CPSC 661: Sampling Algorithms in ML

Andre Wibisono

April 14, 2021

Yale University

Last time

- Wasserstein W_2 metric, Otto calculus
- Langevin dynamics as gradient flow of relative entropy
- SLC $\Rightarrow$ LSI $\Rightarrow$ PI
- Exponential convergence rates of Langevin dynamics

Today: Unadjusted Langevin Algorithm

References

- Dalalyan, *Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent*, COLT 2017
- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018
- Vempala & Wibisono, *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, NeurIPS 2019
- Durmus, Majewski, & Miasojedow, *Analysis of Langevin Monte Carlo via Convex Optimization*, JMLR, 2019

Recap: Langevin dynamics

Langevin dynamics

Want to sample from target distribution $\nu \propto e^{-f}$ on $\mathbb{R}^n$

Relative entropy:

$$H_\nu(\rho) = \mathbb{E}_\rho\left[\log \frac{\rho}{\nu}\right] = \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\nu(x)}\, dx$$

Gradient flow (wrt W_2 metric) is the **Fokker-Planck equation**:

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

Implemented in $\mathbb{R}^n$ by the **Langevin dynamics**:

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

Properties of ν

1. α-**strongly log-concave** if $f = -\log \nu$ is α-strongly convex

$$\nabla^2 f(x) \succeq \alpha I$$

$\Downarrow$

2. α-**log-Sobolev inequality** if

$$J_\nu(\rho) \geq 2\alpha H_\nu(\rho)$$

$\Downarrow$

3. α-**Poincaré inequality** if

$$\mathbb{E}_\nu\left[\left\|\nabla \frac{\rho}{\nu}\right\|^2\right] \geq \alpha \chi^2_\nu(\rho)$$

$h = \frac{\rho}{\nu}$

$\Leftrightarrow \forall\, h : \mathbb{R}^n \to \mathbb{R}$

$\mathbb{E}_\nu [\|\nabla h\|^2] \geq \alpha \cdot \mathrm{Var}_\nu(h)$

Convergence of Langevin dynamics

Target distribution

$$\nu \propto e^{-f}$$

Strong log-concavity:

$$\nabla^2 f(x) \succeq \alpha I$$

Log-Sobolev inequality:

$$J_\nu(\rho) \geq 2\alpha\, H_\nu(\rho)$$

Poincaré inequality:

$$\mathbb{E}_\nu\left[\left\|\nabla \frac{\rho}{\nu}\right\|^2\right] \geq \alpha \chi^2_\nu(\rho)$$

Langevin dynamics

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

$\Longrightarrow$ Exponential contraction:

$$W_2(\rho_t, \tilde{\rho}_t)^2 \leq e^{-2\alpha t} W_2(\rho_0, \tilde{\rho}_0)^2$$

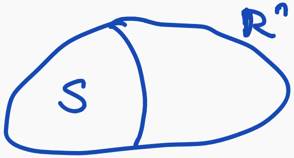
$\Longrightarrow$ Convergence in H_ν:

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

$\Longrightarrow$ Convergence in χ^2_ν:

$$\chi^2_\nu(\rho_t) \leq e^{-2\alpha t} \chi^2_\nu(\rho_0)$$

Isoperimetry



- LSI is equivalent to log-isoperimetry:

$$\psi = \inf_{S \subset \mathbb{R}^n} \frac{\nu(\partial S)}{\min\{\nu(S), \nu(S^c)\} \sqrt{\log \frac{1}{\min\{\nu(S), \nu(S^c)\}}}}$$

ν α-LSI $\Rightarrow \psi = \Omega(\sqrt{\alpha})$

- Poincaré inequality is equivalent to isoperimetry:

$$\phi = \inf_{S \subset \mathbb{R}^n} \frac{\nu(\partial S)}{\min\{\nu(S), \nu(S^c)\}}$$

ν α-PI $\Rightarrow \phi = \Omega(\sqrt{\alpha})$

  - c.f. Cheeger's inequality

LSI and PI beyond log-concave

- LSI and PI stable under bounded perturbation, Lipschitz mapping

  [Holley & Stroock, *Logarithmic Sobolev inequalities and stochastic Ising models*, J. Statist. Phys., 1987]

- LSI and PI of mixture distributions via decomposition into metastable regions

  [Menz & Schlichting, *Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape*, Annals of Probability, 2014]

- PI via Lyapunov function

  [Bakry, Barthe, Cattiaux, Guillin, *A simple proof of the Poincaré inequality for a large class of probability measures*, Electron. Commun Probab, 2008]

Perturbation

Lemma: (Holley-Stroock perturbation lemma)

Suppose ν satisfies α-LSI (resp. α-PI). Let $\tilde{\nu} = \nu \cdot e^{-g}$ with

$$\operatorname{osc}(g) := \sup_x g(x) - \inf_x g(x) < \infty.$$

Then $\tilde{\nu}$ satisfies $\tilde{\alpha}$-LSI (resp. $\tilde{\alpha}$-PI) with

$$\tilde{\alpha} = \alpha \cdot e^{-2\operatorname{osc}(g)}$$

[Holley & Stroock, *Logarithmic Sobolev inequalities and stochastic Ising models*, J. Statist. Phys., 1987]

Lipschitz mapping

Lemma: Suppose ν satisfies α-LSI (resp. α-PI). Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable M-Lipschitz map. The pushforward distribution

$\|T(x) - T(y)\| \leq M \|x-y\|$

$$\tilde{\nu} = T_{\#}\nu$$

also satisfies $\tilde{\alpha}$-LSI (resp. $\tilde{\alpha}$-PI) with

$$\tilde{\alpha} = \frac{\alpha}{M^2}$$

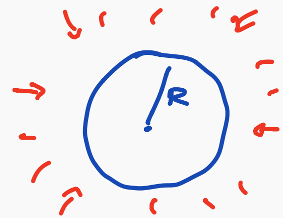
- [Vempala & Wibisono, *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, NeurIPS 2019, Lemma 13, 19]

Poincaré inequality via Lyapunov function

Let $\nu \propto e^{-f}$ on $\mathbb{R}^n$ and **Laplacian** $L = -\Delta + \nabla f \cdot \nabla$

Definition: $V : \mathbb{R}^n \to \mathbb{R}$ is a **Lyapunov function** if $V(x) \geq 1$ and if there exist $\theta > 0$, $b \geq 0$, $R > 0$ such that for all x,

$$-LV(x) \leq -\theta V(x) + b1_{B(0,R)}(x)$$



Theorem: If a Lyapunov function V exists, then ν satisfies α-PI with

$$\alpha = \frac{\theta}{(1 + b/\alpha_R)}$$

where α_R is the Poincaré constant of the restriction $\nu\big|_{B(0,R)}$.

- [Bakry, Barthe, Cattiaux, Guillin, *A simple proof of the Poincaré inequality for a large class of probability measures*, Electron. Commun Probab, 2008]

Corollary: The theorem holds for $\nu \propto e^{-f}$ in either cases below:

1. If there exist $a > 0$, $R \geq 0$, such that for all $\|x\| \geq R$,

$$\langle x, \nabla f(x) \rangle \geq a\|x\|.$$

2. If there exist $a \in (0,1)$, $c > 0$, $R \geq 0$, such that for $\|x\| \geq R$,

$$a\|\nabla f(x)\|^2 - \Delta f(x) \geq c.$$

In particular, holds when f is convex (ν log-concave).

- **KLS conjecture:** Poincaré constant of a log-concave distribution ν on $\mathbb{R}^n$ is independent of n

Rényi divergence

Rényi divergence

Rényi divergence of order $q > 0, q \neq 1$ with respect to $\nu \propto e^{-f}$ is

$$R_{q,\nu}(\rho) = \frac{1}{q-1} \log \mathbb{E}_\nu \left[\left(\frac{\rho}{\nu} \right)^q \right]$$

- **Divergence:** $R_{q,\nu}(\rho) \geq 0$ for all ρ, and $R_{q,\nu}(\rho) = 0$ iff $\rho = \nu$
- **Ordered:** $q \mapsto R_{q,\nu}(\rho)$ is increasing

Rényi divergence

Rényi divergence of order $q > 0, q \neq 1$ with respect to $\nu \propto e^{-f}$ is

$$R_{q,\nu}(\rho) = \frac{1}{q-1} \log \mathbb{E}_\nu \left[\left(\frac{\rho}{\nu}\right)^q\right]$$

- As $q \to 1$, recovers *relative entropy*

$$\lim_{q\to 1} R_{q,\nu}(\rho) = \mathbb{E}_\rho \left[\log \frac{\rho}{\nu}\right] = H_\nu(\rho)$$

- $q = 2$: recovers χ^2-divergence

$$R_{2,\nu}(\rho) = \log(1 + \chi^2_\nu(\rho)) \leq \chi^2_\nu(\rho)$$

- As $q \to \infty$, recovers *warmness*

$$\lim_{q\to\infty} R_{q,\nu}(\rho) = \log\left(\sup_x \frac{\rho(x)}{\nu(x)}\right) = \log(1 + M^\infty_\nu(\rho))$$

Rényi divergence between Gaussian distributions

Let $\rho = \mathcal{N}(0, \sigma^2 I)$ and $\nu = \mathcal{N}(0, \lambda^2 I)$ for some $\sigma^2 > \lambda^2 > 0$

- For $0 < q < \dfrac{\sigma^2}{\sigma^2 - \lambda^2}$:

$$R_{q,\nu}(\rho) = \frac{n}{2} \log \frac{\lambda^2}{\sigma^2} - \frac{n}{2(q-1)} \log \left(\frac{\sigma^2}{\lambda^2} - q\Big(\frac{\sigma^2}{\lambda^2} - 1\Big) \right)$$

- For $q \geq \dfrac{\sigma^2}{\sigma^2 - \lambda^2}$:

$$R_{q,\nu}(\rho) = \infty$$

Rényi divergence along Langevin dynamics

Theorem: Along the Langevin dynamics for $\nu \propto e^{-f}$

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

Rényi divergence of any order $q > 0$ is decreasing:

$$\frac{d}{dt} R_{q,\nu}(\rho_t) \le 0$$

Rényi divergence along Langevin dynamics

Theorem: Along the Langevin dynamics for $\nu \propto e^{-f}$

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

Rényi divergence of any order $q > 0$ is decreasing:

$$\frac{d}{dt} R_{q,\nu}(\rho_t) \le 0$$

Proof: With $h_t = \frac{\rho_t}{\nu}$ we can write

$$\frac{d}{dt} R_{q,\nu}(\rho_t) = -q \frac{\mathbb{E}_\nu[h_t^q \, \|\nabla \log h_t\|^2]}{\mathbb{E}_\nu[h_t^q]} \le 0$$

$q=1$: $\quad R_{1,\nu}(\rho_t) = H_\nu(\rho_t) \quad \Big\} \quad \frac{d}{dt} H_\nu(\rho_t) = -J_\nu(\rho_t)$ $\square$

$$\mathbb{E}_\nu[h_t^1 \, \|\nabla \log h_t\|^2] = \mathbb{E}_{\rho_t}\left[\left\|\nabla \log \frac{\rho_t}{\nu}\right\|^2\right] = J_\nu(\rho_t)$$

Rényi divergence along Langevin dynamics

Theorem: Along the Langevin dynamics for $\nu \propto e^{-f}$

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

1. If $\nu \propto e^{-f}$ satisfies α-LSI, then for $q \geq 1$:

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0)$$

Rényi divergence along Langevin dynamics

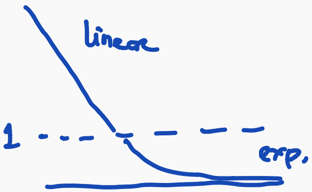
Theorem: Along the Langevin dynamics for $\nu \propto e^{-f}$

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

1. If $\nu \propto e^{-f}$ satisfies α-LSI, then for $q \geq 1$:

$$R_{q,\nu}(\rho_t) \leq e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0)$$

2. If $\nu \propto e^{-f}$ satisfies α-PI, then for $q \geq 2$:



$$R_{q,\nu}(\rho_t) \leq \begin{cases} R_{q,\nu}(\rho_0) - \frac{2\alpha t}{q} & \text{if } R_{q,\nu}(\rho_t) \geq 1 \\ e^{-\frac{2\alpha t}{q}} R_{q,\nu}(\rho_0) & \text{if } R_{q,\nu}(\rho_0) \leq 1 \end{cases}$$

Corollary: To reach $R_{q,\nu}(\rho_t) \le \epsilon$ along Langevin dynamics:

1. If $\nu \propto e^{-f}$ satisfies α-LSI, need

$$t \ge \frac{q}{2\alpha} \log \frac{R_{q,\nu}(\rho_0)}{\epsilon}.$$

2. If $\nu \propto e^{-f}$ satisfies α-PI, need

$$t \ge \frac{q}{2\alpha} \left(R_{q,\nu}(\rho_0) + \log \frac{1}{\epsilon} \right).$$

- [Vempala & Wibisono, *Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices*, NeurIPS 2019, Theorem 3, 5]
- [Cao, Lu, & Lu, *Exponential decay of Rényi divergence under Fokker-Planck equations*, Journal of Statistical Physics, 2018]

Recap: Langevin Dynamics

In continuous time, the Langevin dynamics in $\mathbb{R}^n$

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

which corresponds to the Fokker-Planck equation in $\mathcal{P}(\mathbb{R}^n)$

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

converges to the target distribution $\nu \propto e^{-f}$ exponentially fast in $\{\, W_2,\ H_\nu,\ \chi^2_\nu,\ R_{q,\nu} \}$ under various conditions { SLC, LSI, PI }

- This is in continuous time
- How to implement as an algorithm in discrete time?

Unadjusted Langevin Algorithm

Unadjusted Langevin Algorithm

Goal: Sample from $\nu \propto e^{-f}$ on $\mathbb{R}^n$

In discrete time, the **Unadjusted Langevin Algorithm (ULA)** is

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

where $\eta > 0$ is step size and $Z_k \sim \mathcal{N}(0, I)$ is independent of x_k

Unadjusted Langevin Algorithm

Goal: Sample from $\nu \propto e^{-f}$ on $\mathbb{R}^n$

In discrete time, the **Unadjusted Langevin Algorithm (ULA)** is

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

where $\eta > 0$ is step size and $Z_k \sim \mathcal{N}(0, I)$ is independent of x_k

- Why? Discretization of the Langevin dynamics:

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

$dt = \eta$

$dW_t = \sqrt{dt} = \sqrt{\eta}$

- But biased: does *not* converge to ν

ULA: Standard Gaussian

Let $\nu = \mathcal{N}(0, I)$ on $\mathbb{R}^n$, so $\nabla f(x) = x$

ULA:

$$
\begin{aligned}
x_{k+1} &= x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, Z_k \\
&= (1-\eta) x_k + \sqrt{2\eta}\, Z_k
\end{aligned}
$$

ULA: Standard Gaussian

Let $\nu = \mathcal{N}(0, I)$ on $\mathbb{R}^n$, so $\nabla f(x) = x$

ULA:

$$\begin{aligned} x_{k+1} &= x_k - \eta \nabla f(x_k) + \sqrt{2\eta}\, Z_k \\ &= (1-\eta)x_k + \sqrt{2\eta}\, Z_k \end{aligned}$$

Since $Z_0, \dots, Z_k \sim \mathcal{N}(0, I)$ independent, sum is also Gaussian:

$$\begin{aligned} x_k &= (1-\eta)^k x_0 + \sqrt{2\eta} \sum_{i=0}^{k-1} (1-\eta)^i Z_{k-1-i} \\ &\overset{d}{=} (1-\eta)^k x_0 + \sqrt{\sigma_k^2}\, \tilde{Z}_k, \qquad \tilde{Z}_k \sim \mathcal{N}(0, I) \end{aligned}$$

where

$$\sigma_k^2 = 2\eta \sum_{i=0}^{k-1} (1-\eta)^{2i} = \frac{2\eta(1-(1-\eta)^{2k})}{1-(1-\eta)^2} = \frac{1-(1-\eta)^{2k}}{1-\frac{\eta}{2}}$$

Let $0 < \eta < 2$, so $|1-\eta| < 1$ and $(1-\eta)^{2k} \to 0$. Then as $k \to \infty$,

$$x_k \xrightarrow{d} \sqrt{\frac{1}{1-\frac{\eta}{2}}}\tilde{Z}, \quad \tilde{Z} \sim \mathcal{N}(0, I)$$

Therefore, **ULA** for $\nu = \mathcal{N}(0, I)$ with $0 < \eta < 2$ converges to

$$\nu_\eta = \mathcal{N}\left(0, \frac{1}{1-\frac{\eta}{2}} I\right)$$

- **ULA** is biased: $\nu_\eta \neq \nu$ for all $\eta > 0$
- Bias scales with η: $W_2(\nu, \nu_\eta) = \sqrt{n}\left(\frac{1}{\sqrt{1-\frac{\eta}{2}}} - 1\right) = \Theta(\sqrt{n} \cdot \eta)$ ← $p=1$

ULA vs OU: Standard Gaussian

For $\nu = \mathcal{N}(0, I)$ on $\mathbb{R}^n$

1. Exact gradient flow is the solution to the Ornstein-Uhlenbeck (OU) process at time $t = \eta$:

$$x_{k+1} = e^{-\eta} x_k + \sqrt{(1 - e^{-2\eta})}\, Z_k$$

   This is unbiased, converges to ν exponentially fast

2. **ULA** is using approximation $e^{-\eta} \approx 1 - \eta$:

$$x_{k+1} = (1 - \eta) x_k + \sqrt{2\eta}\, Z_k$$

   This is biased, converges to $\nu_\eta \neq \nu$

ULA: General Gaussian

Let $\nu = \mathcal{N}(\mu, \Sigma)$

- **ULA:**

$$x_{k+1} = (I - \eta\Sigma^{-1})x_k + \eta\Sigma^{-1}\mu + \sqrt{2\eta}\, Z_k$$

- If $0 < \eta < 2\lambda_{\min}(\Sigma)$, then **ULA** converges to

$$\nu_\eta = \mathcal{N}\left(\mu, \Sigma\left(I - \frac{\eta}{2}\Sigma^{-1}\right)^{-1}\right)$$

- Bias:

$$W_2(\nu, \nu_\eta) = \left\| \Sigma^{\frac{1}{2}} - \Sigma^{\frac{1}{2}}\left(I - \tfrac{\eta}{2}\Sigma^{-1}\right)^{-\frac{1}{2}} \right\|_{\mathrm{HS}} = \frac{\eta}{4}\sqrt{\mathrm{Tr}(\Sigma^{-1})} + O(\eta^2)$$

- c.f. exact solution to OU (unbiased):

$$x_{k+1} = e^{-\eta\Sigma^{-1}}x_k + (I - e^{-\eta\Sigma^{-1}})\mu + \sqrt{\Sigma(1 - e^{-2\eta\Sigma^{-1}})}\, Z_k$$

Bias of ULA

Bias of ULA

ULA with step size $\eta > 0$ is *biased*: Converges to $\nu_\eta \neq \nu$

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

Why biased?

- **ULA** is discretization of Langevin dynamics

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

  Discretization introduces error of order $O(\eta)$

- But want convergence like **gradient descent** $\Leftrightarrow$ gradient flow

Convergence rate for optimization

Recall for optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

1. In continuous time, gradient flow:

$$\dot{X}_t = -\nabla f(X_t)$$

- If f is α-strongly convex, then

$$\|X_t - x^*\|^2 \le e^{-2\alpha t}\|X_0 - x^*\|^2$$

- To reach $\|X_t - x^*\|^2 \le \epsilon$, need

$$t \ge \frac{1}{2\alpha} \log \frac{\|X_0 - x^*\|^2}{\epsilon}$$

2. In discrete time, **gradient descent**:

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

- If f is α-strongly convex and L-smooth, and $\eta = \frac{2}{\alpha + L}$, then

$\alpha I \preceq \nabla^2 f(x) \preceq L I$

$$\|x_k - x^*\|^2 \leq e^{-\frac{2\alpha k}{L}} \|x_0 - x^*\|^2$$

- To reach $\|x_k - x^*\|^2 \leq \epsilon$, need

$$k \geq \frac{L}{2\alpha} \log \frac{\|X_0 - x^*\|^2}{\epsilon}$$

$\kappa = \frac{L}{\alpha}$

2. In discrete time, **gradient descent**:

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

- If f is α-strongly convex and L-smooth, and $\eta = \frac{2}{\alpha + L}$, then

$$\|x_k - x^*\|^2 \le e^{-\frac{2\alpha k}{L}} \|x_0 - x^*\|^2$$

- To reach $\|x_k - x^*\|^2 \le \epsilon$, need

$$k \ge \frac{L}{2\alpha} \log \frac{\|X_0 - x^*\|^2}{\epsilon}$$

In particular, **gradient descent** is unbiased: $x_k \to x^*$ as $k \to \infty$

◦ Since gradient flow is a special dynamics, converging to x^*

Dissipativity counteracts discretization error in gradient descent

Convergence rate for sampling

For sampling from $\nu \propto e^{-f}$ on $\mathbb{R}^n$

1. In continuous time, Langevin dynamics:

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

- This is implementing the gradient flow of relative entropy
- If f is α-strongly convex, then

$\Leftrightarrow \nu$ α-SLC

$\Rightarrow H_\nu$ α-SC

$$W_2(\rho_t, \nu)^2 \le e^{-2\alpha t} W_2(\rho_0, \nu)^2$$

- To reach $W_2(\rho_t, \nu)^2 \le \epsilon$, need

$$t \ge \frac{1}{2\alpha} \log \frac{W_2(\rho_0, \nu)^2}{\epsilon}$$

2. In discrete time, **Unadjusted Langevin Algorithm**:

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

- *Biased:* Converges to $\nu_\eta \neq \nu$

2. In discrete time, **Unadjusted Langevin Algorithm**:

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

- *Biased:* Converges to $\nu_\eta \neq \nu$
- If f is α-strongly convex and L-smooth, for small η, can show

$$W_2(\rho_k, \nu_\eta) \leq e^{-c\alpha\eta k} W_2(\rho_0, \nu_\eta)$$

- Suppose bias is $W_2(\nu_\eta, \nu) \leq \eta^p$. for some $p > 0$.
- Then by triangle inequality,

$$\begin{aligned} W_2(\rho_k, \nu) &\leq W_2(\rho_k, \nu_\eta) + W_2(\nu_\eta, \nu) \\ &\leq e^{-c\alpha\eta k} W_2(\rho_0, \nu_\eta) + \eta^p \end{aligned}$$

$k \to \infty$: $\leq \eta^p$

2. In discrete time, **Unadjusted Langevin Algorithm**:

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

- *Biased:* Converges to $\nu_\eta \neq \nu$
- If f is α-strongly convex and L-smooth, for small η, can show

$$W_2(\rho_k, \nu_\eta) \leq e^{-c\alpha\eta k} W_2(\rho_0, \nu_\eta)$$

- Suppose bias is $W_2(\nu_\eta, \nu) \leq \eta^p$. ← for some p : Gaussian : $p=1$
  $W_2(\nu_\eta, \nu) \leq \eta$
  $k = \tilde{\Omega}\left(\frac{1}{\alpha\varepsilon}\right)$
- Then by triangle inequality,

$$\begin{aligned} W_2(\rho_k, \nu) &\leq W_2(\rho_k, \nu_\eta) + W_2(\nu_\eta, \nu) \\ &\leq e^{-c\alpha\eta k} W_2(\rho_0, \nu_\eta) + \eta^p \end{aligned}$$

in general can show
$p = \frac{1}{2}$
$W_2(\nu_\eta, \nu) \leq \sqrt{\eta}$
$k = \tilde{\Omega}\left(\frac{1}{\alpha\varepsilon^2}\right)$

- To reach $W_2(\rho_k, \nu) \leq 2\epsilon$, set $\eta = \epsilon^{1/p}$ and need

$$k \geq \frac{1}{c\alpha\eta} \log \frac{W_2(\rho_0, \nu)}{\epsilon} = \tilde{\Omega}\left(\frac{1}{\alpha\epsilon^{1/p}}\right)$$

$\therefore$ Bias in discrete time $\Rightarrow$ ϵ-mixing time $=$ poly$(1/\epsilon)$
(v.s. in continuous time, ϵ-mixing time $= \log(1/\epsilon)$)

Bias of ULA

Why is **ULA** biased?

- **ULA** is discretization of the Langevin dynamics:

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

- Should try to discretize the Fokker-Planck equation:

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

  - This is the gradient flow $\dot{\rho}_t = -\mathrm{grad}\, H_\nu(\rho_t)$ of relative entropy
  - Want to run gradient descent $\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta\, \mathrm{grad}\, H_\nu(\rho_k))$

- **ULA** is *not* the gradient descent of relative entropy

Gradient descent of relative entropy

Gradient descent of relative entropy

Relative entropy: $H_\nu(\rho) = \mathbb{E}_\rho[\log \frac{\rho}{\nu}]$, grad $H_\nu(\rho) = -\nabla\cdot\left(\rho\ \nabla \log \frac{\rho}{\nu}\right)$

Gradient descent:

$$\begin{aligned}\rho_{k+1} &= \mathrm{Exp}_{\rho_k}(-\eta\,\mathrm{grad}\,H_\nu(\rho_k))\\ &= \mathrm{Exp}_{\rho_k}\left(\nabla\cdot\left(\rho_k\,\eta\,\nabla\log\frac{\rho_k}{\nu}\right)\right)\end{aligned}$$

Gradient descent of relative entropy

Relative entropy: $H_\nu(\rho) = \mathbb{E}_\rho[\log \frac{\rho}{\nu}]$

Gradient descent:

$$\begin{aligned}\rho_{k+1} &= \mathrm{Exp}_{\rho_k}(-\eta\, \mathrm{grad}\, H_\nu(\rho_k)) \\ &= \mathrm{Exp}_{\rho_k}\left(\nabla \cdot \left(\rho_k\, \eta\, \nabla \log \frac{\rho_k}{\nu}\right)\right)\end{aligned}$$

- Suppose ρ_k is M-log-semiconcave wrt ν: $-\nabla^2 \log \frac{\rho_k}{\nu} \succeq MI$. For $\eta \le \frac{1}{\max\{0, -M\}}$, the gradient descent above is given by:

  $$\rho_{k+1} = \left(I - \eta \nabla \log \frac{\rho_k}{\nu}\right)_{\#} \rho_k$$

  This is implemented by

  in $\mathbb{R}^n$:
  $$x_{k+1} = x_k - \eta \nabla f(x_k) - \eta \nabla \log \rho_k(x_k)$$

- Requires knowing ρ_k; not implementable in general
- Can implement in Gaussian case with Gaussian data; unbiased

Recall continuity eq: $\dot{X}_t = v_t(X_t)$

$X_t \sim \rho_t$ satisfies

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$$

Heat equation: $\dfrac{\partial \rho_t}{\partial t} = \Delta \rho_t$ — Brownian motion: $dX_t = \sqrt{2}\, dW_t$

$(\rho_t = \rho_0 * \mathcal{N}(0, 2tI))$

$$= \nabla \cdot (\rho_t \nabla \log \rho_t) = \nabla \cdot \left(\cancel{\rho_t} \frac{\nabla \rho_t}{\cancel{\rho_t}}\right) = \nabla \cdot (\nabla \rho_t)$$

$$\Leftrightarrow \frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t \nabla(-\log \rho_t))$$

this is continuity equation of

$$\dot{X}_t = -\nabla \log \rho_t(X_t) \quad \text{where } X_t \sim \rho_t$$

Gradient descent of relative entropy: Gaussian case

Let $\nu = \mathcal{N}(\mu, \Sigma)$

Let $\rho_0 = \mathcal{N}(\mu, \Sigma_0)$ with $\Sigma_0 \preceq \Sigma$, so $\rho_k = \mathcal{N}(\mu, \Sigma_k)$ stays Gaussian

Gradient descent of relative entropy:

$$x_{k+1} = \left(I + \eta(\Sigma_k^{-1} - \Sigma^{-1})\right) x_k - \eta(\Sigma_k^{-1} - \Sigma^{-1})\mu$$

Covariance:

$$\Sigma_{k+1} = \Sigma_k \left(I + \eta(\Sigma_k^{-1} - \Sigma^{-1})\right)^2 \to \Sigma$$

Gradient descent is unbiased:

$$\rho_k = \mathcal{N}(\mu, \Sigma_k) \to \mathcal{N}(\mu, \Sigma) = \nu$$

[Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018]

Proximal method of relative entropy

Proximal method of relative entropy:

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \left\{ H_\nu(\rho) + \frac{1}{2\eta} W_2(\rho, \rho_k)^2 \right\}$$

- Suppose $-\nabla^2 \log \frac{\rho_{k+1}}{\nu} \preceq LI$ and $\eta \leq \frac{1}{L}$, then

$$\left(I + \eta \nabla \log \frac{\rho_{k+1}}{\nu} \right)_{\#} \rho_{k+1} = \rho_k$$

- Requires knowing ρ_k; not implementable in general
- Can implement in Gaussian case with Gaussian data; unbiased

Proximal method of relative entropy: Gaussian case

Let $\nu = \mathcal{N}(\mu, \Sigma)$

Let $\rho_0 = \mathcal{N}(\mu, \Sigma_0)$, so $\rho_k = \mathcal{N}(\mu, \Sigma_k)$ stays Gaussian

Proximal method of relative entropy:

$$\big(I - \eta(\Sigma_{k+1}^{-1} - \Sigma^{-1})\big)\,(x_{k+1} - \mu) = x_k - \mu$$

Covariance:

$$\big(I - \eta(\Sigma_{k+1}^{-1} - \Sigma^{-1})\big)^2\,\Sigma_{k+1} = \Sigma_k \;\to\; \Sigma$$

Proximal method is unbiased:

$$\rho_k = \mathcal{N}(\mu, \Sigma_k) \to \mathcal{N}(\mu, \Sigma) = \nu$$

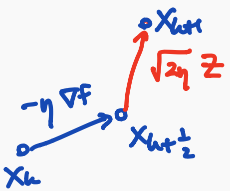
ULA as Forward-Flow

ULA

ULA:

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

Write as a composition of two steps:



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

$$x_{k+1} = x_{k+\frac{1}{2}} + \sqrt{2\eta}\, Z_k \qquad (2)$$

(1) is gradient descent (forward method) of f with step size η

(2) is Brownian motion at time η

$dX_t = \sqrt{2}\, dW_t$

$X_t = X_0 + \sqrt{2}\, W_t$, from $X_0 = x_{k+\frac{1}{2}}$

ULA in the space of distributions

ULA:

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

$$x_{k+1} = x_{k+\frac{1}{2}} + \sqrt{2\eta}\, Z_k \tag{2}$$

Let $x_k \sim \rho_k$ and $x_{k+\frac{1}{2}} \sim \rho_{k+\frac{1}{2}}$. Then

$$\rho_{k+\frac{1}{2}} = (I - \eta \nabla f)_{\#} \rho_k \tag{1}$$

$$\rho_{k+1} = \rho_{k+\frac{1}{2}} * \mathcal{N}(0, 2\eta I) \tag{2}$$

ULA in the space of distributions

ULA:

$$\rho_{k+\frac{1}{2}} = (I - \eta \nabla f)_{\#} \rho_k \tag{1}$$

$$\rho_{k+1} = \rho_{k+\frac{1}{2}} * \mathcal{N}(0, 2\eta I) \tag{2}$$

In $\mathcal{P}(\mathbb{R}^n)$ with W_2 metric:

(1) is the gradient descent of potential energy $F(\rho) = \mathbb{E}_\rho[f]$

$$\rho_{k+\frac{1}{2}} = \mathrm{Exp}_{\rho_k}(-\eta \,\mathrm{grad}\, F(\rho_k))$$

Denote this by $\rho_{k+\frac{1}{2}} = \mathrm{GD}_{F,\eta}(\rho_k)$

ULA in the space of distributions

ULA:

$$\rho_{k+\frac{1}{2}} = (I - \eta \nabla f)_{\#} \rho_k \tag{1}$$

$$\rho_{k+1} = \rho_{k+\frac{1}{2}} * \mathcal{N}(0, 2\eta I) \tag{2}$$

In $\mathcal{P}(\mathbb{R}^n)$ with W_2 metric:

(1) is the gradient descent of potential energy $F(\rho) = \mathbb{E}_\rho[f]$

$$\rho_{k+\frac{1}{2}} = \mathrm{Exp}_{\rho_k}(-\eta \, \mathrm{grad}\, F(\rho_k))$$

Denote this by $\rho_{k+\frac{1}{2}} = \mathrm{GD}_{F,\eta}(\rho_k)$

(2) is the gradient flow of negative entropy $-H(\rho) = \mathbb{E}_\rho[\log \rho]$

$$\dot{\widetilde{\rho}}_t = \mathrm{grad}\, H(\widetilde{\rho}_t) = \Delta \widetilde{\rho}_t$$

at time $t = \eta$, starting from $\widetilde{\rho}_0 = \rho_{k+\frac{1}{2}}$.

Denote this by $\rho_{k+1} = \mathrm{GF}_{-H,\eta}(\rho_{k+\frac{1}{2}})$

Relative entropy as a composite objective

Relative entropy:

$$H_\nu(\rho) = \mathbb{E}_\rho\left[\log \frac{\rho}{\nu}\right] = \int_{\mathbb{R}^n} \rho(x) \overbrace{\log \frac{\rho(x)}{\nu(x)}}^{= \log \rho - \log \nu = \log \rho + f} dx$$

Let $\nu = e^{-f} \Leftrightarrow f = -\log \nu$. Recall decomposition:

$$H_\nu(\rho) = F(\rho) - H(\rho)$$

where

(1) F is potential energy

$$F(\rho) = \mathbb{E}_\rho[f] = \mathbb{E}_\rho[-\log \nu]$$

(2) $-H$ is negative entropy

$$-H(\rho) = \mathbb{E}_\rho[\log \rho]$$

ULA as Forward-Flow

ULA is the Forward-Flow algorithm:

$$\textbf{ULA}_\eta = \mathrm{GF}_{-H,\eta} \circ \mathrm{GD}_{F,\eta}$$

for minimizing relative entropy as a composite objective function:

$$H_\nu(\rho) = F(\rho) - H(\rho)$$

= Forward method

(1) $\mathrm{GD}_{F,\eta}$ is the gradient descent of $F(\rho) = \mathbb{E}_\rho[f]$ with step size η

$$\mathrm{GD}_{F,\eta}(\rho) = (I - \eta \nabla f)_{\#}\rho$$

which can be implemented via gradient descent map of f

$$x \mapsto x - \eta \nabla f(x)$$

(2) $\mathrm{GF}_{-H,\eta}$ is the gradient flow of $-H(\rho) = \mathbb{E}_\rho[\log \rho]$ at time η

$$\mathrm{GF}_{-H,\eta}(\rho) = \rho * \mathcal{N}(0, 2\eta I)$$

can be implemented via Brownian motion / Gaussian noise

$$x \mapsto x - \sqrt{2\eta}\, Z, \qquad Z \sim \mathcal{N}(0, I)$$

However, Forward-Flow is biased for composite optimization!

Composite optimization

Composite optimization

Suppose want to minimize a composite objective $f = g + h$

$$\min_{x \in \mathbb{R}^n} g(x) + h(x)$$

- Suppose can run algorithms $\{$ GD, GF, PG $\}$ for g, h individually. How to minimize $g + h$?
- Minimizer $x^* = \arg\min_{x \in \mathbb{R}^n} g(x) + h(x)$ satisfies

  $$\nabla g(x^*) + \nabla h(x^*) = 0$$

  but $\nabla h(x^*) \neq 0$ in general
- Consistency requirement: x^* is a stationary point of the algorithm