CPSC 486/586: Probabilistic Machine Learning	February, 27, 2023
Lecture 13	
Lecturer: Andre Wibisono	*Scribe: Tatsuhiro Shimizu*

1 Instruction

In the last lecture, we have seen the rejection sampling which can take exponential time to sample from the target distribution $\nu(x)$ in high dimensions. Today's lecture tries to improve the sampling speed by introducing Metropolis Random Walk and Metropolis-Hasting Algorithm.

2 Metropolis Random Walk (MRW)

Suppose we have target distribution $\nu(x) \propto e^{-f(x)}$.

Algorithm 1 Metropolis Random Walk (MRW)

Given: We start from $x_0 \sim \rho_0$ on $\mathcal{X}$, given step size $\eta > 0$.
for $k = 0, 1, \cdots, K$ **do**:

$$y_k \leftarrow x_k + \sqrt{\eta} z_k, \quad z_k \sim \mathcal{N}(0, I)$$

Accept y_k with probability

$$\min\left\{1, \frac{\nu(y_k)}{\nu(x_k)}\right\}.$$

if Accept **then**

$$x_{k+1} = y_k$$

else

$$x_{k+1} = x_k$$

end if
end for
Return $x_K \sim \rho_K$.

Note that $x_k \in \mathbb{R}^d$ is the random variable with distribution $\rho_k \in P(\mathbb{R}^d)$.

Example 1 (Ball Walk)**.** *Suppose target ν is the uniform distribution on convex body $\mathcal{X} \subset \mathbb{R}^d$.*

Our goal is to sample uniformly form $\mathcal{X}$. It is mainly applicable for estimating the volume of $\mathcal{X}$ practically.

We sample by using MRW. For $k = 0, 1, \cdots, K$, Let

$$y_k = x_k + \sqrt{\eta} z_k \quad z_k \sim \mathcal{N}(0, I)$$

If $y_k \in \mathcal{X}$, then set $x_{k+1} = y_k$. Otherwise, set $x_{k+1} = x_k$.

For this example, we can say that the sampling speed is more efficient than rejection sampling but to show that, we introduce the notion of isotropic.

Definition 1. *A convex body $\mathcal{X} \subseteq \mathbb{R}^d$ is in isotropic position if its center of mass is at 0, i.e.,*

$$\mathbb{E}_\nu[X] = 0$$

and its covariance matrix is the identity

$$\mathrm{Cov}_\nu[X] = I$$

where ν here is the uniform distribution over $\mathcal{X}$. Equivalently, for any unit vector $\|v\| = 1$,

$$\frac{1}{\mathsf{Vol}(\mathcal{X})} \int_{\mathcal{X}} (v^\top x)^2 dx = 1.$$

In general, we can say that a distribution ν (not necessarily the uniform distribution) is in isotropic position if it satisfies the above conditions, where the expectations are taken with respect to ν (e.g., see Bertsimas and Vempala 2004; Rudelson 1999). Then, using the notion of isotropic, we have the following statement about the sampling speed.

Theorem 1. *If $\mathcal{X}$ is isotropic, then the MRW is a polytime algorithm.*

For ball walk to produce ρ_k with $TV(\rho_k, \nu) \leq \epsilon$, we need $K = \tilde{O}(d^2)$ where $\tilde{O}$ hides the logarithm terms and warm start, that is $\sup_{x \in \mathcal{X}} \frac{\rho_0(x)}{\nu(x)} \leq M$ where M is a constant.

3 Metropolis-Hastings Algorithm (MH)

We now introduce a generalization of MRW, called the **Metropolis-Hastings** algorithm. Suppose we have a target distribution ν on $\mathcal{X}$. In this algorithm, we need to choose $p(x \mid y)$.

Algorithm 2 Metropolis-Hasting Algorithm (MH)

Given: We start from any $x_0 \sim \rho_0$ on $\mathcal{X}$.
for $k = 0, 1, \cdots, K$ **do**:
Draw
$$y_k \mid x_k \sim p(y_k \mid x_k)$$
Accept y_k with probability
$$\min\left\{1, \frac{\nu(y_k)p(x_k \mid y_k)}{\nu(x_k)p(y_k \mid x_k)}\right\}$$
end for
Return x_K

Note that MRW is a special case of the MH algorithm where

$$p(y \mid x) = \mathcal{N}(x, \eta I)(y)$$

$$p(y \mid x) = p(x \mid y)$$

Question: why does the MH algorithm work, that is $x_K \sim \rho_K \to \nu$ as $k \to \infty$? The short answer is that this Markov Chain is reversible on $\mathcal{X}$. To answer the question mathematically, we first check the definition of the Markov Chain.

Definition 2. *Markov Chain (MC) on $\mathcal{X}$ is specified by a family of probability distributions*

$$Q = \{Q_x = p(\cdot|x) \quad | \, x \in \mathcal{X}\}$$

Then, given $x_0 \in \rho_0$ on $\mathcal{X}$, we can get Markov Chain $X_0, X_1, X_2, \cdots$ where

$$x_{k+1} \mid x_k \sim Q_{x_k}(x_{k+1}) = p(x_{k+1} \mid x_k).$$

Thus, MC $Q = \{Q_x : x \in \mathcal{X}\}$ defines a map

$$Q : p(\mathcal{X}) \to p(\mathcal{X})$$

by

$$\rho_{k+1}(y) = \int_{\mathcal{X}} \rho_k(x)p(y|x)dx$$

Now we can introduce the stationary and reversible properties.

Definition 3. *A probability distribution ν is stationary for the MC defined by*

$$Q = \{Q_x : x \in \mathcal{X}\}$$

if

$$Q(\nu) = \nu,$$

equivalently, that is

$$\nu(y) = \int_{\mathcal{X}} \nu(x)p(y|x)dx$$

Definition 4. *A MC Q is reversible with respect to some $\nu \in P(\mathbb{R}^d)$ if*

$$\nu(x)p(y|x) = \nu(y)p(x|y) \quad \forall x, y \in \mathcal{X},$$

which is equivalent to that the two joint distributions are symmetric

$$\rho(x,y) = \rho(y,x) \quad \rho(x,y) := \nu(x)p(y|x) \quad \forall x, y \in \mathcal{X}$$

Then we have the lemma about the relationship between two properties.

Lemma 1. *If Q is reversible with respect to ν, then ν is stationary for Q.*

Proof. By reversibility, for all $x, y \in \mathcal{X}$, we have

$$\nu(x)p(y \mid x) = \nu(y)p(x)$$

Integrating over $\mathcal{X}$, we have

$$\begin{aligned}
\int_{\mathcal{X}} \nu(x)p(y \mid x)dx &= \int_{\mathcal{X}} \nu(y)p(x \mid y)dx \\
\int_{\mathcal{X}} \nu(x)p(y \mid x)dx &= \nu(y) \int_{\mathcal{X}} p(x \mid y)dx \\
Q(\nu)(y) &= \nu(y)
\end{aligned}$$

□

Then we have a lemma about the properties of MH algorithm.

Lemma 2. *Let $Q = \{Q_x = p(\cdot \mid x; \quad x \in \mathcal{X})\}$ be an arbitrary MC. If we apply MH to get another MC $\hat{Q}$, then $\hat{Q}$ is reversible with respect to ν.*

We define $\alpha_x(y)$ as the acceptance probability in MH algorithm.

$$\alpha_x(y) = \min\left\{1, \frac{\nu(y_k)p(x_k \mid y_k)}{\nu(x_k)p(y_k \mid x_k)}\right\}$$

Now our question is what $\hat{Q}$ is.

$$\hat{Q}_x(y) = \underbrace{Q_x(y)}_{\text{original MC}} \times \underbrace{\alpha_x(y)}_{\text{acceptance probability}} + \underbrace{\delta_x(y)}_{\text{indicator}} \times \underbrace{A(x)}_{\text{probability distribution on } \mathcal{X}}$$

where

$$\begin{aligned}
A(x) &= Q_x(x) + \int_{\mathcal{X}\setminus\{x\}} (1 - \alpha_x(y))Q_x(y)dy \\
&= 1 - \int_{\mathcal{X}\setminus\{x\}} \alpha_x(y)Q_x(y)dy
\end{aligned}$$

Lemma 3. *Given MC Q, let $\hat{Q} = MH_\nu(Q)$. Then ν is reversible with respect to $\hat{Q}$. Thus, ν is stationary for $\hat{Q}$.*

Proof. We want to show that

$$\nu(x)\hat{Q}_x(y) = \nu(y)\hat{Q}_y(x)$$

If $x = y$, then the lemma holds obviously, so suppose $x \neq y$. Then,

$$\begin{aligned}
\nu(x)\hat{Q}_x(y) &= \nu(x)Q_x(y)\alpha_x(y) \\
&= \nu(x)Q_x(y)\min\left\{1, \frac{\nu(y)Q_y(x)}{\nu(x)Q_x(y)}\right\} \\
&= \min\{\nu(x)Q_x(y), \nu(y)Q_y(x)\} \\
&= \nu(y)\hat{Q}_y(x)
\end{aligned}$$

$\square$

References

Bertsimas, Dimitris and Santosh Vempala (July 2004). "Solving Convex Programs by Random Walks". In: *J. ACM* 51.4, pp. 540–556. ISSN: 0004-5411. DOI: 10.1145/1008731.1008733. URL: https://doi.org/10.1145/1008731.1008733.

Rudelson, M. (1999). "Random Vectors in the Isotropic Position". In: *Journal of Functional Analysis* 164.1, pp. 60–72. ISSN: 0022-1236. DOI: https://doi.org/10.1006/jfan.1998.3384. URL: https://www.sciencedirect.com/science/article/pii/S0022123698933845.