = \theta_0 - \theta^\star - \gamma \left(\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star) + \mathcal{R}_1(\theta_0) + \epsilon_1(\theta_0) \right) + \delta\sqrt{2\gamma} Z_1 \ ,$$

where by **H**1, $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d} \left\{ \|\mathcal{R}_1(\theta)\| \, / \, \|\theta - \theta^\star\|^2 \right\} \leq L/2 \ . \tag{27}$$

Taking the tensor product and the expectation, and using that $\theta_0, \epsilon_1, Z_1$ are mutually independent, we obtain

$$\begin{aligned} \mathrm{H}\,\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] &= 2\delta\,\mathrm{Id} + \gamma\mathbb{E}\left[\epsilon_1(\theta_0)^{\otimes 2}\right] + \mathbb{E}\left[\mathcal{R}_1(\theta_0) \otimes \{\theta_0 - \theta^\star\} + \{\theta_0 - \theta^\star\} \otimes \mathcal{R}_1(\theta_0)\right] \\ &+ \gamma\mathbb{E}\left[\mathcal{R}_1(\theta_0)^{\otimes 2} + \{\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star)\} \otimes \mathcal{R}_1(\theta_0) + \mathcal{R}_1(\theta_0) \otimes \nabla^2 U(\theta^\star)(\theta_0 - \theta^\star)\right] \ . \end{aligned} \tag{28}$$

For LMC, ϵ_1 is the null function and by Corollary 10-i), (27) and (28), we obtain (10). Regarding SGLDFP, SGLD and SGD, by a Taylor expansion of ϵ_1 around $\theta^\star$, we get for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s. ,

$$\epsilon_1(\theta) = \epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star) + \mathcal{R}_2(\theta)$$

where by **H**1,$\mathcal{R}_2 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d} \left\{ \|\mathcal{R}_2(\theta)\| \, / \, \|\theta - \theta^\star\|^2 \right\} \leq L/2 \ . \tag{29}$$

Therefore, taking the tensor product and the expectation, we obtain

$$\mathbb{E}\left[\epsilon_1(\theta_0)^{\otimes 2}\right] = \mathbb{E}\left[\epsilon_1(\theta^\star)^{\otimes 2}\right] + (\nabla\epsilon_1(\theta^\star))^{\otimes 2}\,\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] + \mathbb{E}\left[\mathcal{R}_3(\theta_0)\right] \tag{30}$$

where $\mathcal{R}_3 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is defined for all $\theta \in \mathbb{R}^d$, $\mathbb{P}$-a.s. ,

$$\begin{aligned} \mathcal{R}_3(\theta) &= \epsilon_1(\theta^\star) \otimes \{\nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} + \{\nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} \otimes \epsilon_1(\theta^\star) \\ &+ \{\epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} \otimes \mathcal{R}_2(\theta) + \mathcal{R}_2(\theta) \otimes \{\epsilon_1(\theta^\star) + \nabla\epsilon_1(\theta^\star)(\theta - \theta^\star)\} + \mathcal{R}_2^{\otimes 2}(\theta) \ . \end{aligned} \tag{31}$$

Note that $\mathrm{K} = \mathbb{E}\left[(\nabla\epsilon_1(\theta^\star))^{\otimes 2}\right]$. For SGLDFP, $\epsilon_1(\theta^\star) = 0$ a.s. By Corollary 10-i), (27), (28), (29), (30) and (31), we obtain (11).

Regarding SGLD and SGD, we have $\mathbb{E}\left[\epsilon_1(\theta^\star)^{\otimes 2}\right] = (N/p)\,\mathrm{M}$ where M is defined in (14). By Corollary 10-ii), (27), (28), (29), (30) and (31), we obtain (12) and (13).

For the mean of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD}, by a Taylor expansion around $\theta^\star$ for ∇U of order 3, we obtain

$$\theta_1 - \theta^\star = \theta_0 - \theta^\star - \gamma\left(\nabla^2 U(\theta^\star)(\theta_0 - \theta^\star) + (1/2)\,\mathrm{D}^3\,U(\theta^\star)(\theta_0 - \theta^\star)^{\otimes 2} + \mathcal{R}_4(\theta_0) + \epsilon_1(\theta_0)\right) + \delta\sqrt{2\gamma}Z_1\ ,$$

where by **H**1, $\mathcal{R}_4 : \mathbb{R}^d \to \mathbb{R}^d$ satisfies

$$\sup_{\theta\in\mathbb{R}^d}\left\{\|\mathcal{R}_4(\theta)\| \,/\, \|\theta - \theta^\star\|^3\right\} \le L/6\ . \tag{32}$$

Taking the expectation and using that θ_1 is distributed according to $\bar{\pi}$, we get

$$\mathbb{E}\left[\theta_0\right] - \theta^\star = -(1/2)\nabla^2 U(\theta^\star)\,\mathrm{D}^3\,U(\theta^\star)[\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right]] - \nabla^2 U(\theta^\star)^{-1}\mathbb{E}\left[\mathcal{R}_4(\theta_0)\right]\ .$$

(10), (11), (12),(13), (32) and Corollary 10 conclude the proof.

C Means and covariance matrices of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} in the Bayesian linear regression

In this Section, we provide explicit expressions of the covariance matrices of $\pi_{\mathrm{LMC}}, \pi_{\mathrm{FP}}, \pi_{\mathrm{SGLD}}$ and π_{SGD} in the context of the Bayesian linear regression. In this setting, the algorithms are without bias, *i.e.*

$$\int_{\mathbb{R}^d}\theta\pi_{\mathrm{LMC}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{FP}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi_{\mathrm{SGD}}(\mathrm{d}\theta) = \int_{\mathbb{R}^d}\theta\pi(\mathrm{d}\theta) = \theta^\star\ . \tag{33}$$

Before giving the expressions of the variances in Theorem 11, we define $\mathrm{T} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ for all $A \in \mathbb{R}^{d\times d}$ by

$$\mathrm{T}(A) = \mathbb{E}\left[\left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S} x_i x_i^{\mathrm{T}} - \Sigma\right)^{\otimes 2} A\right] = \frac{N}{p}\sum_{i=1}^N\left(\frac{x_i x_i^{\mathrm{T}}}{\sigma_y^2} + \frac{\mathrm{Id}}{N\sigma_\theta^2} - \frac{\Sigma}{N}\right)^{\otimes 2} A\ , \tag{34}$$

where S is a random subsample of $\{1,\dots,N\}$ with replacement of size $p \in \mathbb{N}^*$. Note that, in this setting, $\tilde{L} = \max_{i\in\{1,\dots,N\}}\|x_i\|^2$ and m is the smallest eigenvalue of Σ. There exists $r \in [0, L/(\sqrt{p}m)]$ such that

$$\mathrm{T} \preceq r^2\Sigma^{\otimes 2} \tag{35}$$

i.e. for all $A \in \mathbb{R}^{d\times d}$, $\mathrm{Tr}(A^{\mathrm{T}}\,\mathrm{T}\cdot A) \le r^2\,\mathrm{Tr}(A^{\mathrm{T}}\Sigma^{\otimes 2}A)$. Assuming that $\liminf_{N\to+\infty} N^{-1}m > 0$, r can be chosen independently of N.

Theorem 11. *Consider the case of the Bayesian linear regression. We have for all $\gamma \in (0, 2/L)$*

$$\int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_{\mathrm{LMC}}(\mathrm{d}\theta) = \left(\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma\Sigma\otimes\Sigma\right)^{-1}(2\,\mathrm{Id})\ ,$$

and for all $\gamma \in \left(0, 2/\{(1+r^2)L\}\right)$,

$$\int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_{\mathrm{FP}}(\mathrm{d}\theta) = \left\{\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1}(2\,\mathrm{Id})\ ,$$

$$\int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_{\mathrm{SGLD}}(\mathrm{d}\theta) = \left\{\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1} \cdot\left\{2\,\mathrm{Id} + \frac{\gamma N}{p}\sum_{i=1}^N\left(\frac{(x_i^{\mathrm{T}}\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2}\right)^{\otimes 2}\right\}\ ,$$

$$\int_{\mathbb{R}^d}(\theta - \theta^\star)^{\otimes 2}\pi_{\mathrm{SGD}}(\mathrm{d}\theta) = \left\{\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}^{-1} \cdot\frac{\gamma N}{p}\sum_{i=1}^N\left(\frac{(x_i^{\mathrm{T}}\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2}\right)^{\otimes 2}\ .$$

Proof. We prove the result for SGLD, the adaptation to the other algorithms is immediate. Let $\gamma \in \left(0, 2/\{(1+r^2)L\}\right)$, θ_0 be distributed according to π_{SGLD} and θ_1 be given by (3). By definition of π_{SGLD}, θ_1 is distributed according to π_{SGLD}. We have

$$\mathbb{E}\left[(\theta_1 - \theta^\star)^{\otimes 2}\right] = \mathbb{E}\left[\left[\left\{\mathrm{Id} - \gamma\left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S_1} x_i x_i^{\mathrm{T}}\right)\right\}(\theta_0 - \theta^\star) - \gamma\left(\frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S_1}(x_i^{\mathrm{T}}\theta^\star - y_i)x_i\right) + \sqrt{2\gamma}Z_1\right]^{\otimes 2}\right].$$

Using that θ_0, S_1, Z_1 are mutually independent, we obtain

$$\left\{\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma\mathbb{E}\left[\left(\frac{\mathrm{Id}}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S_1} x_i x_i^{\mathrm{T}}\right)^{\otimes 2}\right]\right\}\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] = 2\,\mathrm{Id} + \gamma\mathbb{E}\left[\left(\frac{\theta^\star}{\sigma_\theta^2} + \frac{N}{p\sigma_y^2}\sum_{i\in S_1}(x_i^{\mathrm{T}}\theta^\star - y_i)x_i\right)^{\otimes 2}\right]$$

and

$$\left\{\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})\right\}\mathbb{E}\left[(\theta_0 - \theta^\star)^{\otimes 2}\right] = 2\,\mathrm{Id} + \frac{\gamma N}{p}\sum_{i=1}^{N}\left(\frac{(x_i^{\mathrm{T}}\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2}\right)^{\otimes 2}.$$

On $\mathbb{R}^{d\times d}$ equipped with the Hilbert-Schmidt inner product, $\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})$ is a positive definite operator. Indeed, by (35),

$$\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T}) \succeq \mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(1+r^2)\Sigma^{\otimes 2} = \left(\mathrm{Id} - \gamma\frac{1+r^2}{2}\Sigma\right)\otimes\Sigma + \Sigma\otimes\left(\mathrm{Id} - \gamma\frac{1+r^2}{2}\Sigma\right) \succ 0$$

for $\gamma \in \left(0, 2/\{(1+r^2)L\}\right)$. $\mathrm{Id}\otimes\Sigma + \Sigma\otimes\mathrm{Id} - \gamma(\Sigma^{\otimes 2} + \mathrm{T})$ is thus invertible, which concludes the proof. □

The covariance matrices make clearly visible the different origins of the noise. The Gaussian noise is responsible of the term $2\,\mathrm{Id}$, while the multiplicative and additive parts of the stochastic gradient (see (6)) are related to the operator T and to the term

$$\frac{\gamma N}{p}\sum_{i=1}^{N}\left(\frac{(x_i^{\mathrm{T}}\theta^\star - y_i)x_i}{\sigma_y^2} + \frac{\theta^\star}{\sigma_\theta^2}\right)^{\otimes 2} \tag{36}$$

respectively.

Denote by

$$\eta_1 = \inf_{N\geq 1}\left\{\frac{2N}{L} \wedge \frac{2N}{(1+r^2)L}\right\} > 0\,. \tag{37}$$

Corollary 12. *Consider the case of the Bayesian linear regression. Set $\gamma = \eta/N$ with $\eta \in (0, \eta_1)$ and assume that $\liminf_{N\to+\infty} N^{-1}m > 0$.*

$$\int_{\mathbb{R}^d}\|\theta - \theta^\star\|^2\,\pi_{\text{LMC}}(\mathrm{d}\theta) = d\Theta_{N\to+\infty}(N^{-1})\,,\quad \int_{\mathbb{R}^d}\|\theta - \theta^\star\|^2\,\pi_{\text{FP}}(\mathrm{d}\theta) = d\Theta_{N\to+\infty}(N^{-1})\,,$$
$$\int_{\mathbb{R}^d}\|\theta - \theta^\star\|^2\,\pi_{\text{SGLD}}(\mathrm{d}\theta) = \eta d\Theta_{N\to+\infty}(1)\,,\quad \int_{\mathbb{R}^d}\|\theta - \theta^\star\|^2\,\pi_{\text{SGD}}(\mathrm{d}\theta) = \eta d\Theta_{N\to+\infty}(1)\,.$$

Recall that, according to the Bernstein-von Mises theorem, the variance of π is of the order d/N when N is large. The corollary confirms that π_{SGLD} is very far from π when the constant step size γ is chosen proportional to $1/N$.

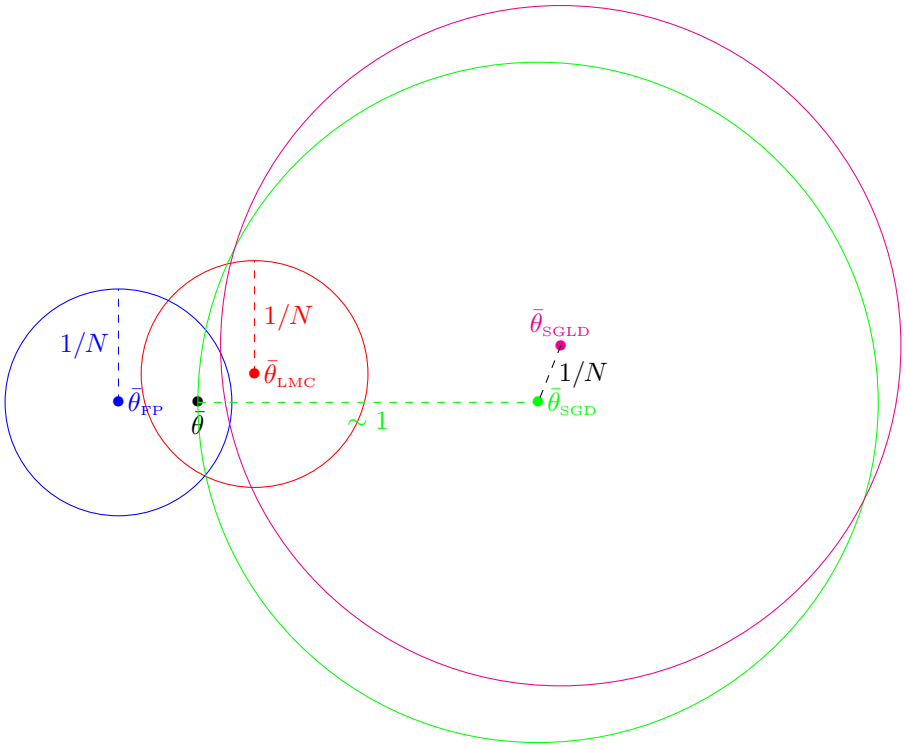

Figure 5: Illustration of Proposition 5, Theorem 6 and Theorem 7 in the asymptotic $N \to +\infty$. $\bar{\theta}$, $\bar{\theta}_{\text{SGD}}$, $\bar{\theta}_{\text{LMC}}$, $\bar{\theta}_{\text{FP}}$ and $\bar{\theta}_{\text{SGLD}}$ are the means under the stationary distributions π, π_{SGD}, π_{LMC}, π_{FP} and π_{SGLD}, respectively. The associated circles indicate the order of magnitude of the covariance matrix. While LMC and SGLDFP concentrate to the posterior mean $\bar{\theta}$ with a covariance matrix of the order $1/N$, SGLD and SGD are at a distance of order ~ 1 of $\bar{\theta}$ and do not concentrate as $N \to +\infty$.

D Illustration of Proposition 5, Theorem 6 and Theorem 7

We provide in Figure 5 an illustration of the results of Section 3.2 as the number of data items N goes to infinity.

References

[1] S. Ahn, A. K. Balan, and M. Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[2] S. Ahn, B. Shahbaba, and M. Welling. Distributed stochastic gradient MCMC. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1044–1052, Bejing, China, 22–24 Jun 2014. PMLR.

[3] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth. Control variates for stochastic gradient MCMC. *ArXiv e-prints 1706.05439*, June 2017.

[4] R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

[5] N. Brosse, A. Durmus, É. Moulines, and S. Sabanis. The Tamed Unadjusted Langevin Algorithm. *ArXiv e-prints*, Oct. 2017.

[6] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett, and M. I. Jordan. On the theory of variance reduction for stochastic gradient Monte Carlo. *ArXiv e-prints 1802.05431*, Feb. 2018.

[7] C. Chen, N. Ding, and L. Carin. On the convergence of Stochastic Gradient MCMC algorithms with high-order integrators. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2278–2286. Curran Associates, Inc., 2015.

[8] C. Chen, W. Wang, Y. Zhang, Q. Su, and L. Carin. A convergence analysis for a class of practical variance-reduction stochastic gradient MCMC. *ArXiv e-prints 1709.01180*, Sept. 2017.

[9] T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian Monte Carlo. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1683–1691, 2014.

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

[11] A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 678–689, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

[12] A. S. Dalalyan and A. G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *ArXiv e-prints 1710.00095*, Sept. 2017.

[13] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *ArXiv e-prints*, July 2017.

[14] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, and H. Neven. Bayesian sampling using stochastic gradient thermostats. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3203–3211, Cambridge, MA, USA, 2014. MIT Press.

[15] K. A. Dubey, S. J. Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing. Variance reduction in stochastic gradient Langevin dynamics. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1154–1162. Curran Associates, Inc., 2016.

[16] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *ArXiv e-prints 1605.01559*, May 2016.

[17] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 06 2017. doi: 10.1214/16-AAP1238.

[18] M. Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203. doi: 10.1002/mana.19901470121.

[19] U. Grenander. Tutorial in pattern theory. Division of Applied Mathematics, Brown University, Providence, 1983.

[20] U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56(4):549–603, 1994. ISSN 0035-9246. With discussion and a reply by the authors.

[21] L. Hasenclever, S. Webb, T. Lienart, S. Vollmer, B. Lakshminarayanan, C. Blundell, and Y. W. Teh. Distributed Bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *Journal of Machine Learning Research*, 18(106):1–37, 2017.

[22] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001.

[23] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Graduate Texts in Mathematics. Springer New York, 1991. ISBN 9780387976556.

[24] A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: cutting the Metropolis-hastings budget. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–181–I–189. JMLR.org, 2014.

[25] C. Li, C. Chen, D. Carlson, and L. Carin. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 1788–1794. AAAI Press, 2016.

[26] W. Li, S. Ahn, and M. Welling. Scalable MCMC for mixed membership stochastic blockmodels. In *Artificial Intelligence and Statistics*, pages 723–731, 2016.

[27] Y.-A. Ma, T. Chen, and E. Fox. A complete recipe for stochastic gradient MCMC. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2917–2925. Curran Associates, Inc., 2015.

[28] T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis. The true cost of stochastic gradient Langevin dynamics. *ArXiv e-prints 1706.02692*, June 2017.

[29] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi: 10.1137/070704277.

[30] S. Patterson and Y. W. Teh. Stochastic gradient riemannian Langevin dynamics on the probability simplex. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3102–3110. Curran Associates, Inc., 2013.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 1350-7265. doi: 10.2307/3318418.

[33] I. Sato and H. Nakagawa. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 982–990, Bejing, China, 22–24 Jun 2014. PMLR.

[34] Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient Langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

[35] C. Villani. *Optimal transport : old and new*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

[36] S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. *Journal of Machine Learning Research*, 17 (159):1–48, 2016.

[37] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 681–688, USA, 2011. Omnipress. ISBN 978-1-4503-0619-5.

[38] D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM J. on Optimization*, 6(3):714–726, Mar. 1996. ISSN 1052-6234. doi: 10.1137/S1052623494250415.