Penalized Reflected Stochastic Gradient Langevin Dynamics for Deep Learning

Amine Boukardagha [1] Advised by Ludovic Tangpi [2]

[1]Columbia University [2]Princeton University

Introduction

Stochastic gradient Langevin dynamics (SGLD) [6] is a widely used Markov Chain Monte Carlo (MCMC) method in deep learning. The sampler smoothly transitions from stochastic optimization to sampling as the injected noise enables exploration for posterior sampling and uses the Langevin diffusion. However, in the case of constrained optimization problems, the Langevin diffusion without adjustments fail. One way to adjust the Langevin diffusion for constrained cases is to consider reflected stochastic differential equations like in [1]. Additionally, reflected stochastic differential equations can be simulated numerically via projected Euler methods and are used in Bayesian Learning and stochastic adaptative control, see [13], [8], [2], [14]. Our goal is to develop an unajusted penalization-based Reflected Langevin Monte Carlo Algorithm (pRLMC) for deep learning using properties of reflected stochastic differential equations and their penalizations.

Stochastic Gradient Langevin Dynamics

Stochastic Gradient Langevin Dynamics stems from the Langevin Diffusion which obeys the following stochastic differential equation

$$dX_t = -\nabla f(X_t)dt + \sigma dW_t \tag{1}$$

where $X_t \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$ is the energy function, W_t a d-dimensional Weiner process, and $\sigma^2/2$ the temperature. Let $\langle , \rangle$ and $|| \cdot ||$, respectively, denote the Euclidean inner product and the corresponding L^2 norm on $\mathbb{R}^d$. If we assume that for some finite a, b, and N we have $-\langle \nabla f(X_t), X_t \rangle \leq a||X_t||^2 + b$, $||X_t|| > N$, then according to Theorem 2.1 in [9] the solution X_t of the diffusion equation (1) converges to the unique invariant Gibbs distribution:

$$d\pi(x) = \frac{1}{Z} exp(-\frac{f(x)}{\sigma^2/2})dx \tag{2}$$

such that the normalization constant Z is the following

$$Z = \int_D exp(-\frac{f(x)}{\sigma^2/2})\, dx \tag{3}$$

Applying the Euler-Maruyama scheme to (1), we get the Langevin Monte Carlo (LMC)

$$X_{k+1} = X_k - \eta \nabla f(X_k) + \sigma\sqrt{\eta}\epsilon_k \tag{4}$$

where η is the step size and $\epsilon \sim \mathcal{N}(0, I_{d\times d})$ such that $I_{d\times d}$ is the $d \times d$ unit matrix.

Reflected Stochastic Gradient Langevin Dynamics

For constrained convex stochastic optimization, instead of using the standard Langevin diffusion, we consider the Reflected Stochastic Differential Equation (RSDE) of the form:

$$X_t = X_0 + \int_0^t \sigma dW_s - \int_0^t \nabla g(X_s)ds + K_t \tag{5}$$

with reflecting boundary condition on closed convex domain D. Here, $X_0 \in D$, X is a reflecting process on D, K is a bounded variation process with variation $||K||$ increasing only, when $X_t \in \partial D$, W is a d-dimentional standard Weiner process, $\sigma \in \mathbb{R}$, and $g : \mathbb{R}^d \to \mathbb{R}$ differentiable.
Consider the penalized Stochastic Differential Equation

$$X_t^n = X_0 + \int_0^t \sigma dW_s^n - \int_0^t \nabla g(X_s^n)ds + K_t^n \tag{6}$$

where $\Pi(x)$ is the metric projection of x onto the convex body D which is the point of ∂D where the minimum distance from x to points from D is attained. Additionally, K_t^n is the following

$$K_t^n = -n\int_0^t X_s^n - \Pi(X_s^n)\, ds \tag{7}$$

Assume that the following conditions are satisfied for convex D and for some $R > 0$ and $L > 0$:

$$\sigma^2 + ||\nabla g(x)||^2 \leq R(1 + ||x||^2) \tag{8}$$

and

$$||\nabla g(x) - \nabla g(y)||^2 \leq L||x - y||^2 \tag{9}$$

Then according to Theorem 4.2 of [1], for X_t strong solution of (5) there exists some $C > 0$ such that:

$$\mathbb{E}[sup||X_t^n - X_t||^2]^{\frac{1}{2}} \leq C(\frac{ln(n)}{n})^{\frac{1}{4}} \tag{10}$$

Theorem 1. The penalized SDE (6) could be rewritten in the following differential form

$$dX_t^n = -\nabla f_n(X_t^n)dt + \sigma dW_t^n \tag{11}$$

such that the penalized energy function $f_n : D \to \mathbb{R}$ is defined as follows:

$$f_n(x) = g(x) + \frac{n}{2}dist^2(x, \partial D)$$

Theorem 2. If D is convex domain and g is strongly convex then the penalized energy function f_n is also strongly convex for $x \in D$ and $n \in \mathbb{N}$.

Wasserstein Contraction of the Penalized Reflected Stochsatic Differential Equation

According to [11], for all μ, ν probability measures on $\mathbb{R}^d$, the Wasserstein distance of order 2 is:

$$W_2(\mu, \nu) = \inf_{\pi\in\Omega(\mu,\nu)} \mathbb{E}_{(X_t,Y_t)\sim\pi}[||X_t - Y_t||^2]^{\frac{1}{2}} \tag{12}$$

Such that μ and ν respectively are the probability laws of X_t and Y_t and where $\Omega(\mu, \nu)$ is the set of all couplings between μ and ν.

Theorem 3. Let μ_0 and ν_0 be two measures in $\mathbb{R}^d$ and $X_t^n \in D$ and $Y_t^n \in D$ the solutions of (6) starting from X_0 and Y_0 of respective laws μ_0 and ν_0, both driven by the same Weiner process. If we assume that D is convex domain and g strongly convex, then we have the following contraction result:

$$W_2(\mu_t^n, \nu_t^n) \leq e^{-Ct}W_2(\mu_0, \nu_0) \tag{13}$$

Corollary 1. Assume that the conditions of Theorem 3 are satisfied and for finite constants a,b,N we have the following condition on g for $||X_t^n|| > N$:

$$-\langle \nabla g(X_t^n), X_t^n \rangle \leq (a + n)||X_t^n||^2 - n\langle X_t^n, \Pi(X_t^n)\rangle + b \tag{14}$$

Then, the solution of (6) converges to the unique invariant Gibbs distribution dependent on the penalization term n

$$d\pi(x) = \frac{1}{Z_n} exp(-\frac{g(x) + \frac{n}{2}dist^2(x, \partial D)}{\sigma^2/2})dx \tag{15}$$

such that the penalized normalization constant Z_n is the following

$$Z_n = \int_D exp(-\frac{g(x) + \frac{n}{2}dist^2(x, \partial D)}{\sigma^2/2})\, dx \tag{16}$$

and we have the following contraction result:

$$W_2(\mu_t^n, \pi^n) \leq e^{-Ct}W_2(\mu_0, \pi^n) \tag{17}$$

Wasserstein Contraction of Reflected Stochastic Gradient Langevin Dynamics

Wasserstein Contraction of Reflected Stochastic Gradient Langevin Dynamics In this portion, we investigated the Wasserstein contraction properties of the reflected stochastic differential equation as well as the convergence to a unique invariant distribution.

Theorem 4. Let $X^n \in D$ and $Y^n \in D$ satisfy (6), $n \in \mathbb{N}$ and let $X_t \in D$ and $Y_t \in D$ solutions of (5) with respective laws μ_t and ν_t. Assume that the conditions (8) and (9) are satisfied. Then, we have the following contraction result for the laws of X_t and Y_t:

$$W_2(\mu_t, \nu_t) \leq e^{-Ct}W_2(\mu_0, \nu_0) + K(\frac{ln(n)}{n})^{\frac{1}{4}} \tag{18}$$

Corollary 2. Let $X_t \in D$ and $Y_t \in D$ solutions of (5) with respective laws μ_t and ν_t. Assume that the conditions (8) and (9) are satisfied. Then, we have the following contraction result for the laws of X_t and Y_t:

$$W_2(\mu_t, \nu_t) \leq e^{-Ct}W_2(\mu_0, \nu_0) \tag{19}$$

Corollary 3. Under the generalized one-sided Lipschitz condition, the geometric drift condition, and growth condition of Corollary 1 of [8] and given the contraction result of Corollary 2, there exists a unique stationary distribution $\pi \in \mathcal{P}_V$, such that $\int_D V(x)\, d\pi(x) < \infty$. Let X_t be a solution to (5) with law μ_t such that $\int_D V(x)\, d\mu_0(x) < \infty$. If $V(x) \geq 1$ for all $x \in D$, then μ_t has the following contraction property:

$$W_2(\mu_t, \pi) \leq \chi e^{-Ct}W_2(\mu_o, \pi) \tag{20}$$

such that $\chi = \frac{1}{2}diam(D)\xi^{-1}$ and $\xi^{-1} > 0$.

Theorem 5. Assuming the generalized one-sided Lipschitz condition, the geometric drift condition, and growth condition of Corollary 1 of [8], we have the following convergence result for the invariant measure and its penalized counterpart for some $C > 0$:

$$W_2(\pi, \pi^n) \leq C(\frac{ln(n)}{n})^{\frac{1}{4}} \tag{21}$$

Result

Applying the Euler-Maruyama scheme to (11), we get the Penalized Reflected Langevin Monte Carlo (pRLMC)

$$X_{k+1}^n = X_k^n - \eta\nabla f_n(X_k^n) + \sigma\sqrt{\eta}\epsilon_k^n \tag{22}$$

where $k \in \{0, ..., T\}$, η is the step size, penalization number n is sufficiently large, $\epsilon_k^n \sim \mathcal{N}(0, I_{d\times d})$ such that $I_{d\times d}$ is the $d \times d$ unit matrix, and

$$f_n(X_k^n) = g(X_k^n) + \frac{n}{2}dist^2(X_k^n, \partial D)$$

.

Unadjusted pRLMC Algorithm

Input: starting guess X_0, step size $\eta > 0$, penalization number n, volatility σ, number of epochs T, convex set D
Output: $X_0^n, ..., X_T^n$
for $t = 0$ to T **do**
compute $\nabla f_n(X_t^n) = \nabla[g(X_t^n) + \frac{n}{2}dist^2(X_t^n, \partial D)]$
sample $\epsilon_t^n \sim \mathcal{N}(0, I_{d\times d})$
compute $X_{t+1}^n = X_t^n - \eta\nabla f_n(X_t^n) + \sigma\sqrt{\eta}\epsilon_t^n$
end for
return $X_0^n, ..., X_T^n$

References

[1] L. Słomiński, *Weak and strong approximations of reflected diffusions via penalization methods*, Stochastic Processes and their Applications, Volume 123, Issue 3, 2013, Pages 752-763.

[2] L. Słomiński, *Euler's approximations of solutions of sdes with reflecting boundary*, Stochastic processes and their applications, vol. 94, no. 2, pp. 317-337, 2001.

[3] F. Bolley, I. Gentil, and A. Guillin, *Convergence to equilibrium in Wasserstein distance for Fokker–Planck equations*, Journal of Functional Analysis, Volume 263, Issue 8, 2012, Pages 2430-2457.

[4] V. Balestro, H. Martini, and R. Teixeira, *Convex analysis in normed spaces and metric projections onto convex bodies*, arXiv, p7 2019.

[5] J. Bouvrie and J. Slotine, *Wasserstein Contraction of Stochastic Nonlinear Systems*, arXiv: Optimization and Control, 2019.

[6] M. Welling and Y. Teh, *Bayesian Learning via Stochastic Gradient Langevin Dynamics*,In Proc. of the International Conference on Machine Learning (ICML), pp. 681–688, 2011.

[7] A. Eberle, A. Guillin, and R. Zimmer, *Quantitative harris-type theorems for diffusions and mckean– vlasov processes*, Transactions of the American Mathematical Society, vol. 371, no. 10, pp. 7135– 7173, 2019.

[8] T. Lekang and A. Lamperski, *Wasserstein Contraction Bounds on Closed Convex Domains with Applications to Stochastic Adaptive Control*, 2021 60th IEEE Conference on Decision and Control (CDC), Austin, TX, USA, 2021, pp. 366-371.

[9] G. Roberts and R. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, pages 341–363, 1996.

[10] M. Chae and S. Walker, *Wasserstein upper bounds of the total variation for smooth densities*, Statistics and Probability Letters, Volume 163, 2020.

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

[12] A. Domokos, J. Ingram, and M. Marsh, *Projections onto closed convex sets in Hilbert spaces*, Acta Math. Hungar. 152, 114–129, 2017. https://doi.org/10.1007/s10474-017-0691-9

[13] R. Laumont, V. De Bortoli, A Almans, J. Delon, A. Durmus, and M. Pereyra, *Bayesian Imaging Using Plug and Play Priors: When Langevin Meets Tweedie*, SIAM Journal on Imaging Sciences, 2022.

[14] H. Tanaka, *Stochastic differential equations with reflecting boundary condition in convex regions*, Hiroshima Mathematical Journal, vol. 9, no. 1, pp. 163-177, 1979.

mab2505@columbia.edu