Convergence of the unadjusted Langevin algorithm

Simon Xiang

University of Texas at Austin

July 15, 2023

Motivation

Question

For $f : \mathbb{R}^n \to \mathbb{R}$, assuming we have oracle access to ∇f, how would we sample from the probability distribution $\nu \sim e^{-f}$?

- As the step size $\eta \to 0$, the **discretized unadjusted Langevin algorithm** recovers **Langevin dynamics**, a continuous time stochastic process converging to ν.
- If the **log-Sobolev inequality** is satisfied, this algorithm converges at an exponential rate, or $\mathcal{O}(\log n)$ time.

KL Divergence and the log-Sobolev inequality

Let ρ, ν be probability distributions on $\mathbb{R}^n$.

Definition

The **Kullback-Liebler (KL) divergence** of ρ with respect to ν is defined by

$$H_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx.$$

Definition

For all smooth $g \colon \mathbb{R}^n \to \mathbb{R}$ and $\alpha < 0$ with $\mathbb{E}_\nu[g^2] < \infty$, the **log-Sobolev inequality (LSI)** is given by

$$\mathbb{E}_\nu[g^2 \log g^2] - \mathbb{E}_\nu[g^2] \log \mathbb{E}_\nu[g^2] \leq \frac{2}{\alpha} \mathbb{E}_\nu[\|\nabla g\|^2].$$

Equivalence

Definition

The *relative Fisher information* is given by

$$J_\nu(\rho) = \int_{\mathbb{R}^n} \rho(x) \left\| \nabla \log \frac{\rho(x)}{\nu(x)} \right\|^2 dx.$$

Proposition

The log-Sobolev inequality is equivalent to the following relation between KL divergence and Fisher information for all ρ:

$$H_\nu(\rho) \leq \frac{1}{2\alpha} J_\nu(\rho).$$

To obtain this inequality from LSI, choose $g^2 = \frac{\rho}{\nu}$. To obtain LSI from this inequality, choose $\rho = \frac{g^2 \nu}{\mathbb{E}_\nu[g^2]}$.

Langevin dynamics and the Fokker-Planck equation

Definition

For a target distribution $\nu \sim e^{-f}$, the **Langevin dynamics** is a continuous time stochastic process $(X_t)_{t\geq 0}$ in $\mathbb{R}^n$ that evolves by the following SDE:

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t$$

where $(W_t)_{t\geq 0}$ is the n-dimensional Brownian motion.

Definition

If $(X_t)_{t\geq 0}$ evolves following the Langevin dynamics, then its pdf $(\rho_t)_{t\geq 0}$ evolves by the **Fokker-Planck equation**

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu}\right)$$

where $\nabla\cdot$ is divergence and Δ is the Laplacian.

Convergence

Lemma

Along the Langevin dynamics (or Fokker-Planck equation),

$$\frac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t).$$

Proof.

Recall that the time derivative of KL divergence along any flow is

$$\frac{d}{dt} H_\nu(\rho_t) = \frac{d}{dt} \int_{\mathbb{R}^n} \rho_t \log \frac{\rho_t}{\nu}\, dx = \int_{\mathbb{R}^n} \frac{\partial \rho_t}{\partial t} \log \frac{\rho_t}{\nu}\, dx$$

as the second part from the chain rule vanishes. Then along the Fokker-Planck equation this integrates by parts to $-J_\nu(\rho_t)$. ☒

Since $J_\nu(\rho) \geq 0$, the KL divergence with respect to ν is decreasing and $\rho_t \to \nu$.

Rate of convergence

Theorem

Suppose ν satisfies LSI with $\alpha > 0$. Then along the Langevin dynamics,

$$H_\nu(\rho_t) \le e^{-2\alpha t} H_\nu(\rho_0).$$

Proof.

Applying our lemma and the fact that LSI is equivalent to $H_\nu(\rho) \le \frac{1}{2\alpha} J_\nu(\rho)$, we have

$$\frac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t) \le -2\alpha H_\nu(\rho_t).$$

Integrating yields the bound $H_\nu(\rho_t) \le e^{-2\alpha t} H_\nu(\rho_0)$ as desired. ☒

The discretized unadjusted Langevin algorithm

Definition

The **Unadjusted Langevin Algorithm (ULA)** with step size $\eta > 0$ is the discrete-time algorithm

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

where $z_k \sim N(0, 1)$ is an independent Gaussian random variable in $\mathbb{R}^n$. Let ρ_k denote the probability distribution of x_k evolving along ULA.

- Comparing to $dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t$, as $\eta \to 0$ we see how ULA recovers the Langevin dynamics.
- For fixed $\eta > 0$, ULA converges to a biased limiting distribution $\nu_\eta \neq \nu$. So KL divergence sadly does not converge to 0 along ULA, and has an asymptotic bias $H_\nu(\nu_\eta) > 0$.

Conclusion

Thank you for listening! Reference:
`https://arxiv.org/abs/1903.08568`