Underdamped Langevin MCMC

Yuze Zhou

Apr.29th

Underdamped Langevin Diffusion Process

- Underdamped Langevin Diffusion Process:

$$dv_t = -\gamma v_t dt - u\nabla(f(x_t))dt + (\sqrt{2\gamma u})dB_t$$
$$dx_t = v_t dt$$

- $(x_t, v_t) \in \mathbb{R}^{2d}$ and f is twice continuously-differentiable
- Unique invariant distribution proportional to

$$exp(-(f(x) + \|v\|_2^2/2u))$$

- More specifically, for the marginal distribution of x from the invariant distribution would be proportional to:

$$exp(-f(x))$$

Assumptions

- **Lipschitz Gradients**

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

- **m-strongly convex**

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{m}{2}\|x - y\|_2^2$$

- Notations for $f(x)$:

$$\kappa = L/m$$
$$x^* = \arg\min f(x)$$

Notations

- Let μ and υ be two probability measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$
- $\Gamma(\mu, \upsilon)$ is the set of all couplings of measure μ and υ
- **Wasserstein Distance**

$$W_2(\mu, \upsilon) = (\inf_{\zeta \in \Gamma(\mu,\upsilon)} \int \|x - y\|_2^2 d\zeta(x, y))^{1/2}$$

- $\Gamma_{opt}(\mu, \upsilon)$ is the optimal coupling that achieves the infimum of the Wasserstein distance

Notations

- For the Langevin Diffusion Process defined by the SDE before, with initial condition $(x_0, v_0) \sim p_0$ for some distribution p_0 on $\mathbb{R}^{2d}$, let p_t be the distribution for (x_t, v_t) and Φ_t be the operator that maps p_0 to p_t

$$\Phi_t p_0 = p_t$$

Previous Work

- **Durmus & Moulines 2016** For $f(x)$ with same properties, Unadjusted Langevin MCMC:

$$x_{k+1} = x_k - \delta \nabla f(x_k) + \sqrt{2\delta} Z_{k+1}$$

- Converge in W2 with $\mathcal{O}(d\kappa^2/\epsilon^2)$ iterations

Underdamped Langevin MCMC

- **Discretized Underdamped Langevin Process**
- Starting from $(\tilde{x}_0, \tilde{v}_0) \in \mathbb{R}^{2d}$, the discretized Underdamped Langevin Process will evolve according to the following SDE:

$$d\tilde{v}_t = -\gamma\tilde{v}_t - u\nabla f(\tilde{x}_0)dt + (\sqrt{2\gamma u})dB_t$$
$$d\tilde{x}_t = \tilde{v}_t dt$$

- Denote $Z^\delta(\tilde{x}_0, \tilde{v}_0)$ as the distribution of $(\tilde{x}_\delta, \tilde{v}_\delta)$ evolving according to the discretized process starting from $(\tilde{x}_0, \tilde{v}_0)$
- The discretized version of Underdamped Langevin Process has an explicit form of solution for $(\tilde{x}_t, \tilde{v}_t)$ and easy to implement
- Parallelly, for the discretized SDE with initial condition $(\tilde{x}_0, \tilde{v}_0) \sim p_0$ for some distribution $\tilde{p}_0$ on $\mathbb{R}^{2d}$, let $\tilde{p}_t$ be the distribution for $(\tilde{x}_t, \tilde{v}_t)$ and $\tilde{\Phi}_t$ be the operator that maps $\tilde{p}_0$ to $\tilde{p}_t$

$$\tilde{\Phi}_t p_0 = \tilde{p}_t$$

The algorithm for underdamped Langevin MCMC now goes as followed

Algorithm 1: Underdamped Langevin MCMC

Initialisation: Choose step-size δ, iteration times n and starting point (x^0, v^0);

for $0 \leq i \leq n-1$ **do**

  Sample $(x^{i+1}, v^{i+1}) \sim Z^{\delta}(x^i, v^i)$

end

Convergence

- For the following analysis on convergence, denote by p^* the unique invariant distribution of the Underdamped Langevin Process such that
$$p^* \propto exp(-(f(x) + \frac{1}{2u}\|v\|_2^2))$$
- Let $g(x, v) = (x, x + v)$, and q_t be the distribution of $g(x_t, x_t + v_t)$, q^* be the distribution of $g(x, v)$ when $(x, v) \sim p^*$
- In the following analysis, we would set $u = 1/L$ and $\gamma = 2$

Convergence

Theorem (Cheng et.al. 2018)

Let (x_0, v_0) and (y_0, w_0) be two arbitrary points in $\mathbb{R}^{2d}$. Let p_0 and p_0' be two dirac measures concentrated on (x_0, v_0) and (y_0, w_0). If we set $u = 1/L$ and $\gamma = 2$, then for every $t > 0$, there exists a $\zeta_t(x_0, v_0, y_0, w_0) \in \Gamma(\Phi_t p_0, \Phi_t p_0')$ such that:

$$\mathbb{E}_{(x_t,v_t,y_t,w_t)\sim\zeta_t(x_0,v_0,y_0,w_0)}[\|x_t - y_t\|_2^2 + \|(x_t + v_t) - (y_t + w_t)\|_2^2] \leq e^{-t/\kappa}[\|x_0 - y_0\|_2^2 + \|(x_0 + v_0) - (y_0 + w_0)\|_2^2]$$

- From the theorem above, let $(x_0, v_0) \sim p_0$ and $g(x_0, v_0) \sim q_0$, we could easily get:

$$W_2(\Phi_t q_0, q^*) \leq e^{-t/2\kappa} W_2(q_0, q^*)$$
$$W_2(\Phi_t p_0, p^*) \leq 4e^{-t/2\kappa} W_2(p_0, p^*)$$

Convergence

- The previous theorem only guarantees the convergence of the continuous process. However, for the discretized version, more assumptions are needed.
- Let δ represents a single step of the Underdamped Langevin MCMC algorithm, p_t be the probability distribution of (x_t, v_t) from the continuous process and $\tilde{p}_t$ be the distribution from the discretized version. Recall the definitions of $\tilde{\Phi}_t$ and Φ_t aforementioned.
- **Assumptions:** For the continuous time process, there exists $\mathcal{E}_\kappa$ such that:

$$\forall t \in [0, \delta] \qquad \mathbb{E}_{p_t}[\|v\|_2^2] \leq \mathcal{E}_\kappa$$

- $\mathcal{E}_k$ could be explicitly bounded by function of paramters m, L and d

Convergence

The following theorem bounds the distance between the continuous process and the discretized process with one step.

Theorem (Cheng et.al. 2018)

Let $\tilde{\Phi}_t$ and Φ_t be the probability transfer operator as aforementioned. Let p_0 be any arbitrary distributions and the step-size $\delta < 1$. If we choose $u = 1/L$ and $\gamma = 2$, the Wasserstein distance between the continuous process and the discretized process is upper bounded by:

$$W_2(\tilde{\Phi}_\delta p_0, \Phi_\delta p_0) \leq \delta^2 \sqrt{\frac{2\mathcal{E}_\kappa}{5}}$$

Convergence

Theorem (Cheng et.al. 2018)

Let $p^{(n)}$ be the distribution of the Underdamped Langevin MCMC algorithm after n steps starting from initial distribution $p^{(0)} = \mathbf{1}_{\{x=x^{(0)},v=0\}}$, and the initial distribution satisfies $\|x^{(0)} - x^*\| \leq D^2$. If we set the step-size to be:

$$\delta = \frac{\epsilon}{104\kappa}\sqrt{\frac{1}{d/m+D^2}}$$

and run the algorithm for:

$$n \geq (\frac{52\kappa^2}{\epsilon}) \cdot (\sqrt{\frac{d}{m} + D^2}) \cdot log(\frac{24(\frac{d}{m}+R^2)}{\epsilon})$$

we shall have the guarantee that:

$$W_2(p^{(n)}, p^*) \leq \epsilon$$

- converge with $\mathcal{O}(\sqrt{d}\kappa^2/\epsilon)$ iterations

Comparisons to traditional MCMC

- Strong assumptions for the invariant distribution proportional to $exp(-f(x))$
- Converge in Wasserstein Distance rather than total variation distance

References I

- Cheng.X, Chatterji.N.S, Barlett.P, Jordan.M.I(2018)
  Underdamped Langevin MCMC, a non-asymptotic analysis
- Durmus.A, Moulines.E(2016)
  Sampling from a strongly log-concave distribution with the Unadjusted Langevin Algorithm
- Pavliotis.G.A(2014)
  Stochastic Process and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations