Langevin MC,
or the curse of dimensionality

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

March 9, 2019

1 Motivation

Monte Carlo methods are a class of algorithms that use repeated random sampling to calculate numerical approximations of integrals.

In this report we consider the problem of sampling from a distribution of the form

$$\pi(x) \propto \exp(-U(x))$$

for some potential function U. In statistical mechanics, this distributions is called a Gibbs distribution, and is often used to model the postions and velocities of particles in a gas.

2 Langevin Monte Carlo Algorithms

The Langevin equation is a stochastic differential equation (SDE) originally developed to model the movement of a Brownian particle [3]. 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 section/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

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

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

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 [6].

Theorem 2.1 (Theorem 2.1, [6]). *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 $N, a, b < \infty$,*

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

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

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

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

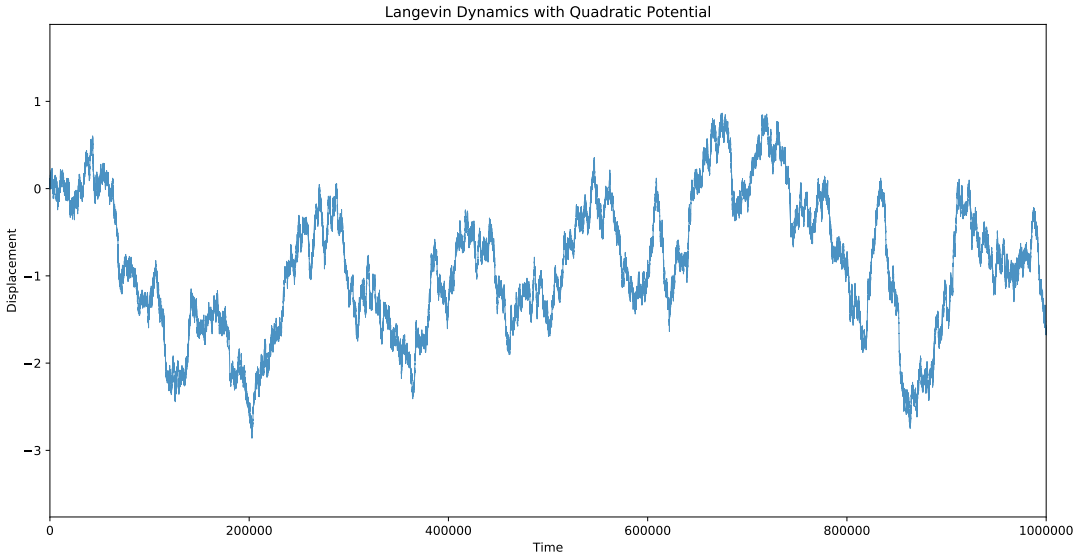

Figure 1: Simulating Langevin dynamics in one dimension with a quadratic potential $U(x) = x^2/2$

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 the masure 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 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 [6]. 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.

3 MALA

Before describing the Metropolis-adjusted Langevin algorithm `MALA`, it is pertinent at this point to recall the random walk Metropolis-Hastings algorithm `RWM` [2, 5]. 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

$$\alpha(X_n, Y_{n+1}) = 1 \wedge \frac{\pi(Y_{n+1})q(Y_{n+1}, X_n)}{\pi(X_n)q(X_n, Y_{n+1})}.$$[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 [6]. 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 3.1 (Theroem 4.2, [6]). *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.

[b] I_d denotes the $d \times d$ identity matrix.
[c] Here $t \wedge s = \min\{t, s\}$.

4 Taming the Gradient

We have seen that both `ULA` and `MALA` run into issues when the gradient of the potential is superlinear. Taming the drift coefficient is a method to reduce the superlinearity whilst maintaining the invariant distribution of the SDE [1, 6, 7]. Recall that the Langevin dynamics are governed by

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

We have seen that problems arise when the drift coefficient is superlinear, that is

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

For this reason a family of drift functions $(G_h)_{h>0},\ G_h : \mathbb{R}^d \to \mathbb{R}^d$ and the Markov chain

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

Here we consider three choices of G_h, each of which leads to a convergent scheme.

$$T_h(x) = \frac{\nabla U(x)}{1 + h\|\nabla(x)\|}, \qquad\qquad T_h^{\mathrm{RT}} = \frac{D\nabla U(x)}{2(D \vee \|\nabla U(x)\|)}, \quad D > 0$$

We name T_h the taming function from [1] and T_h^{RT} the RT-taming after [6] where it was introduced.

4.1 tULA/c

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

where $T_h(x) = \frac{\nabla U(x)}{1+\|\nabla U(x)\|}$ or $T_h(x) = \left(\frac{\nabla U(x)}{1+|\partial_i U(x)|}\right)_{i=\{1,\dots,d\}}$

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

+++Show ill condition/stiff bad behaviour, how coordinatewise fixes+++

4.2 tMALA/c

Use the same taming T as in `tULA`. Is this sensible? Could compare with `MALTA`.

4.3 MALTA?

[6]

Tame with

$$T = \frac{\nabla U(x)}{1 \vee h|\nabla U(x)|}$$

for some constant $D > 0$

5 Discretise Differently

5.1 tHOLA

[8] 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 $\sigma_h(x) = \mathrm{diag}\left(\left(\sigma_h^{(k)}(x)\right)_{k\in\{1,\dots,d\}}\right)$ with,

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

5.2 LM

[4] Non-Markovian scheme,

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

6 Other Methods

6.1 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,

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

7 Beyond Moments

References

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

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

[3] 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.

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

[5] 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.

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

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

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