CPSC 486/586: Probabilistic Machine Learning March 1st, 2023

Lecture 14

Lecturer: Andre Wibisono *Scribe: Alec Xiang*

1 Metropolis-Hastings Algorithm (continued)

1.1 Definitions

Given any proposal distribution $Q = \{Q_x : x \in R^d\}$ and that we want to sample from a target distribution $\nu \in \mathcal{P}(\mathbb{R}^d)$, we will output $\hat{Q} = \{\hat{Q}_x : x \in R^d\}$ such that $\hat{Q}$ is reversible with respect to ν. Recall that reversibility means $\nu(x)\hat{Q}_X(y) = \nu(y)\hat{Q}_y(x) \forall x, y \in \mathbb{R}^d$ and that it implies ν is stationary for $\hat{Q}$, i.e. $\nu(x) = \int_{\mathbb{R}^d} \nu(y)\hat{Q}_y(x)dy$ (which can be interpreted "if you start at ν, after one step you're still at ν").

Note that a proposal distribution (transition probability) is $Q = \{Q_x : x \in R^d\}$ where Q_x is a probability distribution on $\mathbb{R}^d$ that represents our next distribution ρ_1, if we start from $\rho_0 = \delta_x$ and jump according to Q. Let's write this as a map:

$$Q_\sharp(\delta_x) = Q_x$$

where $Q_\sharp : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$. $\rho_1 = Q_\sharp(\rho_0)$ is given by the following:

$$X_0 \sim \rho_0$$

$$X_1 \mid X_0 \sim Q_{X_0}$$

then $X_1 \sim \rho_1 = Q_\sharp(\rho_0)$. So any given Q defines a Markov chain:

$$X_0 \sim \rho_0 \to X_1 \sim \rho_1 \to X_2 \sim \rho_2$$

where $X_0 \sim \rho_0$ (arbitrary) and $X_{k+1} \mid X_k \sim Q_{X_k}(x_{k+1}) = p(x_{k+1} \mid x_k)$—then $X_k \sim \rho_k$, where $\rho_{k+1} = Q_\sharp(\rho_k)$. Typically, as $k \to \infty$, ρ_k will converge to some distribution, i.e. $\rho_k \to \rho_\infty$ where ρ_∞ is stationary with respect to Q:

$$Q_\sharp(\rho_\infty) = \rho_\infty$$

We are typically interested in how fast $\rho_k \to \rho_\infty$ (faster is better, because it allows us to run the Markov chain for less time to obtain one sample).

1.2 Constructing Q such that ν is stationary with respect to Q

The goal now is to construct $Q = \{Q_x : x \in \mathbb{R}^d\}$ such that ν (our target distribution) is stationary with respect to Q and the Markov chain $X_k \to X_{k+1} \to ...$ converges fast to ν, $\lim_{k\to\infty} d(\rho_k, \nu) = 0$. Note that this defines an algorithm to sample from ν:

1. Start from $X_0 \sim \rho_0$
2. Sample: $X_{k+1} \mid X_k \sim Q_{X_k}(X_{k+1})$ for $k = 0, 1, ..., T-1$
3. Output $X_t \sim \rho_T$

From now until the end of the semester, we will look at a lot of different Markov chains and look at how fast they converge.

What if we have Q that converges to some $\rho_\infty \neq \nu$? e.g. $Q_X = \mathcal{N}(x, \eta I), \eta > 0$. That's why we use the MH algorithm (Metropolis filter) $\hat{Q}$, which gives us reversibility with respect to ν ("detailed balance"), by "forcing" Q to be reversible with respect to ν (through the accept/reject step on top of Q). Let $\hat{Q} = MH(Q, \nu)$ given by:

1. From $x \in \mathbb{R}^d$, draw $y \mid x \sim Q_x$ (original proposal)
2. Set $x_{new} = y$ with probability $A(x, y)$, $x_{new} = x$ with probability $1 - A(x, y)$. Here,

$$A(x, y) = \min\left\{1, \frac{\nu(y)Q_y(x)}{\nu(x)Q_x(y)}\right\}.$$

1.3 MH is "optimal"

We can check that $\hat{Q} = MH(Q, V)$ is reversible with respect to ν (we did this last time). In fact, MH is a projection of Q to the space of ν-reversible Markov chains. Let $\mathcal{R}(\nu) = \{Q \in \mathcal{Q} :$ Q is reversible with respect to $\nu\}$. Define $d(Q, Q') = \int_{\mathbb{R}^d \times \mathbb{R}^d \setminus \text{diag}} |Q_x(y) - Q'_x(y)|\nu(x)dxdy$ where diag $= \{(x, x) : x \in \mathbb{R}^d\}$. Then we have the theorem from [Billera, Diaconis '01]:

$$\boxed{\hat{Q} = MH(Q, \nu) \in \underset{R\in\mathcal{R}(\nu)}{\arg\min}\, d(Q, R)}$$

but you can also see Chewi, §7.2, 7.3.

2 Other choices for Q besides $Q_x = \mathcal{N}(x, 2\eta I)$

1. $Q_x = \mathcal{N}(x, 2\eta I)$ (Gaussian Noise) $\implies$ $\hat{Q} = MH(Q, \nu)$ is the Metropolis Random Walk (MRW) algorithm

2. $Q_x = \mathcal{N}(x - \eta\nabla f(x), 2\eta I)$, $y = x - \eta\nabla f(x) + \sqrt{2\eta}Z$ (Unadjusted Langevin Algorithm) with an independent $Z \sim \mathcal{N}(0, I) \implies \hat{Q} = MH(Q, \nu)$ is the MALA (Metropolis-Adjusted Langevin Algorithm)

Recall that the MRW algorithm to sample from $\nu(x) \propto \exp(-f(x))$ is:

1. Start $x_0 \sim \rho_0$
2. For $k = 0, 1, ..., T-1$:
   (a) $y_k = x_k + \sqrt{2\eta}Z_k$, $Z_k \sim \mathcal{N}(0, I)$ indep.
   (b) $x_{k+1} = y_k$ with probability $A(x_k, y_k)$, $x_{k+1} = x_k$ with probability $1 - A(x_k, y_k)$ where $A(x, y) = \min\{1, \frac{\nu(y)}{\nu(x)}\}$ (because $Q_x(y) = Q_y(x)$).

Since the normalizing constants cancel, we have $A(x, y) = \min\{1, \exp(f(x) - f(y))\}$—so if $f(y_k) \leq f(x_k)$, always accept, but otherwise you accept with probability $\exp(-(f(y_k) - f(x_k)))$. Let's try to plug in a different Q.

2.1 MALA (Metropolis-Adjusted Langevin Algorithm)

Algorithm:

1. $X_0 \sim \rho_0$
2. For $k = 0, 1..., T-1$
   (a) $y_k = x_k - \eta\nabla f(x_k) + \sqrt{2\eta}Z_k$ where $Z_k \sim \mathcal{N}(0, I)$
   (b) $x_{k+1} = y_k$ with probability $A(x_k, y_k)$, $X_{k+1} = x_k$ with probability $1 - A(x_k, y_k)$

where $A(x, y) = \min\{1, \frac{\nu(y)Q_y(x)}{\nu(x)Q_x(y)}\}$—we can compute this in practice. Note that MRW is called a 0-order algorithm (only needs $\nu(x)$ up to a constant), whereas MALA is called a 1st-order algorithm (needs $\nu(x)$ and also $\nabla \log \nu(x)$, i.e. $-f(x), -\nabla f(x)$). So if you assume that we know this gradient (may or may not be difficult depending on your problem), you can show that MALA works better than MRW (for some nice class of distributions).

2.2 Convergence guarantees

Assume that $\nu(x) \propto \exp(-f(x))$ where f is strongly convex and smooth, so $\alpha I \preceq \nabla^2 f(x) \preceq LI$. Define the condition number:

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

Recall that chi-square divergence is $\chi^2(\rho||\nu) = \mathrm{Var}_\nu(\frac{\rho}{\nu}) = \int \nu(x)(\frac{\rho(x)}{\nu(x)} - 1)^2 dx$:

Theorem 1 (If we start from computable ρ_9, i.e. feasible/easy)**.** *If $\rho_0 = \mathcal{N}(x^*, \frac{1}{L}I)$, then to get ρ_T with $\sqrt{\chi^2(\rho_t||\nu)} \leq \epsilon$, (a) MRW needs $T = \tilde{O}(\kappa^2 d \cdot polylog\frac{1}{\epsilon})$ and (b) MALA needs $T = \tilde{O}(\kappa d \cdot polylog\frac{1}{\epsilon})$, i.e. MALA is better by κ.*

Theorem 2 (Warm start)**.** *If ρ_0 with $\sqrt{\chi^2(\rho_0||\nu)} = O(1)$ then to get to $\sqrt{KL(\rho_T||\nu)} \leq \epsilon$, MALA needs $T = \tilde{O}(\kappa d^{\frac{1}{2}} polylog\frac{1}{\epsilon})$*

That is, under the feasible start case, going from MRW to MALA gives us a factor of κ improvement. Going from feasible start to warm start for MALA gives us a $d^{\frac{1}{2}}$ improvement.