CPSC 661: Sampling Algorithms in ML

Andre Wibisono

April 26, 2021

Yale University

Last time

- Langevin dynamics
- Unadjusted Langevin Algorithm
- Variants of Langevin dynamics (first-order):
  {Weighted, Mirror, Newton, Interacting} Langevin Dynamics

Today: Underdamped Langevin Dynamics

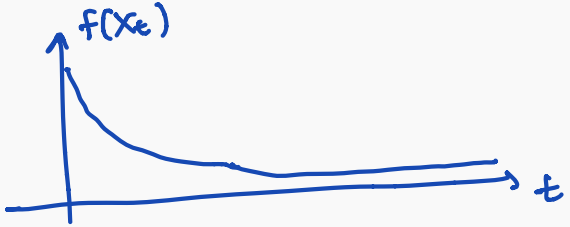
First vs. Second-Order Dynamics

$$\dot{x}_t = -\nabla f(x_t)$$

- First order (in time) dynamics
- Gradient flow
- Dissipative: Minimizes f

$$\frac{d}{dt} f(x_t) = -\|\nabla f(x_t)\|^2 \leq 0$$

- Greedy descent flow



$$\frac{d^2}{dt^2} x_t = \ddot{x}_t = -\nabla f(x_t)$$

- Second-order (in time) dynamics
- Newton's Law:

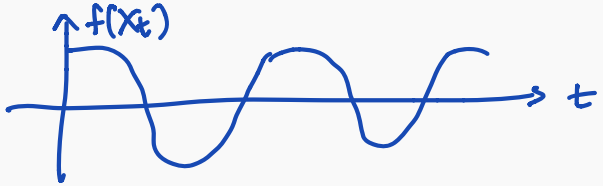
$$\underbrace{\text{mass}}_{m=1} \times \underbrace{\text{Acceleration}}_{\ddot{x}_t} = \underbrace{\text{Force}}_{-\nabla f(x_t)}$$

- Conservative: Preserves Hamiltonian

$$H(t) = \frac{1}{2}\|\dot{x}_t\|^2 + f(x_t)$$

$$\frac{d}{dt} H(t) = 0$$

- Function value $f(x_t)$ is oscillatory



Eg. $f(x) = \frac{1}{2} x^T A x$

$\nabla f(x) = Ax, \quad A = A^T > 0$

Gradient flow:

$$\dot{x}_t = -\nabla f(x_t) = -Ax_t$$

$$\Rightarrow x_t = e^{-At} x_0 \longrightarrow 0$$

Hamiltonian flow:

$$\ddot{x}_t = -\nabla f(x_t) = -Ax_t$$

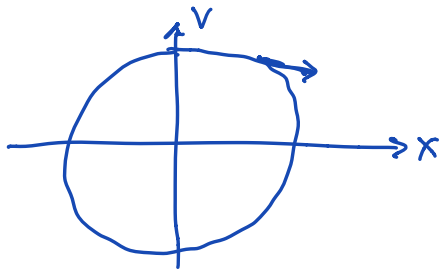
$$x_t = (\cos At)\,\beta_1 + (\sin At)\,\beta_2$$

where $\beta_1, \beta_2 \in \mathbb{R}^n$ determined by

$$x_0, \dot{x}_0 \in \mathbb{R}^n$$

in phase space:

$$\left.\begin{aligned} \dot{x}_t &= v_t \\ \dot{v}_t &= -Ax_t \end{aligned}\right\} \Leftrightarrow \ddot{x}_t = -Ax_t$$



Optimization: Acceleration

Newton's Law with friction / damping

(HB) $\ddot{X}_t + \gamma \dot{X}_t + \nabla f(X_t) = 0$, $\gamma > 0$ damping

(also called * Heavy ball [Polyak, Nesterov]

* accelerated gradient flow)

* no damping : $\gamma = 0 \Rightarrow$ Newton's Law $\ddot{X}_t + \nabla f(X_t) = 0$

* overdamped : $\gamma \to \infty$: $\tilde{X}_t = X_{\gamma t}$ recovers gradient flow

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

because $\tilde{X}_t$ satisfies

$\frac{1}{\gamma} \ddot{\tilde{X}}_t + \dot{\tilde{X}}_t + \nabla f(\tilde{X}_t) = 0$

Eg. $f(x) = \frac{\alpha}{2}x^2$ on $x \in \mathbb{R}$ $\rightarrow$ strongly convex w. param α

then (HB): $\ddot{x}_t + \gamma \dot{x}_t + \alpha x_t = 0$

Characteristic eq: $r^2 + \gamma r + \alpha = 0$

$$\Leftrightarrow r_{12} = \frac{-\gamma \pm \sqrt{\gamma^2 - 4\alpha}}{2}$$

Solution to (HB) is $x_t = c_1 \cdot e^{r_1 t} + c_2 \cdot e^{r_2 t}$, $c_1, c_2 \in \mathbb{R}$

Case 1: Underdamped case: ($i = \sqrt{-1}$)

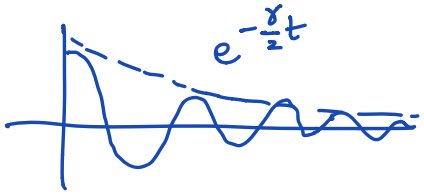
if $\gamma < 2\sqrt{\alpha}$, then $r_{12} = \frac{-\gamma \pm i\omega}{2}$ $\quad \omega = \sqrt{4\alpha - \gamma^2} \in \mathbb{R}$

$$e^{r_{12}t} = e^{-\left(\frac{\gamma \pm i\omega}{2}\right)t}$$

$\left(e^{i\theta} = \cos\theta + i\sin\theta\right)$

$$= e^{-\frac{\gamma}{2}t}\left(\cos\left(\frac{\omega t}{2}\right) \pm i \sin\left(\frac{\omega t}{2}\right)\right)$$

solution is $x_t = e^{-\frac{\gamma}{2}t}\left(c_1 \cdot \cos\left(\frac{\omega t}{2}\right) + c_2 \cdot \sin\left(\frac{\omega t}{2}\right)\right)$



* best rate is $\gamma = 2\sqrt{\alpha}$ (critical damping)

$\Rightarrow O\left(e^{-\sqrt{\alpha}t}\right)$ convergence

(vs. Gradient flow $\dot{x}_t = -\alpha x_t$

$\Rightarrow x_t = e^{-\alpha t} x_0 \Rightarrow O(e^{-\alpha t})$ convergence

when $0 < \alpha \ll 1$, $\sqrt{\alpha} \gg \alpha$

Heavy ball : $\ddot{x}_t + \gamma \dot{x}_t + \nabla f(x_t) = 0$

* Hamiltonian is decreasing:

$$\frac{d}{dt}\left(\frac{1}{2}\|\dot{x}_t\| + f(x_t)\right) = -\gamma \|\dot{x}_t\|^2 \le 0$$

* Theorem:

Assume f is α-strongly convex $(\nabla^2 f(x) \succeq \alpha I)$

and $0 < \gamma \le 2\sqrt{\alpha}$

then along heavy ball,

$$\mathcal{E}_t = f(x_t) - f(x^*) + \frac{\alpha}{2}\left\| x_t - x^* + \frac{2}{\gamma}\dot{x}_t \right\|^2$$

is a Lyapunov function, and it converges to 0 exp. fast:

$$\frac{d}{dt}\mathcal{E}_t \le -\frac{\gamma}{2}\mathcal{E}_t$$

$$\Rightarrow \quad \mathcal{E}_t \le e^{-\frac{\gamma}{2}t}\mathcal{E}_0$$

$$\Rightarrow \quad f(x_t) - f(x^*) \le \mathcal{E}_t \le e^{-\frac{\gamma}{2}t}\mathcal{E}_0$$

Then best rate is $\gamma = 2\sqrt{\alpha}$:

$$\Rightarrow f(x_t) - f(x^*) \le O\left(e^{-\sqrt{\alpha}t}\right)$$

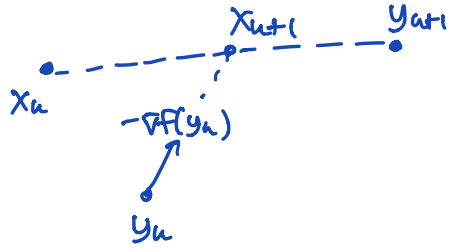
(c.f. Gradient flow

$f(x_t) - f(x^*) \le e^{-\alpha t}$)

↓ discretize:

Accelerated gradient descent (momentum method)

(AGD)
$$x_{k+1} = y_k - \eta \nabla f(y_k)$$
$$y_{k+1} = x_{k+1} + \beta \cdot (x_{k+1} - x_k)$$



Thm: If f is α- strongly convex, L-smooth

$$\alpha I \preceq \nabla^2 f(x) \preceq L I, \qquad \kappa = \frac{L}{\alpha}$$

then with $\eta = \frac{1}{L}$, $\beta = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$:

AGD has $f(x_k) - f(x^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \cdot C_0$

$\Rightarrow$ complexity is $O(\sqrt{\kappa})$

(c.f. gradient descent: complexity $O(\kappa)$

& AGD $O(\sqrt{\kappa})$ is optimal, $\exists$ lower bound $\Omega(\sqrt{\kappa})$)

Note:
- HB: $\ddot{X}_t + \gamma \dot{X}_t + \nabla f(X_t) = 0$ for f strongly convex
- can use $\ddot{X}_t + \frac{3}{t} \dot{X}_t + \nabla f(X_t) = 0$ for f convex

(gets $O\left(\frac{1}{t^2}\right)$ convergence rate vs. $O\left(\frac{1}{t}\right)$ for GF)

Underdamped Langevin Dynamics

$$\text{(ULD)} \quad \ddot{X}_t + \gamma \dot{X}_t + \nabla f(X_t) + \sqrt{2\gamma}\, \xi_t = 0$$

where $\gamma > 0$ is damping

$$\xi_t = \frac{dW_t}{dt} \quad \text{is "white noise"}$$

$$\text{(ULD)} \iff \boxed{\begin{aligned} dX_t &= V_t\, dt \\ dV_t &= (-\gamma V_t - \nabla f(X_t))\, dt + \sqrt{2\gamma}\, dW_t \end{aligned}}$$

Facts: * In (x, v) space, stationary distribution is

$$\rho_\infty(x, v) = \nu(x) \cdot \mathcal{N}(v)$$

where $\nu(x) \propto e^{-f(x)}$ for $x \in \mathbb{R}^n$

$$\mathcal{N}(v) = \mathcal{N}(v; 0, I) \propto e^{-\frac{1}{2}\|v\|^2} \quad \text{for } v \in \mathbb{R}^n$$

Let $H(x,v) = f(x) + \frac{1}{2}\|v\|^2$

then $\rho_\infty(x,v) \propto e^{-H(x,v)}$

* As $\gamma \to \infty$, $\tilde{X}_t = X_{\gamma t}$ recovers (overdamped) Langevin dynamics:

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

* Hypoelliptic diffusion, not uniformly elliptic

(diffusion only on v variable)

* Lemma: If $(X_t, V_t) \sim \rho_t(x,v)$ follows ULD

then ρ_t follows the "kinetic Fokker-Planck equation":

$$\frac{\partial \rho_t}{\partial t} = -\langle v, \nabla_x \rho_t(x,v)\rangle + \langle \nabla f(x), \nabla_v \rho_t(x,v)\rangle \quad \} \text{ comes from Hamiltonian flow}$$

$$+ \gamma\left(\nabla_v \cdot (v\,\rho_t(x,v)) + \Delta_v \rho_t(x,v)\right) \quad \} \text{ comes from Langevin dynamics}$$

because ULD is a mixture:

$$\begin{pmatrix} dX_t \\ dV_t \end{pmatrix} = \underbrace{\begin{pmatrix} v\,dt \\ -\nabla f(x)\,dt \end{pmatrix}}_{\text{Hamiltonian flow}} + \underbrace{\begin{pmatrix} 0 \\ -\gamma v\,dt + \sqrt{2\gamma}\,dW_t \end{pmatrix}}_{\text{Langevin dynamics for } v \text{ for target } \mathcal{N}(0,I)}$$

$\dot{x} = v$, $\dot{v} = -\nabla f(x)$ $\} \Leftrightarrow \ddot{x}_t = -\nabla f(x_t)$

* So for example,

$$\frac{d}{dt}\,\chi^2_{\pi_\infty}(\pi_t) = -2\gamma\,\mathbb{E}_{\pi_\infty}\left[\left\|\nabla_v \frac{\pi_t(x,v)}{\pi_\infty(x,v)}\right\|^2\right] \le 0$$

How to get convergence rate ?

* Villani, Hypocoercivity, 2009
* Eberle, Guillin, Zimmer, Synchronous + Reflection coupling, 2019

Will see :

* Cao, Lu, Wang, Poincaré inequality
* Dalalyan & Riou-Durand, Synchronous coupling

  → discrete-time algorithm

(1) Convergence Rate via Poincaré:

Theorem: [Cao, Lu, Wang, 2020]

Assume:
- $\nu(x) \propto e^{-f(x)}$ satisfies α-Poincaré inequality
- $\|\nabla^2 f(x)\| \le M(1 + \|\nabla f(x)\|)$ for all $x \in \mathbb{R}^n$, some $M \ge 1$
- f is super linear:

$$\exists\, \beta > 1 : \lim_{x \to \infty} \frac{f(x)}{\|x\|^\beta} = \infty$$

($\Rightarrow$ embedding $H^1(\nu) \hookrightarrow L^2(\nu)$ is compact)

Then ULD:

$$dX = V\,dt$$
$$dV = (-\gamma V - \nabla f(X))\,dt + \sqrt{2\gamma}\,dW$$

has:

$$\chi^2_{g_\infty}(g_t) \le C_0 \cdot e^{-\lambda t}\, \chi^2_{g_\infty}(g_0)$$

where

$$\lambda = \sqrt{\alpha} \cdot \log\left(1 + \frac{\gamma\sqrt{\alpha}}{C_1(\sqrt{\alpha} + R + \gamma)^2}\right)$$

where: 1) if f convex, then $R = 0$

then with $\gamma = \sqrt{\alpha}$: $\lambda = \sqrt{\alpha} \cdot \log\left(1 + \frac{1}{4C_1}\right) = \Theta(\sqrt{\alpha})$

2) if $\nabla^2 f(x) \ge -K \cdot I$, $K > 0$, then $R = \sqrt{K}$ for $\alpha \ll 1$:

then with $\gamma = \sqrt{\alpha} + K$: $\lambda = \sqrt{\alpha} \cdot \log\left(1 + \frac{\sqrt{\alpha}}{\sqrt{\alpha} + K}\right) \approx \frac{\alpha}{\sqrt{\alpha} + K}$

② Convergence Rate by Synchronous Coupling

[Cheng et al. 2018, Dalalyan & Riou-Durand 2020]

Theorem: Assume f is α-strongly convex, L-smooth

Then ULD: $dX_t = V_t\,dt$

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

with $\gamma \geq \sqrt{L+\alpha}$,

ULD has exponential contraction, in particular to π_∞:

$$W_2(\pi_t, \pi_\infty) \leq \sqrt{2} \cdot e^{-\frac{\alpha}{\gamma}t}\, W_2(\pi_0, \pi_\infty)$$

in discrete time: approximate ULD by

$$dX_t = V_t\,dt$$

$$dV_t = (-\gamma V_t - \nabla f(X_0))\,dt + \sqrt{2\gamma}\,dW_t$$

Can integrate, get algorithm:

Kinetic Langevin Monte Carlo (KMLC):

$$X_{k+1} = X_k + \psi_1(\eta)\,V_k - \psi_2(\eta)\,\nabla f(X_k) + \sqrt{2\gamma}\,\xi_k$$

$$V_{k+1} = \psi_0(\eta)\,V_k - \psi_1(\eta)\,\nabla f(X_k) + \sqrt{2\gamma}\,\xi_k'$$

where $\psi_0(\eta) = e^{-\gamma\eta}$ $(\approx 1-\gamma\eta)$

$$\psi_1(\eta) = \int_0^\eta \psi_0(t)\,dt = \frac{1-e^{-\gamma\eta}}{\gamma} \quad (\approx \eta)$$

$$\psi_2(\eta) = \int_0^\eta \psi_1(t)\,dt = \frac{e^{-\gamma\eta} - 1 + \gamma\eta}{\gamma^2} \quad \left(\approx \frac{\eta^2}{2}\right)$$

and $\xi_k = (\xi_{k,1}, \dots, \xi_{k,n})$

$\xi'_k = (\xi'_{k,1}, \dots, \xi'_{k,n})$

$\forall i: \quad (\xi_{k,i}, \xi_{k,i'}) \sim \mathcal{N}(0, \Sigma)$

where $\Sigma = \int_0^\gamma \begin{pmatrix} \psi_0(t) \\ \psi_1(t) \end{pmatrix} (\psi_0(t) \quad \psi_1(t)) \, dt$

Theorem: [Dalalyan & Riou-Durand]

Assume f is α-strongly convex, L-smooth

$(\alpha I \le \nabla^2 f(x) \le L I) \qquad \kappa = \frac{L}{\alpha}$

then KLMC from any $X_0 \sim \rho_0$

$v_0 \sim \mathcal{N}(0, I)$

with $\gamma \ge \sqrt{L+\alpha}$

satisfies $W_2(\rho_k, \rho_\infty) \le \sqrt{2} \left(1 - \frac{\alpha\eta}{\gamma}\right)^k W_2(\rho_0, \rho_\infty) + \kappa \cdot \eta \sqrt{n}$

$\Rightarrow$ to reach $W_2(\rho_k, \rho_\infty) \le \varepsilon$

mixing time of KLMC is $\tilde{O}\left(\frac{\kappa^{3/2} \cdot \sqrt{n}}{\sqrt{\alpha} \cdot \varepsilon}\right)$

compare with ULA: $\tilde{O}\left(\frac{\kappa^2 \cdot n}{\alpha \cdot \varepsilon^2}\right)$ from coupling

$\Rightarrow \tilde{O}\left(\frac{\kappa \cdot n}{\alpha \cdot \varepsilon^2}\right)$ from convex opt.

References

- Villani, *Hypocoercivity*, Memoirs of the American Mathematical Society, 2009
- Eberle, Guillin, & Zimmer, *Couplings and quantitative contraction rates of Langevin dynamics*, The Annals of Probability, 2019

References

- Cheng, Chatterji, Bartlett, & Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, COLT 2018
- Shen & Lee, *The Randomized Midpoint Method for Log-Concave Sampling*, NeurIPS 2019
- Dalalyan & Riou-Durand, *On sampling from a log-concave density using kinetic Langevin diffusions*, Bernoulli, 2020
- Cao, Lu, & Wang, *On explicit ℓ_2-convergence rate estimate for underdamped Langevin dynamics*, arXiv 2019
- Ma, Chatterji, Cheng, Flammarion, Bartlett, & Jordan, *Is there an analog of Nesterov acceleration for MCMC?*, arXiv 2019
- Li, Zha, & Tao, *Hessian-Free High-Resolution Nesterov Acceleration for Sampling*, arXiv 2020