CPSC 661: Sampling Algorithms in ML

Andre Wibisono

April 21, 2021

Yale University

Last time

- Langevin dynamics for sampling from $\nu \propto e^{-f}$

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

- Unadjusted Langevin Algorithm

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

Plan

Variations of Langevin dynamics

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

1. **First-order dynamics**
   Weighted Langevin dynamics
   Mirror Langevin dynamics
   Newton Langevin dynamics
   Interacting Langevin dynamics

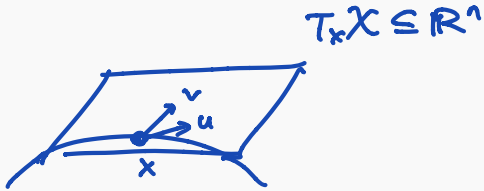
2. **Second-order dynamics**
   Underdamped Langevin dynamics
   Hamiltonian Monte Carlo

1. First-order dynamics

Optimization 1: Natural gradient flow

On manifold $\mathcal{X}$, want $\min_{x \in \mathcal{X}} F(x)$

Natural gradient flow $\dot{X}_t = -\operatorname{grad} F(X_t)$



$\langle u, v \rangle_x = u^T g(x) v$

- if $\mathcal{X} \subseteq \mathbb{R}^n$, metric (as a matrix) $g(x) > 0$

NGF: $\dot{X}_t = -\underbrace{g(X_t)^{-1}}_{M(X_t) > 0} \nabla F(X_t)$

- dissipative:

$$\begin{aligned}
\frac{d}{dt} F(X_t) &= \langle \operatorname{grad} F(X_t), \dot{X}_t \rangle_{X_t} \\
&= \langle g(X_t)^{-1} \nabla F(X_t), -g(X_t)^{-1} \nabla F(X_t) \rangle_{X_t} \\
&= -\| g(X_t)^{-1} \nabla F(X_t) \|^2_{X_t} \\
&= -\| \nabla f(X_t) \|^2_{g(X_t)^{-1}} \leq 0
\end{aligned}$$

- exp. convergence rate if Gradient Domination: $F(x) - F(x^*) \leq \frac{1}{2\alpha} \| \nabla F(x) \|^2_{g(x)^{-1}}$

Sampling 1: Weighted Langevin Dynamics

For $M(x) \succ 0$, Weighted Langevin Dynamics:

(WLD) $$dX_t = \left(\nabla \cdot M(X_t) - M(X_t) \nabla f(X_t) \right) dt + \sqrt{2 M(X_t)}\, dW_t$$

- eg. $M(x) = I$, this is Langevin dynamics

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

- Lemma: (WLD) has $\nu \propto e^{-f}$ as stationary distribution
- Fokker-Planck eq: for density ρ_t of X_t

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

$\|v\|_M^2 = v^T M(x) v$

- de Bruijn's identity:

$$\frac{d}{dt} H_\nu(\rho_t) = -\mathbb{E}_{\rho_t} \left[\left\| \nabla \log \frac{\rho_t}{\nu} \right\|_M^2 \right]$$

$$\frac{d}{dt} \chi^2_\nu(\rho_t) = -2\, \mathbb{E}_\nu \left[\left\| \nabla \frac{\rho_t}{\nu} \right\|_M^2 \right]$$

- So if we have M-weighted LSI / Poincaré:

(M-LSI) $$H_\nu(f) \le \frac{1}{2\alpha} \mathbb{E}_f\left[\|\nabla \log \frac{f}{\nu}\|_M^2\right]$$

(M-PI) $$\mathrm{Var}_\nu(g) \le \frac{1}{\alpha} \mathbb{E}_\nu\left[\|\nabla g\|_M^2\right]$$

then we have exponential convergence rate along (WLD):

M-LSI $\Rightarrow$ $$H_\nu(f_t) \le e^{-2\alpha t} H_\nu(f_0)$$

M-PI $\Rightarrow$ $$\chi^2_\nu(f_t) \le e^{-2\alpha t} \chi^2_\nu(f_0)$$

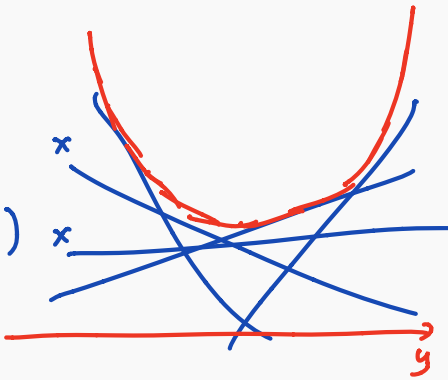
Optimization 2: Mirror flow

Hessian manifold: $g(x) = \nabla^2 \phi(x)$ for some $\phi : \mathcal{X} \to \mathbb{R}$ convex

Natural gradient flow: $\dot{x}_t = -g(x_t)^{-1} \nabla f(x_t)$

$\hookrightarrow \dot{x}_t = -(\nabla^2 \phi(x_t))^{-1} \nabla f(x_t)$

$\Leftrightarrow \frac{d}{dt} \nabla \phi(x_t) = \nabla^2 \phi(x_t) \cdot \dot{x}_t = -\nabla f(x_t)$



Mirror flow: $\boxed{\frac{d}{dt} \nabla \phi(x_t) = -\nabla f(x_t)}$

(discretize:

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

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

Convex Dual $\phi^*(y) = \sup_{x \in \mathcal{X}} \langle x, y \rangle - \phi(x)$

$\nabla \phi^*(y) = \operatorname{argmax}_{x \in \mathcal{X}} \langle x, y \rangle - \phi(x)$

$\nabla \phi^* = (\nabla \phi)^{-1}$

Bregman divergence $D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle$

(Mirror Descent) $= \operatorname{argmin}_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\eta} D_\phi(x, x_k) \right\}$

Sampling 2: Mirror Langevin Dynamics

$M(x) = \nabla^2\phi(x)^{-1}$ for some $\phi : \mathcal{X} \to \mathbb{R}$ convex

Weighted Langevin Dynamics:

$$dX = \left(\nabla \cdot (\nabla^2\phi(x)^{-1}) - \nabla^2\phi(x)^{-1}\nabla f(X) \right) dt + \sqrt{2\,\nabla^2\phi(x)^{-1}}\, dW$$

$\Rightarrow$ Stationary distribution is $\nu \propto e^{-f}$

Itô's Lemma (change of variable for SDE):

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

Mirror Langevin Dynamics

$Y = \nabla\phi(X) \sim \tilde{\nu}$ if $X \sim \nu$

Stationary distribution is $\tilde{\nu} = (\nabla\phi)_{\#}\nu$

Can get convergence rates if have LSI / PI with $M = (\nabla^2\phi)^{-1}$

$\Rightarrow$ Mirror LSI

Mirror PI

In discrete time, can discretize Mirror Langevin Dynamics:

1) $\nabla\phi(x_{k+1}) = \nabla\phi(x_k) - \eta\,\nabla f(x_k) + \sqrt{2\eta\,\nabla^2\phi(x_k)}\,z_k, \quad z_k \sim \mathcal{N}(0, I)$

$\Leftrightarrow \quad x_{k+1} = \nabla\phi^*\left(\nabla\phi(x_k) - \eta\,\nabla f(x_k) + \sqrt{2\eta\,\nabla^2\phi(x_k)}\,z_k\right)$

$= \underset{x\in\mathcal{X}}{\operatorname{argmin}} \left\{ \langle \eta\,\nabla f(x_k) - \sqrt{2\eta\,\nabla^2\phi(x_k)}\,z_k,\, x - x_k\rangle + D_\phi(x, x_k) \right\}$

[Zhang et al. 2020]

2) Mirror Langevin Algorithm [Ahn & Chewi 2020]

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

$\Leftrightarrow \quad x_{k+\frac{1}{2}} = \underset{x\in\mathcal{X}}{\operatorname{argmin}} \left\{ \langle \nabla f(x_k), x - x_k\rangle + \frac{1}{\eta} D_\phi(x, x_k) \right\}$

$x_{k+1} = \nabla\phi^*(Y_\eta)$ where $dY_t = \sqrt{2\,\nabla^2\phi^*(Y_t)^{-1}}\,dW_t$

$Y_0 = \nabla\phi(x_{k+\frac{1}{2}})$

$\Rightarrow$ Convergence analysis using self-concordance of ϕ

& Relative smoothness & strong convexity of f wrt ϕ

Optimization 3: Newton flow

Choose $g(x) = \nabla^2 f(x)$ for $\min_{x \in \mathbb{R}^n} f(x)$ $\quad (\phi = f)$

Natural gradient flow: $\dot{X}_t = -(\nabla^2 f(X_t))^{-1} \nabla f(X_t)$

$\Updownarrow$

Newton flow : $\frac{d}{dt} \nabla f(X_t) = -\nabla f(X_t)$

$\Leftrightarrow \quad \nabla f(X_t) = e^{-t} \nabla f(X_0)$

Convergence $\to 0$ exp. fast with uniform rate

Ex: $f(x) = \frac{1}{2} x^T A x$

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

$\Rightarrow X_t = e^{-At} X_0 \to 0$ with rate $\lambda_{\min}(A)$

NF: $\dot{X}_t = -(\nabla^2 f(X_t))^{-1} \nabla f(X_t)$

$= -A^{-1} \cdot AX_t = -X_t$

$\Rightarrow X_t = e^{-t} X_0 \to 0$ with rate 1.

in discrete time:

Newton's method: $x_{k+1} = x_k - \eta \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$

Analysis via self-concordance + barrier properties

Sampling 3: Newton Langevin Dynamics

Mirror Langevin Dynamics with $\phi = f$:

(NLD): $d\nabla f(X_t) = -\nabla f(X_t)\,dt + \sqrt{2\,\nabla^2 f(X_t)}\,dW_t$

- Affine invariant
- Exponential convergence rate for any log-concave $\nu \propto e^{-f}$

Theorem [Brascamp-Lieb inequality '76]

$\forall$ strictly convex $f: \mathbb{R}^n \to \mathbb{R}$, $\nu \propto e^{-f}$ satisfies:

$$\mathrm{Var}_\nu(h) \le 1 \cdot \mathbb{E}_\nu\left[\|\nabla h\|^2_{(\nabla^2 f)^{-1}}\right]$$

$\Leftrightarrow$ ν satisfies Poincaré ineq. with constant 1 wrt. metric $\nabla^2 f$

$\Rightarrow$ Along (NLD), $\chi^2_\nu(\rho_t) \le e^{-2t}\,\chi^2_\nu(\rho_0)$.

References 1

- Brascamp & Lieb, *On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation*, *Journal of Functional Analysis*, 1976
- Hsieh, Kavis, Rolland, & Cevher, *Mirrored Langevin dynamics*, NeurIPS 2018
- Chewi, Le Gouic, Lu, Maunu, Rigollet, & Stromme, *Exponential ergodicity of mirror-Langevin diffusions*, NeurIPS 2020
- Zhang, Peyré, Fadili, & Pereyra, *Wasserstein control of mirror Langevin Monte Carlo*, COLT 2020
- Ahn & Chewi, *Efficient constrained sampling via the mirror-Langevin algorithm*, arXiv 2020

Sampling 4: Interacting Langevin Dynamics

Another way to get affine invariance:

Precondition by covariance

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

where $C_t = \mathrm{Cov}(\rho_t) = \mathbb{E}_{\rho_t}\left[(X_t - \mu_t)(X_t - \mu_t)^T\right]$, $X_t \sim \rho_t$

$$\mu_t = \mathbb{E}_{\rho_t}[X_t]$$

Fokker-Planck eq: now nonlinear

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

This is "Kalman-Wasserstein gradient flow"

↳ metric $\phi = -\nabla \cdot (\rho\, C(\rho)\, \nabla u)$

$$\|\phi\|_\rho^2 = \mathbb{E}_\rho\left[\|\nabla u\|^2_{C(\rho)}\right]$$

↑

this is mean-field ($N \to \infty$)

of Affine-Invariant Langevin Dynamics (ALDI)

have particles $X_t = (X_t^{(1)}, \dots, X_t^{(N)})$, $X_t^{(i)} \in \mathbb{R}^n$

$\uparrow$

$\mathbb{R}^{n \times N}$

empirical mean $\mu(X_t) = \frac{1}{N} \sum_{i=1}^{N} X_t^{(i)}$

covariance $C(X_t) = \frac{1}{N} \sum_{i=1}^{N} (X_t^{(i)} - \mu(X_t))(X_t^{(i)} - \mu(X_t))^T$

ALDI: $dX_t^{(i)} = -C(X_t)\left(\nabla f(X_t^{(i)}) - \frac{n+1}{2} \nabla \log\det C(X_t)\right) dt$

$+ \sqrt{2 C(X_t)}\, dW_t^{(i)}$

for $i = 1, \dots, N$

$\Leftrightarrow dX_t^{(i)} = \left(-C(X_t)\nabla f(X_t) + \frac{n+1}{N}(X_t^{(i)} - \mu(X_t))\right) dt$

$+ \sqrt{2 C(X_t)}\, dW_t^{(i)}$

References 2

- Garbuno-Inigo, Hoffmann, Li, & Stuart, *Interacting Langevin Diffusions: Gradient Structure And Ensemble Kalman Sampler*, SIAM Journal on Applied Dynamical Systems, 2020

- Garbuno-Inigo, Nüsken, & Reich, *Affine invariant interacting Langevin dynamics for Bayesian inference*, SIAM Journal on Applied Dynamical Systems, 2020

Example: Gaussian $\nu = \mathcal{N}(0, \Sigma)$ $\rightarrow$ $\nabla f(x) = \Sigma^{-1} x$

$\nabla^2 f(x) = \Sigma^{-1}$

1) OU (Langevin Dynamics)

$$dX = -\Sigma^{-1} X\, dt + \sqrt{2}\, dW$$

$$\Rightarrow X_t \overset{d}{=} e^{-\Sigma^{-1} t} X_0 + \sqrt{\Sigma\left(I - e^{-2\Sigma^{-1} t}\right)}\, Z$$

$\rho_t \rightarrow \nu$ with rate $\lambda_{\min}(\Sigma^{-1}) = \dfrac{1}{\lambda_{\max}(\Sigma)}$

2) Newton Langevin Dynamics:

$$d\nabla f(X) = -\nabla f(X)\, dt + \sqrt{2\nabla^2 f(x)}\, dW$$

$$d(\Sigma^{-1} X) = -\Sigma^{-1} X\, dt + \sqrt{2\Sigma^{-1}}\, dW$$

$$dX = -X\, dt + \sqrt{2\Sigma}\, dW$$

$$d(e^t X) = e^t (dX + X\, dt) = \sqrt{2\Sigma}\, e^t\, dW_t$$

$$\Rightarrow X_t \overset{d}{=} e^{-t} X_0 + \sqrt{(1 - e^{-2t})\Sigma}\, Z$$

$\rho_t \rightarrow \nu$ with rate 1.

3) Interacting Langevin Dynamics

If $X_0 \sim \rho_0 = \mathcal{N}(0, \Sigma_0)$

Then $X_t \sim \rho_t = \mathcal{N}(0, \Sigma_t)$

$$\Sigma_t^{-1} = (\Sigma_0^{-1} - \Sigma^{-1}) e^{-2t} + \Sigma^{-1}$$

$\rho_t \rightarrow \nu$ with rate 1.