Gradient-Based Markov chain Monte Carlo Sampling

UT Austin, Foundations of Data Science Spring ’22

April 20, 2022

Outline

- Overdamped Langevin Dynamics
  - Continuous-time properties
  - Discrete-time convergence
- Metropolis-Hastings Adjustment
- Connection to Optimization (GF interpretation)
- Other 1st order method + Bigger Picture

Disclaimer

1. Vast area involving many fields (incomplete even for what concerns this topic) but the goal is to convey the flavor of result out there
2. Some of the calculations are formal derivations (e.g., exchange differentiation and integral) but every step can be made rigorous

Goal & Setup for MCMC Sampling

- Potential function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Given gradient access $\nabla f(\cdot)$

Goal & Setup for MCMC Sampling

- Potential function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Given gradient access $\nabla f(\cdot)$
- Draw samples from $\pi \propto e^{-f}$ (normalizing constant involving high-dimensional integral usually unknown)

Goal & Setup for MCMC Sampling

- Potential function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Given gradient access $\nabla f(\cdot)$
- Draw samples from $\pi \propto e^{-f}$ (normalizing constant involving high-dimensional integral usually unknown)
- Construct Markov chain that has π as stationary distribution:

$$\pi = \pi P = \pi P^k \quad \forall k \geq 1$$

  for time-homogeneous transition matrix (uniquely characterize the chain)

$$P_{ij} = P(X_{k+1} = j | X_k = i) \quad k \geq 0$$

Goal & Setup for MCMC Sampling

- Potential function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Given gradient access $\nabla f(\cdot)$
- Draw samples from $\pi \propto e^{-f}$ (normalizing constant involving high-dimensional integral usually unknown)
- Construct Markov chain that has π as stationary distribution:

$$\pi = \pi P = \pi P^k \quad \forall k \geq 1$$

  for time-homogeneous transition matrix (uniquely characterize the chain)

$$P_{ij} = P(X_{k+1} = j | X_k = i) \quad k \geq 0$$

- Hope is that $\rho_k = \rho_0 P^k \to \pi$ as $k \to \infty$ and simulate the process

Goal & Setup for MCMC Sampling

- Potential function $f(x) : \mathbb{R}^d \to \mathbb{R}$. Given gradient access $\nabla f(\cdot)$
- Draw samples from $\pi \propto e^{-f}$ (normalizing constant involving high-dimensional integral usually unknown)
- Construct Markov chain that has π as stationary distribution:

$$\pi = \pi P = \pi P^k \quad \forall k \geq 1$$

  for time-homogeneous transition matrix (uniquely characterize the chain)

$$P_{ij} = P(X_{k+1} = j | X_k = i) \quad k \geq 0$$

- Hope is that $\rho_k = \rho_0 P^k \to \pi$ as $k \to \infty$ and simulate the process
- One of top 10 most influential algorithms of the 20th century by SIAM (others include Simplex for LP, FFT, Krylov Subspace, Fast Multipole ...). Widely used across the sciences.

Some Notations and Definitions

For vector field $v : \mathbb{R}^d \to \mathbb{R}^d$ and function $f : \mathbb{R}^d \to \mathbb{R}$,

- Divergence: $(\nabla \cdot v)(x) = \sum_{i=1}^{d} \frac{\partial v_i(x)}{\partial x_i}$
- Laplacian:

$$\Delta f(x) = \mathrm{Tr}(\nabla^2 f(x)) = \sum_{i=1}^{d} \frac{\partial^2 f(x)}{\partial x_i^2}$$

  therefore $(\nabla \cdot \nabla f)(x) = \Delta f(x)$
- Wasserstein-2 distance:

$$W_2^2(\rho, \pi) = \inf_{x \sim \rho, y \sim \pi} \mathbb{E}[\|x - y\|_2^2]$$

- KL Divergence:

$$D_{KL}(\rho || \pi) = \int \rho(x) \log \frac{\rho(x)}{\pi(x)} dx = \mathbb{E}_\rho[f] + \mathrm{NegEnt}(\rho)$$

- β-smoothness & α-strong convexity: $\alpha \cdot I \preceq \nabla^2 f \preceq \beta \cdot I$

Overdamped Langevin Dynamics

- Unadjusted Overdamped Langevin

$$dX_t = \underbrace{-\nabla f(X_t)}_{\text{drift}}\,dt + \underbrace{\sqrt{2}dW_t}_{\text{diffusion}} \tag{1}$$

Overdamped Langevin Dynamics

- Unadjusted Overdamped Langevin

$$dX_t = \underbrace{-\nabla f(X_t)}_{\text{drift}} dt + \underbrace{\sqrt{2} dW_t}_{\text{diffusion}} \tag{1}$$

- Stochastic Differential Equation (SDE) with Brownian motion:

$$W_0 = 0, W_{s+t} - W_s \sim \mathcal{N}(0, t), \text{indep increments}$$

Overdamped Langevin Dynamics

- Unadjusted Overdamped Langevin

$$dX_t = \underbrace{-\nabla f(X_t)}_{\text{drift}} dt + \underbrace{\sqrt{2} dW_t}_{\text{diffusion}} \tag{1}$$

- Stochastic Differential Equation (SDE) with Brownian motion:

$$W_0 = 0, W_{s+t} - W_s \sim \mathcal{N}(0, t), \text{indep increments}$$

- Came out of physics (not related to sampling in its original context), as many other ideas in this area

Overdamped Langevin Dynamics

- Unadjusted Overdamped Langevin

$$dX_t = \underbrace{-\nabla f(X_t)}_{\text{drift}} dt + \underbrace{\sqrt{2} dW_t}_{\text{diffusion}} \tag{1}$$

- Stochastic Differential Equation (SDE) with Brownian motion:

$$W_0 = 0, W_{s+t} - W_s \sim \mathcal{N}(0, t), \text{indep increments}$$

- Came out of physics (not related to sampling in its original context), as many other ideas in this area
- Fokker-Planck (forward Kolmogorov) equation governs evolution of density $X_t \sim \rho_t$ from which it is clear $\rho_t = \pi \propto e^{-f}$ is the right invariant measure

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

  This is a PDE. Connection between SDE/PDE goes much deeper!

Derivation of Fokker-Planck Equation (1D)

Take $h : \mathbb{R} \to \mathbb{R}$ smooth, compactly supported. Ignore $\mathcal{O}(\delta^2)$ terms:

$$h(X_{t+\delta}) = h(X_t) + h'(X_t)(-\nabla f(X_t)\delta + \sqrt{2\delta}Z) + \frac{1}{2}h''(X_t)2\delta Z^2$$

Let $E(t) = \mathbb{E}[h(X_t)] = \int h(x)\rho(x,t)dx$ therefore

$$\dot{E}(t) = \lim_{\delta \to 0} \frac{1}{\delta}(E(t+\delta) - E(\delta)) = \int h(x)\frac{\partial}{\partial t}\rho(x,t)dt$$

Take expectation ($Z \sim \mathcal{N}(0,1)$ independent from X_t),

$$\mathbb{E}[h(X_{t+\delta})] = \mathbb{E}[h(X_t)] + \underbrace{\mathbb{E}[-h'(X_t)\nabla f(X_t) + h''(X_t)]}_{\dot{E}(t)}\,\delta$$

Therefore

$$\begin{aligned}\int h(x)\frac{\partial}{\partial t}\rho(x,t)dt &= \int \rho(x,t)[-h'(x)\nabla f(x) + h''(x)]dx \\ &= \int h(x)[\frac{\partial}{\partial x}(\nabla f(x)\rho(x,t)) + \frac{\partial^2}{\partial x^2}\rho(x,t)]dx\end{aligned}$$

where we used IBP twice and conclude by noting h is arbitrary.

Convergence in Continuous Time

Under assumption f is α-strongly convex. Synchoronus coupling: same Brownian motion for two dynamics

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

$$dY_t = -\nabla f(Y_t)dt + \sqrt{2}dW_t$$

Therefore

$$\begin{aligned}\frac{d}{dt}\|X_t - Y_t\|_2^2 &= 2\langle X_t - Y_t, \nabla f(Y_t) - \nabla f(X_t)\rangle \\ &\leq -2\alpha\|X_t - Y_t\|_2^2\end{aligned}$$

So if we start $X_0 \sim \rho_0$, $Y_0 \sim \pi$, let $X_t \sim \rho_t, Y_t \sim \pi$

$$W_2^2(\rho_t, \pi) \leq \mathbb{E}[\|X_t - Y_t\|_2^2] \leq \exp(-2\alpha t) \cdot \mathbb{E}[\|X_0 - Y_0\|_2^2]$$

Min over all couplings (ρ_0, π) gives Wasserstein contraction for (1).

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$

converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$

  converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.
- Rate: $\mathcal{O}(\text{poly}(\frac{1}{\epsilon}))$ w/o warm-start (dictated by the stepsize)

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:
$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$
converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.
- Rate: $\mathcal{O}(\text{poly}(\frac{1}{\epsilon}))$ w/o warm-start (dictated by the stepsize)
- Assumption: α-strong convexity + β-smoothness. Weaker assumption than strong-log-concavity based on isoperemtry inequality exists (Log-Sobolev, Poincaré, ...) but more technical.

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$

  converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.
- Rate: $\mathcal{O}(\text{poly}(\frac{1}{\epsilon}))$ w/o warm-start (dictated by the stepsize)
- Assumption: α-strong convexity + β-smoothness. Weaker assumption than strong-log-concavity based on isoperemtry inequality exists (Log-Sobolev, Poincaré, ...) but more technical.
- Other discretization can be considered: (proximal-type)

$$x_{k+1} = \arg\min_x \; f(x) + \frac{1}{2h}\|x - (x_k + \sqrt{2h} \cdot z_{k+1})\|_2^2$$

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$

converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.

- Rate: $\mathcal{O}(\text{poly}(\frac{1}{\epsilon}))$ w/o warm-start (dictated by the stepsize)
- Assumption: α-strong convexity + β-smoothness. Weaker assumption than strong-log-concavity based on isoperemtry inequality exists (Log-Sobolev, Poincaré, ...) but more technical.
- Other discretization can be considered: (proximal-type)

$$x_{k+1} = \arg\min_x \ f(x) + \frac{1}{2h}\|x - (x_k + \sqrt{2h} \cdot z_{k+1})\|_2^2$$

- One useful fact from optimization:

$$\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \frac{\alpha\beta}{\alpha + \beta}\|x - y\|^2 + \frac{1}{\alpha + \beta}\|\nabla f(x) - \nabla f(y)\|^2$$

Langevin Monte Carlo

- Euler-Maruyama Discretization with stepsize h:

$$x_{k+1} = x_k - h\nabla f(x_k) + \sqrt{2h} \cdot z_{k+1} \tag{2}$$

  converges to $\pi_h \neq \pi$ but $\pi_h \to \pi$ as $h \to 0$.
- Rate: $\mathcal{O}(\mathrm{poly}(\frac{1}{\epsilon}))$ w/o warm-start (dictated by the stepsize)
- Assumption: α-strong convexity + β-smoothness. Weaker assumption than strong-log-concavity based on isoperemtry inequality exists (Log-Sobolev, Poincaré, ...) but more technical.
- Other discretization can be considered: (proximal-type)

$$x_{k+1} = \arg\min_x \, f(x) + \frac{1}{2h}\|x - (x_k + \sqrt{2h} \cdot z_{k+1})\|_2^2$$

- One useful fact from optimization:

$$\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x-y\|^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|^2$$

- Guarantees in TV, KL, χ^2 also possible

Discrete Time Convergence (Sketch)

Again in W_2 metric, we do synchoronous coupling (same z using (2)).

$$
\begin{aligned}
&\|x_{k+1} - y_{k+1}\|_2^2 = \|x_k - y_k - h(\nabla f(x_k) - \nabla f(y_k))\|_2^2 \\
&= \|x_k - y_k\|_2^2 - 2h\langle x_k - y_k, \nabla f(x_k) - \nabla f(y_k)\rangle + h^2\|\nabla f(x_k) - \nabla f(y_k)\|_2^2
\end{aligned}
$$

Let $W_2^2(\rho_k, \rho_k') = \mathbb{E}[\|x_k - y_k\|_2^2]$ be optimal coupling and $h \leq \frac{2}{\alpha+\beta}$,

$$
\begin{aligned}
W_2^2(\rho_{k+1},\rho_{k+1}') &\leq \mathbb{E}[\|x_{k+1} - y_{k+1}\|_2^2] \\
&\leq (1 - \frac{2h\alpha\beta}{\alpha+\beta})\mathbb{E}[\|x_k - y_k\|_2^2] + h(h - \frac{2}{\alpha+\beta})\mathbb{E}[\|\nabla f(x_k) - \nabla f(y_k)\|_2^2] \\
&\leq (1 - \frac{2h\alpha\beta}{\alpha+\beta})W_2^2(\rho_k, \rho_k') \leq \exp(-\frac{2kh\alpha\beta}{\alpha+\beta})W_2^2(\rho_0, \rho_0')
\end{aligned}
$$

Conclusion: It has *unique* stationary dist π_h but starting from π will step away from π (next slide). Can show $W_2(\pi_h, \pi) = \mathcal{O}(h)$. Therefore to get $W_2(\rho_k, \pi) \leq W_2(\rho_k, \pi_h) + W_2(\pi_h, \pi) \leq \epsilon$ need $h = \mathcal{O}(\epsilon)$ and $k = \mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ iterations $\rightarrow$ exponential slowdown from cts time (1).

Simple Example on Asymptotic Bias for ULA

Take $x_0 \sim \rho_0 = \mathcal{N}(0, I_d)$ and $\nabla f(x) = x$ the quadratic potential:

$$x_{k+1} = x_k - hx_k + \sqrt{2h} \cdot z_k = (1-h)x_k + \sqrt{2h} \cdot z_k$$

i.e., Ornstein–Uhlenbeck process with $\pi \sim \mathcal{N}(0, I_d)$. Distribution evolves as

$$x_\infty \sim \rho_\infty = \mathcal{N}\left(0, 2h \cdot \sum_{i=0}^{\infty} (1-h)^{2i} \cdot I_d\right) \to \mathcal{N}\left(0, \frac{1}{1-h/2} \cdot I_d\right)$$

Hence for

$$\begin{aligned} W_2(\rho_\infty, \pi) &= \mathbb{E}\left[\left\| \frac{1}{\sqrt{1-h/2}} z - z \right\|_2^2\right]^{1/2} = \left| \frac{1}{\sqrt{1-h/2}} - 1 \right| \sqrt{d} \\ &\sim \frac{h}{4}\sqrt{d} \leq \epsilon \end{aligned}$$

Need to take $h = \mathcal{O}(\epsilon d^{-1/2})$.

Metropolis Hastings Adjustment (MALA)

To correct for bias we still use the Langevin proporsal but add an accept-reject step:

for $k = 1, \cdots, T$ **do**

  $\tilde{X}_{k+1} \sim \mathcal{N}(X_k - h\nabla f(X_k), 2h \cdot I)$

  $q(\tilde{X}_{k+1}|X_k) = \mathbb{P}(X_k \to \tilde{X}_{k+1}) = C \cdot \exp(-\frac{1}{4h}\|\tilde{X}_{k+1} - X_k + h\nabla f(X_k)\|_2^2)$

  Compute $\alpha \leftarrow \min\left\{1, \frac{\pi(\tilde{X}_{k+1})q(X_k|\tilde{X}_{k+1})}{\pi(X_k)q(\tilde{X}_{k+1}|X_k)}\right\}$

  Draw $U \sim \text{Unif}([0,1])$

  if $U \leq \alpha$ **then**

    $X_{k+1} = \tilde{X}_{k+1}$

  else

    $X_{k+1} = X_k$

  end if

end for

Remark: (1) right stationary distribution thanks to detailed balance (next slide); (2) no need for normalizing constant; (3) guarantee polylog(ϵ^{-1}) but usually need some warmness

Detailed Balance

Let $P(X, \tilde{X})$ denote the induced Markov Chain transition probabilities from state X to $\tilde{X}$ for MALA (wlog assume second term below is $\min$)

$$P(X, \tilde{X}) = \underbrace{q(\tilde{X}|X)}_{\text{proposal}} \cdot \underbrace{\min\left\{1, \frac{\pi(\tilde{X})q(X|\tilde{X})}{\pi(X)q(\tilde{X}|X)}\right\}}_{accept/reject} = \frac{\pi(\tilde{X})q(X|\tilde{X})}{\pi(X)}$$

and

$$P(\tilde{X}, X) = q(X|\tilde{X}) \cdot \min\left\{1, \frac{\pi(X)q(\tilde{X}|X)}{\pi(\tilde{X})q(X|\tilde{X})}\right\} = q(X|\tilde{X})$$

Therefore $\pi(X)P(X, \tilde{X}) = \pi(\tilde{X})P(\tilde{X}, X)$ for all $X, \tilde{X}$. This is the DB condition and ensures π is the stationary distribution:

$$\sum_X \pi(X)P(X, \tilde{X}) = \sum_X \pi(\tilde{X})P(\tilde{X}, X) = \pi(\tilde{X}) \sum_X P(\tilde{X}, X) = \pi(\tilde{X})$$

hence $\pi = \pi P$. Aside: Proposal distribution can be more general.

Connection to (Deterministic) Optimization

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{3}$$

- **[JKO '98]** Density $X_t \sim \rho_t$ along dynamics (3) follows gradient flow of minimizing KL divergence with Wasserstein-2 metric in the space of probability measures

$$\text{``}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{''}$$

Connection to (Deterministic) Optimization

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{3}$$

- [JKO '98] Density $X_t \sim \rho_t$ along dynamics (3) follows gradient flow of minimizing KL divergence with Wasserstein-2 metric in the space of probability measures

$$\text{``}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{''}$$

- Can show f strongly convex $\rightarrow$ KL functional strongly convex in density space $\Rightarrow$ linear convergence in continuous time $\mathcal{O}(\log(\frac{1}{\epsilon}))$ as observed earlier.

Connection to (Deterministic) Optimization

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{3}$$

- [JKO '98] Density $X_t \sim \rho_t$ along dynamics (3) follows gradient flow of minimizing KL divergence with Wasserstein-2 metric in the space of probability measures

$$\text{``}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{''}$$

- Can show f strongly convex $\rightarrow$ KL functional strongly convex in density space $\Rightarrow$ linear convergence in continuous time $\mathcal{O}(\log(\frac{1}{\epsilon}))$ as observed earlier.
- *Consequence:* Brownian motion as steepest descent for negative entropy functional $\int \rho \log \rho$ in density space:

$$dX_t = \sqrt{2}dW_t \rightarrow \dot{\rho}_t = \Delta \rho_t$$

Solution is $\rho_t \sim \mathcal{N}(x_0, 2tI)$ if $\rho_0 \sim \delta_{x_0}$.

Connection to (Deterministic) Optimization

$$dX_t = -\nabla f(X_t)dt + \sqrt{2}dW_t \tag{3}$$

- [JKO '98] Density $X_t \sim \rho_t$ along dynamics (3) follows gradient flow of minimizing KL divergence with Wasserstein-2 metric in the space of probability measures

$$\text{“}\dot{\rho}_t = -\nabla_{W_2} KL(\rho_t \| \pi)\text{”}$$

- Can show f strongly convex $\rightarrow$ KL functional strongly convex in density space $\Rightarrow$ linear convergence in continuous time $\mathcal{O}(\log(\frac{1}{\epsilon}))$ as observed earlier.
- *Consequence:* Brownian motion as steepest descent for negative entropy functional $\int \rho \log \rho$ in density space:

  $$dX_t = \sqrt{2}dW_t \rightarrow \dot{\rho}_t = \Delta \rho_t$$

  Solution is $\rho_t \sim \mathcal{N}(x_0, 2tI)$ if $\rho_0 \sim \delta_{x_0}$.
- Optimization as sampling: take temperature to ∞

Underdamped Langevin

Introduce auxiliary variable à la ”Momentum” from optimization:

$$\begin{aligned} dX_t &= V_t dt \\ dV_t &= -\nabla f(X_t)dt - \underbrace{\gamma V_t}_{\text{friction}} dt + \sqrt{2\gamma} dW_t \end{aligned}$$

Can check invariant measure $\pi(X, V) \propto e^{-f(x) - \frac{1}{2}\|v\|^2}$ so take the marginal gives the desired $X \sim \pi$.

Naive discretization wouldn’t work but SOTA scheme gives improvement. In some sense the second-order dynamics with Brownian motion term in the auxiliary variable eases discretization.

Parting Thoughts

- Mostly focused on convergence analysis
  - touches on {probability, numerical analysis, optimization, PDE, optimal transport, physics … }
  - Other algorithms: Hamiltonian Monte Carlo, Stein Variational GD, Gibbs Sampler, Riemannian Manifold Langevin, Schrödinger bridge, Zig-Zag sampler, 0th-order method (hit-and-run, ball walk) …
- References for MCMC Algorithms: (+ practical guidance)
  - Jun Liu, “Monte Carlo Strategies in Scientific Computing”
  - “Handbook of Markov Chain Monte Carlo”
- Software Packages: Stan, TensorFlow Probability, …
- Wasserstein GF as an analysis tool also features *prominently* in mean-field analysis of e.g., Neural Networks (cf. Chizat-Bach ’18, Mei-Montanari-Nguyen ’18)

Thanks! Questions?