Langevin MC,
or the curse of dimensionality

B. Han, T.M. Hodgson, M. Holden & M. Puza

March 13, 2019

Contents

1 Introduction

In Bayesian statistics, we are interested in performing inference on the posterior distribution of a parameter, θ. This is calculated using Bayes rule

$$\pi(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)}$$

where $f(x|\theta)$ is the likelihood function of the data, and $p(\theta)$ is the prior distribution on the parameter. The term on the denominator, $f(x) = \int f(x)|\theta)p(\theta)d\theta$ is a normalising constant, such that π is a distribution. In general, this normalising constant is difficult to calculat analytically, and so the posterior is only known up to proportionality.

$$\pi(\theta|x) \propto f(x|\theta)p(\theta)$$

This means we cannot easily make inferential statements about the parameter θ. To solve this problem, we use Markov chain Monte Carlo methods to draw samples from the posterior distribution, and use these samples to make inferences on the parameter.

1.1 Markov Chain Monte Carlo

Monte Carlo methods are a class of algorithms that replace difficult or impossible analytical probability calculations with numerical approximations. Given a random variable X on a probability space $(\Omega, \mathcal{F}, \pi)$ with distribution given by π consider the problem of calculating

$$\mathbb{E}_\pi[g(X)] = \int_\Omega g(x)d\pi$$

If the distribution π does not take a well-known form, then analytically solving this expectation is very difficult. Instead, Monte Carlo integration uses the strong law of large numbers to approximate the integral. If X_n are a sequence of i.i.d. random variables distributed according to π, then

$$\frac{1}{N}\sum_{n=1}^{N} g(X_n) \to \mathbb{E}_\pi[g(X)] \quad \text{a.s. as } N \to \infty.$$

Hence we can approximate integrals by taking independent samples from the distribution π. However, for an arbitrary distribution, it is not necessarily possible to find sample independently in an efficient way. Markov chain Monte Carlo allows us to dispense with this assumption. Rather than sampling independently, we can instead construct a Markov chain with π as its invariant distribution. This chain can then be used to generate dependent samples, which can be used for Monte Carlo integration provided the chain is ergodic with respect to π.

Definition 1.1 (Ergodicity)**.** Let $T : \Omega \to \Omega$ be a probability-preserving transformation on a probability space $(\Omega, \mathcal{F}, P)$. Then we say T is *ergodic* if for every $F \in \mathcal{F}$ with $T^{-1}(F) = F$, either $P(F) = 0$ or $P(F) = 1$.

Intuitively, this condition means that the process explores the whole space, without becoming stuck in a subregion. The ergodic theorem then provides the theoretical justification that permits the use of dependent samples for Monte Carlo integration.

Theorem 1.2 (Ergodic Theorem)**.** *Let f be measurable, $\mathbb{E}_\pi(|f|) < \infty$, and T be an ergodic probability-preserving transformation. Then with probability 1:*

$$\frac{1}{N}\sum_{n=1}^{N} f(T^n(x)) \to \mathbb{E}_\pi[f] \quad \textit{a.s. as } N \to \infty.$$

In other words, if T is ergodic, time averages converge to space averages. Hence, if we can construct a Markov chain which is ergodic with respect to a target measure π, it can be used to generate Monte Carlo samples. One method of constructing such a Markov chain is by using Langevin dynamics.

2 Langevin Monte Carlo Algorithms

The Langevin equation is a stochastic differential equation (SDE) originally developed to model the movement of a Brownian particle [9]. The form of interest here is the *overdamped* Langevin equation, in which the particle experiences no average acceleration. The equation is thus

$$\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t. \tag{1}$$

Here, W_t is a d-dimensional Wiener process (Brownian motion) and $U : \mathbb{R}^d \to \mathbb{R}$ is the potential function. The equation can be thought of as modelling a particle in a potential well with shape U. As each particle moves randomly, it is natural to ask what is the average position of many particles in such a well? It can be shown that in fact the position of a particle moving according to the above dynamics is exactly π +++Reference to earlier subsection/first mention of distribution+++. For a diffusion process this is called the *stationary distribution*[a] To show that π is indeed the stationary distribution the following lemma

Lemma. *For a one-dimensional Itô diffusion +++add conditions on diffusion/drift+++,*

$$\mathrm{d}X_t = \mu(X_t)\,\mathrm{d}t + \sigma^2(X_t)dW_t,$$

the Fokker-Planck operator, $\mathcal{L}^*$, is

$$\mathcal{L}^* := -\partial_x(\mu(x)\cdot) + \frac{1}{2}\partial_x^2(\sigma^2(x)\cdot).$$

A measure π is invariant for the diffusion if and only if

$$\mathcal{L}^*\pi = 0$$

The proof of this is omitted however it can be seen by forming the Fokker-Planck equation for the probability density of the diffusion. The proof that π is the stationary measure of Equation (1) is given only in the one dimensional case, however it is extendable to higher dimensions. For the Langevin equation, the Fokker-Planck operator is

$$\mathcal{L}^* = \partial_x(U'(x)\cdot) + \partial_{xx}\,\cdot\,.$$

So it remains to calculate $\mathcal{L}^*\pi$.

$$\begin{aligned}
\mathcal{L}^*\pi &= \frac{\partial}{\partial x}\left[U'(x)\pi(x) + \frac{\partial}{\partial x}\pi(x)\right] \\
&= \frac{\partial}{\partial x}\left[U'(x)\mathcal{Z}\mathrm{e}^{-U(x)} + \left(-U'(x)\mathcal{Z}\mathrm{e}^{-U(x)}\right)\right] \\
&= \frac{\partial}{\partial x}[0] \\
&= 0
\end{aligned}$$

Hence π is indeed the invariant measure of (1). ■

Although this shows that the Langevin equation has an invariant measure, the question of convergence to this measure remains unanswered. Roberts and Tweedie give the following restriction [14].

Theorem 2.1 (Theorem 2.1, [14])**.** *Let $P_X^t(x, A) = \mathbb{P}(X_t \in A | X_0 = x_0)$ and suppose that $\nabla U(x)$ is continuously differentiable and that, for some $N, a, b < \infty$,*

$$\nabla U(x) \cdot x \le a|x|^2 + b, \qquad |x| > N.$$

Then the measure π is invariant for the Langevin diffusion X. Moreover, for all $x \in \mathbb{R}^d$ and Borel sets A,

$$\|P_X^t(x,\cdot) - \pi\| = \frac{1}{2}\sup_A |P_X^t(x, A) - \pi(A)| \to 0$$

+++Should this norm be an integral? Add exponentially fast convergence/spectral gap inequality? Figure of $U = x^2/2$? 2d? +++

[a] Another common term is *invariant measure* +++

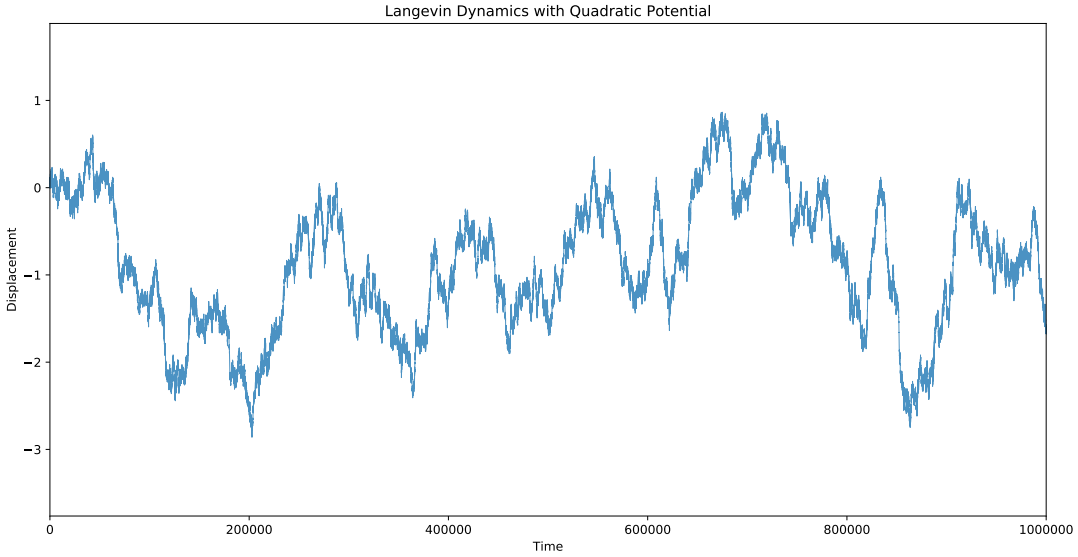

Figure 1: Simulating Langevin dynamics in one dimension with a quadratic potential $U(x) = x^2/2$

The problem of sampling from the high dimensional distribution has been reduced to being able to accurately simulate Langevin dynamics. However, this is not as simple as it sounds. To simulate the continuous process (1), it must first be discretised. However, doing so may not preserve the convergence to the invariant measure. The discretised process may not have the same stationary measure or it may not even exist. This means that the method used to discretise must be chosen carefully to ensure good convergence properties. The most natural way to discretise an SDE is to use the stochastic analogue of the (forward) Euler method used on ordinary differential equations, known as the Euler-Maruyama (EM) method. Doing so leads to the Unadjusted Langevin Algorithm (`ULA`).

2.1 The Unadjusted Langevin Algorithm

Applying the Euler-Maruyama method to Equation (1) gives the following iterative scheme.

$$X_{n+1} = X_n - h\nabla U(X_n) + \sqrt{2h}Z_{n+1}, \qquad X_0 = x_0$$

Here the Z_n are i.i.d. standard normal random variables and h is the step size. This is equivalent to $X_{n+1} \sim N(X_n - h\nabla U(X_n), 2hI_d)$.[b] A simple example shows that this discretisation does not converge to π. Let π be a standard Gaussian distribution, that is $U(x) = |x|^2/2$ and choose $h = 1$. Then the update is given by

$$\begin{aligned} X_{n+1} &\sim N(X_n - \nabla U(X_n), 2) \\ &\sim N(X_n - X_n, 2) \\ &\sim N(0, 2) \nsim \pi. \end{aligned}$$

So the chain converges immediately, but to the wrong distribution. Let π_h^{ULA} denote the stationary distribution of `ULA` with a stepsize h. This is not the only issue that can occur. As well as not converging to the correct distribution, the discretised chain may not be ergodic, even when the continuous diffusion is exponentially ergodic [14]. In particular, the algorithm misbehaves when the gradient of the potential is superlinear. That is,

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{\|x\|} = +\infty.$$

To mitigate these issues there are two main approaches: taming the gradient and Metropolisation. A further third method involves using a different discretisation scheme. Our main focus will be the former, although all three approaches will be discussed.

[b] I_d denotes the $d \times d$ identity matrix.

2.2 MALA

Before describing the Metropolis-adjusted Langevin algorithm `MALA`, it is pertinent at this point to recall the random walk Metropolis-Hastings algorithm `RWM` [8, 11]. This popular variant of the Metropolis-Hastings algorithm *proposes* values and then accepts/rejects them according to some probability α. So given X_n, propose a candidate Y_{n+1} as

$$Y_{n+1} = X_n + \sqrt{2h}Z_{n+1}.$$

Once again, h is the stepsize and Z is a normal random variable. Then, accept or reject this proposal using Metropolis rejection, that is with some probability

$$\alpha(X_n, Y_{n+1}) = 1 \wedge \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})}.^{\text{c}}$$

Here $q(x, y)$ is the transition probability, $\mathbb{P}(Y_{n+1} = y|X_n = x) \sim N(X_n, h^2)$. This rejection step is key in creating a kernel that is reversible and thus invariant for the measure π.

`MALA` can be seen as another variant of the Metropolis-Hastings algorithm, using Langevin dynamics to propose new states. It is perhaps better understood as `ULA` but with an added Metropolis rejection step [14]. Adding this rejection step means the algorithm always has the correct invariant distribution, although convergence is still not guaranteed as the following theorem shows.

Theorem 2.2 (Theroem 4.2, [14])**.** *If π is bounded, and*

$$\liminf_{\|x\|\to\infty} \frac{\|\nabla U(x)\|}{\|x\|} > \frac{4}{h}$$

then the `MALA` chain is not exponentially ergodic. +++define exp ergodic+++

So it can be seen that `MALA` is not without its issues, and does not solve all the problems of `ULA`. The concept of taming was introduced to try and reduce the magnitude of these problems.

2.3 Taming the Gradient

We have seen that both `ULA` and `MALA` run into issues when the gradient of the potential is superlinear. Given an SDE such as (1), taming adjusts the drift coefficient in such a way that preserves the invariant measure and improves speed of convergence [3, 14, 15]. To do this, a family of drift functions $(G_h)_{h>0}$, $G_h : \mathbb{R}^d \to \mathbb{R}^d$ are introduced. The SDE to be discretised is thus

$$\mathrm{d}X_t = -G_h(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t.$$

Applying the Euler-Maruyama method gives the following Markov chain

$$X_{k+1} = X_k - hG_h(X_k) + \sqrt{2h}Z_{k+1}, \qquad X_0 = x_0.$$

To preserve the invariant measure, some restrictions must be placed on $(G_h)_{h>0}$, namely that they are 'close' to ∇U (**A1**) while **A2** ensures ergodicity is preserved and improves stability [3].

A1 For all $h > 0, G_h$ is continuous. There exist $\alpha \geq 0, C_\alpha < +\infty$ such that for all $h > 0$ and $x \in \mathbb{R}^d$,

$$\|G_h(x) - \nabla U(x)\| \leq hC_\alpha(1 + \|x\|^\alpha).$$

A2 For all $h > 0$,

$$\liminf_{\|x\|\to\infty} \left[\left\langle \frac{x}{\|x\|}, G_h(x)\right\rangle - \frac{h}{2\|x\|}\|G_h(x)\|^2\right] > 0$$

[c] Here $t \wedge s = \min\{t, s\}$.

Here we consider two specific taming functions,

$$T_h(x) = \frac{\nabla U(x)}{1 + h\|\nabla U(x)\|}, \qquad T_h^{\text{RT}} = \frac{\nabla U(x)}{1 \vee h\|\nabla U(x)\|}.$$

Brosse et al. introduced and studied T_h whilst Roberts & Tweedie suggested T_h^{RT}, later analysed by Bou-Rabee & Vanden-Eijnden [2, 3, 14]. Both taming functions retain the direction of the gradient, only reducing the magnitude of its effect. The latter is the usual `ULA` until the gradient gets large enough ($\|\nabla U(x)\| > 1/h$), at which point it begins normalising. In contrast, the first will always tame, regardless of size of the gradient. However for the scaling to have noticeable effect, the gradient must be $\mathcal{O}(h^{-1})$. When T_h is the taming function, the algorithm will be referred to as `tULA`, the tamed unadjusted Langevin algorithm. When the second is applied, it will be called `MALTA`, the Metropolis adjusted Langevin truncated algorithm after [14]. Any tamed algorithm using T_h will be prefixed with a lowercase `t`. For a proof that T_h satisfies **A1** and **A2**, see [3, Lemma 2].
When the problem is ill-conditioned, taming the gradient does not help +++Kostas' example, ill-cond Gauss does this motivate coordinatewise?+++.

2.3.1 tULA/c

So far, the gradient has only been tamed globally. This has a negative side effect in that it heavily restricts movement in all dimensions, regardless of whether the gradient is superlinear in that direction. A solution to this is to use coordinatewise taming with the following drift.

$$T_h^c(x) = \left(\frac{\partial_i U(x)}{1 + h|\partial_i U(x)|} \right)_{i=\{1,\dots,d\}}$$

This allows each dimension to be scaled individually. Any algorithm with coordinate-wise taming will be suffixed with a lowercase `c`.

2.4 Discretise Differently

An alternative approach is to use a different discretisation of the SDE (1), which we consider in this section. The first is an extension of the Euler method [16], while the latter uses a non-Markovian scheme developed for use in molecular dynamics [10].

2.4.1 Higher Order Langevin Algorithm

As in the ordinary case, the Euler-Maruyama method is not the only way of discretising an SDE. One can also take a higher order expansion, analagous to the Runge-Kutta method, known as the order 1.5 Wagner-Platen expansion[d]. For a one dimensional Langevin diffusion (1), this is

$$X_{n+1} = X_n - hU'(X_n) + \sqrt{2h}Z_n - \sqrt{2}U''(X_n)\tilde{Z}_n + \frac{h^2}{2}\Big[U'(X_n)U''(X_n) - U'''(X_n)\Big].$$

Here, $\tilde{Z}_n$ is defined as

$$\tilde{Z}_n = \int_{t_n}^{t_{n+1}} \int_{t_n}^{s} \mathrm{d}W_r \, \mathrm{d}s.$$

This is a d-dimensional Gaussian random variable with mean 0_d and covariance $\frac{1}{3}h^3 I_d$. Use an Itô-Taylor expansion

$$X_{n+1} = X_n + \mu_h(X_n)h + \sigma_h(X_n)\sqrt{h}Z_{n+1}$$

where

$$\mu_h(x) = -\nabla U_h(x) + \frac{h}{2}\left(\left(\nabla^2 U \nabla U\right)_h(x) - \vec{\Delta}(\nabla U)_h(x) \right),$$

and $\sigma_h(x) = \operatorname{diag}\left(\left(\sigma_h^{(k)}(x) \right)_{k\in\{1,\dots,d\}} \right)$ with,

$$\sigma_h^{(k)}(x) = \sqrt{2 + \frac{2h^2}{3}\sum_{j=1}^{d} |\nabla^2 U_h^{(k,j)}(x)|^2 - 2h\nabla^2 U_h^{(k,k)}(x)}$$

ALSO need to define the h subscript, i.e. the tamed variables. Although fn depends on gamma it doesn't indicate that taming has occurred.

[d] Or the stochastic Runge-Kutta method [**?**]

2.4.2 LM

[10] Non-Markovian scheme,

$$X_{n+1} = X_n + h\nabla U(X_n) + \sqrt{\frac{h}{2}}\,(Z_n + Z_{n+1})$$

2.5 Other Methods

2.5.1 RWM

Popular variant of the Metropolis-Hastings algorithm (CITE) with a normal proposal.

$$U_{n+1} = X_n + \sqrt{2h}Z_{n+1}$$

Calculate acceptance probability

$$\alpha(X_n, U_{n+1}) = 1 \wedge \frac{\pi(U_{n+1})q(U_{n+1}, X_n)}{\pi(X_n)q(X_n, U_{n+1})}$$

Here $q(x, y)$ is the transition probability, $\mathbb{P}(X_{n+1} = y|X_n = x)$. If `rand`$\leq \alpha$,

$$X_{n+1} = U_{n+1}.$$

That is,

$$X_{n+1} = \mathbb{I}(u \leq \alpha)U_{n+1} + \mathbb{I}(u > \alpha)X_n$$

2.6 Visualization

A demonstration of the above methods has been implemented using the visualization library of [**?**][e]. The visualization dynamically follows the trace of a chosen method applied to a chosen two-dimensional distribution. Distributions of various qualitative properties are available. This can be found at the following URL:

`http://goatleaps.xyz/assets/ULA/ULA.html`

Figure 2: Screenshot from the visualization; tULAc applied to a Gaussian mixture distribution.

[e]With kind permission of Alex Rogozhnikov, `https://arogozhnikov.github.io/about/`.

3 Beyond Moments

While first and second moments give us some idea of the performance of our sampling algorithms, we ideally would like a fuller picture. In this section we compare the performance of algorithms using the total variation distance, L^2-Wasserstein distance and Kullback–Leibler divergence. Using these measures, we can compare the performance to theoretical upper bounds for `ULA`.

3.1 Statistical Distances

Let $\mathcal{B}(\mathbb{R}^d)$ denote the Borel σ-algebra on $\mathbb{R}^d$. Let P and Q be probability measures on the space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Then we define the total variation distance, Kullback–Leibler divergence and Wasserstein metric as follows:

Definition 3.1 (Total Variation)**.** The total variation distance between two probability measures P and Q on $(\Omega, \mathcal{F})$ is defined as

$$\|P - Q\|_{TV} = \sup_{A \in \mathcal{F}} |P(A) - Q(A)| .$$

Proposition 3.2. *If the set Ω is countable then this is equivalent to half the L^1 norm.*

$$\|P - Q\|_{TV} = \frac{1}{2} \|P - Q\|_1 = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|$$

Proof. Let $B = \{\omega : P(\omega) \geq Q(\omega)\}$ and let $A \in \mathcal{F}$ be any event. Then

$$P(A) - Q(A) \leq P(A \cap B) - Q(A \cap B) \leq P(B) - Q(B).$$

The first inequality holds since $P(\omega) - Q(\omega) < 0$ for any $\omega \in A \cap B^c$, and so the difference in probability cannot be greater if these elements are excluded. For the second inequality, we observe that including further elements of B cannot decrease the difference in probability. Similarly,

$$Q(A) - P(A) \leq Q(B^c) - P(B^c) = P(B) - Q(B)$$

Thus, setting $A = B$, we have that $|P(A) - Q(A)|$ is equal to the upper bound in the total variation distance. Hence,

$$\|P - Q\|_{TV} = \frac{1}{2} |P(B) - Q(B) + Q(B^c) - P(B^c)| = \frac{1}{2} \sum_{\omega \in \Omega} |P(x) - Q(x)|$$

■

POSSIBLY TAKE OUT KL DIVERGENCE??

Definition 3.3 (Kullback–Leibler Divergence)**.** Let P and Q be two probability measures on $(\Omega, \mathcal{F})$. If $P \ll Q$, the Kullback–Leibler divergence of P with respect to Q is defined as

$$KL(P|Q) = \int_\Omega \frac{dP}{dQ} log\left(\frac{dP}{dQ}\right) dQ.$$

LITTLE MORE INTUITION ON WASSERSTEIN & TRANSPORT??

Finally we consider the Wasserstein distance. If P and Q are probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we say that γ is a transport plan between two probability measures P and Q if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for any Borel set $A \subset \mathbb{R}^d$, $\gamma(A \times \mathbb{R}^d) = P(A)$ and $\gamma(\mathbb{R}^d \times A) = Q(A)$. We denote the set of all such transport plans by $\Pi(P, Q)$.

Definition 3.4 (Wasserstein distance)**.** For two probability measures, P and Q, the L^p-Wasserstein distance is given by

$$W_p(P, Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, d\gamma(x, y) \right)^{1/p}.$$

We will restrict our attention mainly to L^1-Wasserstein and L^2-Wasserstein distances. Due to practical impossibility of computing higher-dimensional Wasserstein distances, we also introduce a compuatationally more feasible variant, the Sliced Wasserstein distance. First proposed in [12] and further elaborated on, for example, in [**?**], the Sliced Wasserstein distance exploits the fact that the Wasserstein distance between 1-dimensional probability measures P, Q can be computed with an explicit formula $\left|F^{-1}(t) - G^{-1}(t)\right|^p dt$ where F and G are the CDFs of P and Q respectively [13].

Definition 3.5 (Sliced Wasserstein distance)**.** For two probability measures, P and Q, the L^p-Wasserstein distance is given by

$$SW_p(P,Q) = \left(\int_{\mathbb{S}^{d-1}} W_p^p \left(\mathcal{RI}_P(\cdot,\theta), \mathcal{RI}_Q(\cdot,\theta) \right) d\theta \right)^{\frac{1}{p}}$$

where $\mathbb{S}^{d-1}$ is the $(d-1)$-dimensional sphere and $\mathcal{RI}$ denotes the Inverse Radon transform. In the above references, it is also proved that SW_p is indeed a metric. The main reason why we can use the Sliced Wasserstein distance as an approximation to the Wasserstein distance is that these two metrics are equivalent[**?**].

3.2 Theoretical Non-asymptotic Error Bounds

Theoretical bounds on the total variation distance between the distribution of the n^{th} iterate of the unadjusted Langevin Algorithm were first provided in the case of a 'warm start' in [4].

Then [6], [7] improve and consider Wasserstein distance. These papers showed that $O(d/\epsilon)$ iterations are needed for precision level ϵ.

`ULA` [5] `tULA` [3] `HOLA` [16] `MALA` [1]

4 Comparison of methods

4.1 Implementation

- repo
- description
- docs

4.2 Results

- plots plots plots plots plots

4.3 Sampling approaches

4.4 Parameters

4.5 Estimating error

— first/second moment

— trace

— histogram

— KL div

— TV

— sliced W

— sliced W no histo

— KDE KL, TV, SW

4.5.1 The Curse of dimensionality

4.5.2 Comparing continuous and discrete distributions

4.5.3 Histogram vs. KDE

4.5.4 Sliced Wasserstein approximation

The Sliced Wasserstein distance, being defined via a multi-dimensional integral, cannot be computed exactly. Therefore, we resort to a simple Monte Carlo scheme where L samples $\{\theta_i\}$ are drawn uniformly from the $(d-1)$-dimensional sphere $\mathbb{S}^{d-1}$.

$$SW_p(P,Q) \approx \left(\frac{1}{L} \sum_{i=1}^{L} W_p^p \left(\mathcal{RI}_P(\cdot,\theta), \mathcal{RI}_Q(\cdot,\theta) \right) \right)^{\frac{1}{p}}$$

5 SGLD

Stochastic Gradient Langevin Dynamics
In this section, we will closely follow [**?**]

5.1 Introduction

Normally, samples in machine learning are of huge sample sizes, for which most MCMC algorithms are not designed to process. As a result of the computational cost, several new approaches were proposed recently, Stochastic Gradient Langevin Dynamics (SGLD) is a popular one. SGLD is based on the Langevin Monte Carlo (LMC) LMC a discretization of a continuous-time process, it requires to compute the gradient of the log-posterior at the current fit of the parameter and avoid the accept/reject step. SGLD use unbiased estimator of the gradient of log-posterior based on subsampling, suitable for samples of huge size.

5.2 Governing Equation

Recall the following equations: Langevin Stochastic Differential Equation (SDE):

$$d\theta_t = -\nabla U(\theta_t)d + \sqrt{2}dB_t$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Euler discretization of the Langevin SDE:

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

, where $h > 0$ is a constant step size and $(Z_k)_{k\geq 1}$ is a sequence of i.i.d standard d - dimensional Guassian vectors. To reduce the costs of the algorithms, we will switch to SGLD, for which we will replace ∇U with an unbiased estimate $\nabla U_0 + (\frac{N}{p})\sum_{i\in S} \nabla U_i$, where S is a minibatch of 1,...., N with replacement of size p. Our iterations were then updated as

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

Stochastic Gradient Descent (SGD) is characterised by the same recursion as SGLD without the Gaussian noise, (the last term):

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

5.3 Analysis in Wasserstein distance

5.4 Definitions and Notations in Markov chain theory

Recall the following definitions: $\mathcal{P}_2(\mathbb{R}^d)$ the set of probablity measures with finite second momet.
$\mathcal{B}(\mathbb{R}^d)$ the Borel σ - algebra of $\mathbb{R}^d$.
For $\lambda, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we define the Wasserstein distance by

$$W_2(\lambda,\nu) = \inf_{\xi\in\Pi(\lambda,\nu)} \left(\int_{\mathbb{R}\times\mathbb{R}} ||\theta - \vartheta)||^2 \xi(d\theta, d\vartheta) \right)^{\frac{1}{2}}$$

where, $\Pi(\lambda,\nu)$ is the set of probablity measures ξ on $\mathcal{B}(\mathbb{R}^d)\otimes\mathcal{B}(\mathbb{R}^d)$ satisfying for all $A \in \mathcal{B}(\mathbb{R}^d), \xi(A\times\mathbb{R}^d) = \lambda(A)$ and $\xi(\mathbb{R}^d \times A) = \nu(A)$.
For any probablity measure λ on $\mathcal{B}(\mathbb{R}^d)$, we define λR for all $A \in \mathcal{B}(\mathbb{R}^d)$ by $\lambda R(A) = \int_{\mathbb{R}^d} \lambda(d\theta)R(\theta, A)$.
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 $A \in \mathcal{B}(\mathbb{R}^d)$, $R^{k+1}(\theta, A) = \int_{\mathbb{R}^d} R^k(\theta, d\vartheta)R(\vartheta, A)$.

A probablity measure $\bar{\pi}$ is invariant for R if $\bar{\pi}R = \bar{\pi}$.
Our algorithms LMC, SGLD, SGD and SGLDFP algorithms are homogeneous Markov chains with Markov kernels denoted $R_{LMC}, R_{SGLD}, R_{SGD}$ and R_{FP}.

5.5 Results

For lemma 1, Theorem 2 and Corollary 3, we assume H1, H2 and H3.

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

$$W_2^2(R_{SGLD}^k(\theta, \cdot), \pi_{SGLD}) \leq (1 - mh)^k \int_{\mathbb{R}^d} ||\theta - \vartheta||^2 \pi_{SGLD}(d\vartheta)$$

same inequality holds for LMC, SGD and SGLDFP.

Theorem 5.1. *For all $h \in (0, \frac{1}{L}], \lambda, \nu \in \mathcal{P}_2(\mathbb{R}^d) and n \in \mathbb{N}$, we have the following upper- bounds in Wasserstein distance between*

i) LMC and SGLDFP,

$$W_2^2(\lambda R_{LMC}^n, \nu R_{FP}^n) \leq (1-mh)^n W_2^2(\lambda, \nu) + \frac{2L^2hd}{pm^2} + \frac{L^2h^2}{p} n(1-mh)^{n-1} \int_{\mathbb{R}^d} ||\vartheta - \theta*||^2 \mu(d\vartheta)\,, \quad (2)$$

ii) the Langevin diffusion and LMC,

$$\begin{aligned} W_2^2(\lambda R_{LMC}^n, \mu P_{nh}) \leq 2(1 - \frac{mLh}{m+L})^n W_2^2(\lambda, \mu) + dh\frac{m+L}{2m}(3 + \frac{L}{m})(\frac{13}{6} + \frac{L}{m}) \\ + ne^{-(\frac{m}{2})h(n-1)} L^3h^3(1 + \frac{m+L}{2m}) \int_{\mathbb{R}^d} ||\vartheta - \theta*||^2 \mu(d\vartheta)\,, \end{aligned} \quad (3)$$

iii) SGLD and SGD

$$W_2^2(\lambda R_{SGLD}^n, \mu R_{SGD}^n) \leq (1 - mh)^n W_2^2(\lambda, \mu) + \frac{(2d)}{m}. \quad (4)$$

Proof omitted.

Corollary 5.2. *Set $h - \frac{\eta}{N} with \eta \in (0, \frac{1}{(2L)}]$ and assume that $lim \inf_{N\to\infty} mN^{-1} > 0$. Then*

i) for all $n \in N$, we get $W_2(R_{LMC}^n(\theta*, \cdot), R_{FP}^n(\theta*, \cdot)) = \sqrt{d\eta}\mathcal{O}(N^{-\frac{1}{2}})$ and $W_2(\pi_{LMC}, \pi_{FP}) = \sqrt{d\eta}\mathcal{O}(N^{-\frac{1}{2}})$.

ii) for all $n \in \mathbb{N}, we get W_2(R_{SGLD}^n(\theta*, \cdot), R_{SGD}^n(\theta*, \cdot)) = \sqrt{d}\mathcal{O}(N^{-\frac{1}{2}})$, and $W_2(\pi_{SGLD}, \pi_{SGD}) = \sqrt{d}\mathcal{O}(N^{-\frac{1}{2}})$.

References

[1] Nawaf Bou-Rabee and Martin Hairer. Nonasymptotic mixing of the mala algorithm. *IMA Journal of Numerical Analysis*, 33(1):80–110, 2013.

[2] Nawaf Bou-Rabee and Eric Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for sdes. *Communications on Pure and Applied Mathematics*, 63(5):655–696, 2010.

[3] Nicolas Brosse, Alain Durmus, Eric Moulines, and Sotirios Sabanis. The tamed unadjusted langevin algorithm. *Stochastic Processes and their Applications*, 2018.

[4] Arnak S 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.

[5] Arnak S Dalalyan and Avetik Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 2019.

[6] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. *arXiv preprint arXiv:1605.01559*, 2016.

[7] Alain Durmus and Eric Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017.

[8] W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970.

[9] Don S. Lemons and Anthony Gythiel. Paul Langevins 1908 paper On the Theory of Brownian Motion [Sur la thorie du mouvement brownien, C. R. Acad. Sci. (Paris) 146, 530533 (1908)]. *American Journal of Physics*, 65(11):1079–1081, 1997.

[10] Charles Matthews and Benedict Leimkuhler. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 06 2012.

[11] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21:1087–1092, 6 1953.

[12] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[13] Aaditya Ramdas, Nicolás Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[14] Gareth O. Roberts and Richard L. Tweedie. Exponential Convergence of Langevin Distributions and Their Discrete Approximations. *Bernoulli*, 2(4):341–363, 1996.

[15] Sotirios Sabanis. A note on tamed euler approximations. *Electron. Commun. Probab.*, 18:10 pp., 2013.

[16] Sotirios Sabanis and Ying Zhang. Higher Order Langevin Monte Carlo Algorithm. Workingpaper, ArXiv, 8 2018.