CPSC 661: Sampling Algorithms in Machine Learning February 3, 2021

Lecture 2

Lecturer: Andre Wibisono *Scribe: Linyong Nan*

1 Outline

Today's lecture is about Markov chain and it covers three main parts:

- Measure theory formalism of Markov chain
- Stationary distribution and reversibility of Markov chain
- Distance between distributions and mixing time

2 Measure Theory Formalism of Markov Chain

We first introduce the measure theory as a tool to describe and formalize problems of Markov chain that we will analyze in this lecture.
Let space $\mathcal{X}$ be a set with a σ-algebra $\mathcal{A} \subseteq 2^{\mathcal{X}} = \{\text{all subset of } \mathcal{X}\}$. We want to note that for $\mathcal{X}$ that has a σ-algebra, it means:

- $\mathcal{X} \in \mathcal{A}$, $\mathcal{A}$ is closed under complement and countable union
- Elements of $\mathcal{A}$ are measurable subsets of $\mathcal{X}$

Measure A measure on $\mathcal{X}$ is a function $\nu : \mathcal{A} \to [0, \infty]$. Note that:

- $\nu(\emptyset) = 0$ and ν is countably additive, which means that for a set $A = \cup_{i=1}^{\infty} A_i$ (union of disjoint subsets), $\nu(A) = \sum_{i=1}^{\infty} \nu(A_i)$
- ν is a *probability measure* if $\nu(\mathcal{X}) = 1$

Consider an example in discrete space: let $\mathcal{X} = \{1, \ldots, n\}, \mathcal{A} = 2^{\mathcal{X}}$, and consider the following two measures:

- ν = Counting: $\nu(A) = |A|$ (number of elements in A). Note that this is not a probability measure since $\nu(A) = n$.
- ν = Uniform: $\nu(A) = \frac{|A|}{n}$, and this is a probability measure.
- ν = Point mass at $x^* : \nu(A) = \mathbf{1}\{x^* \in A\}$ (1 if $x^* \in A$ else 0)

Consider an example in continuous space: let $\mathcal{X} = \mathbb{R}^n$, $\mathcal{A}$=Borel σ-algebra (generated by open sets), then we have the following measures:

- ν = Lebesgue: $\nu(A) = \text{Vol}(A)$. Note that this is not a probability measure since $\nu(\mathbb{R}^n) = \infty$.
- ν = Lebesgue on K: $\nu(A) = \text{Vol}(A \cap K)$, for $K \subseteq \mathbb{R}^n$.
- ν = Uniform on K: $\nu(A) = \frac{\text{Vol}(A \cap K)}{\text{Vol}(K)}$.
- ν = Gaussian: $\nu(A) = \int_A (2\pi)^{-n/2} e^{-\frac{||x||^2}{2}} dx$
- ν = Point mass at $x^* : \nu(A) = \mathbf{1}\{x^* \in A\}$ (1 if $x^* \in A$ else 0)

Density Assume there is a reference measure μ on space $\mathcal{X}$, and that

- $\mathcal{X} = \{1, \ldots, n\} : \mu$ = counting measure
- $\mathcal{X} = \mathbb{R}^n : \mu$ = Lebesgue measure

We write a measure ν as a density (Radon-Nikodym derivative) with respect to μ:

$$h = \frac{d\nu}{d\mu} \Leftrightarrow \nu(A) = \int_A h(x) d\mu(x)$$

- Requires $\nu \ll \mu$: if $\mu(A) = 0$, then $\nu(A) = 0$
- We usually also denote the density by ν itself

We revisit the above examples in discrete and continuous space, and here are some examples of densities:

1. $\mathcal{X} = \{1, \ldots, n\}, \mu$ = counting measure
   - ν = Uniform: $\nu(x) = \frac{1}{n}$
   - ν = Point mass:
   $$\delta_{x^*}(x) = \begin{cases} 1 & \text{if } x = x^* \\ 0 & \text{else} \end{cases}$$
2. $\mathcal{X} = \mathbb{R}^n, \mu$ = Lebesgue measure
   - ν = Uniform on $K : \nu(x) = \frac{1}{\text{Vol}(K)}$, for $K \subseteq \mathbb{R}^n$
   - ν = Gaussian: $\nu(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{||x||^2}{2}}$
   - ν = Point mass:
   $$\delta_{x^*}(x) = \begin{cases} \infty & \text{if } x = x^* \\ 0 & \text{else} \end{cases}$$
   where $\int_{\mathbb{R}^n} \delta_{x^*}(x) dx = 1$, $\int_{\mathbb{R}^n} \delta_{x^*}(x) \cdot f(x)\, dx = f(x^*)$

2.1 Markov Chain

We can now describe Markov chain with the above formalism and the concept of random walk.

Random Walk Let $\mathcal{X}$ be the state space. A **random walk** on $\mathcal{X}$ is a sequence of random variables

$$X_0, X_1, X_2, \ldots \in \mathcal{X}$$

where the future depends on the past and present.

Markov Chain A **Markov chain** is a random walk where the future is **independent** of the past given the present:

$$X_{t+1} \perp\!\!\!\perp (X_0, \ldots, X_{t-1}) \mid X_t$$

$$X_{t+1} \mid \{X_0, \ldots, X_{t-1}, X_t\} \stackrel{d}{=} X_{t+1} \mid X_t$$

The distribution of X_{t+1} is determined only by X_t:

$$X_{t+1} \mid X_t \sim P_{X_t}(\cdot)$$

We want to note that this definition is for a first-order (only remember 1 previous state), time-homogeneous (transition probability does not change) Markov chain.

Formally, a **Markov chain** on $\mathcal{X}$ is a collection

$$P = (P_x : x \in \mathcal{X})$$

where each P_x is a probability measure on $\mathcal{X}$, and if we start at $X_0 = x$, then P_x is the distribution of the next state X_1:

$$P_x(A) = Pr(X_1 \in A \mid X_0 = x)$$

We can say that P maps an initial distribution $X_0 \sim \rho_0$ to the next distribution $X_1 \sim \rho_1$:

$$\rho_1(A) = \int_{\mathcal{X}} P_x(A)\ d\rho_0(x)$$

Therefore P defines a Markov chain

$$X_0 \to \cdots \to X_k \to X_{k+1} \to \cdots$$

with the distribution of $X_k \sim \rho_k$:

$$\rho_k(A) = \int_{\mathcal{X}} P_x(A)\ d\rho_{k-1}(x)$$

3 Stationary Distribution and Reversibility

One category of Markov chains that we are interested in studying is the stationary Markov chain that has a stationary distribution ν for P:

$$\nu(A) = \int_{\mathcal{X}} P_x(A)\ d\nu(x)$$

If $X_0 \sim \nu$, then $X_1 \sim \nu$ and all $X_k \sim \nu$. Existence and uniqueness of the stationary distribution are followed from irreducibility and recurrence, which always hold within the scope of this lecture. More discussions on existence and uniqueness can be found in Reference [1, 2, 5].

3.1 Discrete Space

In a discrete setting, a Markov chain P on $\mathcal{X} = \{1, \ldots, n\}$ can be represented as a stochastic matrix $P \in \mathbb{R}^{n \times n}$ with:

- non-negative entries: $P(x, y) = P_x(y) \geq 0$
- row sums to 1: $\sum_{y \in \mathcal{X}} P(x, y) = 1$

A probability distribution ρ on $\mathcal{X}$ is a row vector

$$\rho = [\rho(x_1) \cdots \rho(x_n)] \in \mathbb{R}^n$$

Therefore an operation of Markov chain $X_{k-1} \to X_k$ is a right multiplication by P:

$$\rho_k = \rho_{k-1} P$$

In discrete setting, a stationary distribution ν is a left eigenvetor of P:

$$\nu = \nu P$$

Existence and uniqueness of ν is guaranteed by *Perron-Frobenius theorem.*
Next we describe the stationary condition and reversible condition as follows:

- Stationary condition $\Leftrightarrow$ global balance:

$$\forall x : \sum_{y \in \mathcal{X}} \nu(x) P(x, y) = \nu(x) = \sum_{y \in \mathcal{X}} \nu(y) P(y, x)$$

- Reversible condition $\Leftrightarrow$ detailed balance (stronger):

$$\forall x, y : \nu(x) P(x, y) = \nu(y) P(y, x)$$

3.1.1 Example

We consider a discrete example of random walk on Graph. Let $G = (V, E, w)$ be an undirected, weighted graph. Note that for a random walk on G, we start at any vertex $x \in V$. At each time, jump to a neighbor y with probability $\propto w(x, y)$. With this setting, we can define the following Markov chain P:

$$P_x(y) = \frac{w(x, y)}{\deg(x)}$$

where $\deg(x)$ is the total weight of edges from the neighbors of x. We can also write the stationary distribution ν which is proportional to the degree distribution:

$$\nu(x) \propto \deg(x) = \sum_{y \in \mathcal{X}} w(x, y)$$

This stationary distribution also satisfies detailed balance (P is reversible with respect to ν), because:

$$D = \sum_x \deg(x) = 2 \sum_{(x,y) \in E} w(x, y)$$

$$\frac{\deg(x)}{D} \cdot \frac{w(x, y)}{\deg(x)} = \frac{w(x, y)}{D} = \frac{w(y, x)}{D} = \frac{\deg(y)}{D} \cdot \frac{w(y, x)}{\deg(y)}$$

Conversely, we can make the following statements:

- Any Markov chain on a finite space $\mathcal{X}$ can be written as a random walk on a *directed*, weighted graph
- Any reversible Markov chain on a finite space $\mathcal{X} = \{1, \cdots, n\}$ can be written as a random walk on an *undirected*, weighted graph. Given reversible condition, we have:

  $$w(x, y) = \nu(x) \cdot P(x, y) \overset{Rev.}{=} \nu(y) \cdot P(y, x) = w(y, x)$$

  $$\deg(x) = \sum_y w(x, y) = \sum_y \nu(x) \cdot P(x, y) = \nu(x)$$

  So we can define the Markov chain P:

  $$P(x, y) = \frac{w(x, y)}{\deg(x)}$$

3.1.2 Non-Example

We consider the **Non-Backtracking walk on graph**, which is **not** a first-order Markov chain. At time k from $X_k \in V$: jump to a neighbor X_{k+1} which is not the previous state X_{k-1}. Note that:

- This is a *second-order* Markov chain

- It can be written as a first-order Markov chain by enlarging space ($\mathcal{X} = V \times V$), but it loses reversibility
- Second-order/non-reversible Markov chains can mix faster. More discussions of this result can be found in Reference [6, 7].

3.2 Continuous Space

In a continuous setting, a Markov chain P on $\mathcal{X} = \mathbb{R}^n$ is written as a transition density $P : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with:

- non-negative entries: $P(x, y) = P_x(y) = P(y \mid x) \geq 0$
- row sums to 1: $\int_{\mathbb{R}^n} P(x, y)\ dy = 1$

A probability distribution ρ on $\mathbb{R}^n$ is $\rho : \mathbb{R}^n \to [0, \infty]$ with:

$$\int_{\mathbb{R}^n} \rho(x)\ dx = 1$$

Therefore an operation of Markov chain $X_{k-1} \to X_k$ is a linear functional:

$$\rho_k(y) = \int_{\mathbb{R}^n} \rho_{k-1}(x) P(x, y)\ dx$$

In continuous setting, a stationary distribution ν is a *left eigenfunction* of P:

$$\nu(y) = \int_{\mathbb{R}^n} \nu(x) P(x, y)\ dx$$

Similarly, we can describe the following conditions:

- Stationary condition $\Leftrightarrow$ global balance:

$$\int_{\mathbb{R}^n} \nu(y) P(y, x)\ dx = \int_{\mathbb{R}^n} \nu(x) P(x, y)\ dy$$

- Reversible condition $\Leftrightarrow$ detailed balance:

$$\nu(y) P(y, x) = \nu(x) P(x, y)$$

3.2.1 Example

Let $\eta > 0$ be constant (step size), $\mathcal{X} = \mathbb{R}^n$.

Brownian motion (Gaussian walk): From X, move to

$$Y = X + \sqrt{2\eta}Z$$

where $Z \sim \mathcal{N}(0, I)$ is independent. In this setting, we can define Markov chain $P_x = \mathcal{N}(x, 2\eta I)$ with density

$$P_x(y) = \frac{1}{(4\pi\eta)^{n/2}} \exp(-\frac{||y-x||^2}{4\eta})$$

It has the stationary: Lebesgue measure $\nu(x) = 1$, and it is also reversible (because symmetry: $P_x(y) = P_y(x)$).

Ornstein-Uhlenbeck: From X, move to

$$Y = e^{-\eta}X + \sqrt{(1-e^{-2\eta})}Z$$

where $Z \sim \mathcal{N}(0, I)$ is independent. In this setting, we can define Markov chain $P_x = \mathcal{N}(e^{-\eta}x, (1 - e{-}2\eta)I)$ with density

$$P_x(y) = \frac{1}{(2\pi(1-e^{-2\eta}))^{n/2}} \exp(-\frac{||y-e^{-\eta}x||^2}{2(1-e^{-2\eta})})$$

It has the stationary Gaussian distribution $\nu = \mathcal{N}(0, I)$ which is also reversible:

$$\nu(x) = \frac{1}{(2\pi)^{n/2}} \exp(-\frac{||x||^2}{2})$$

4 Distance between Distributions and Mixing Time

We are interested in how fast (mixing time) distributions will converge to a stationary measure. Formally, given a Markov chain $X_0 \to X_1 \to \cdots X_k \to \cdots$, $X_k \sim \rho_k$, we first choose a distance/divergence d between distributions. We want to note that there are many ways to measure distance between distributions ρ and ν:

- Total Variation (TV) distance ($L^1(\nu)$)
- χ^2-divergence ($L^2(\nu)$)
- KL divergence (relative entropy)
- more including f-divergence, Wasserstein distance, Rényi divergence, ...

Then we define the mixing time as:

$$\tau(\epsilon) = \inf\{K : d(\rho_k, \nu) \leq \epsilon, \forall k \geq K\}$$

Typically, we see exponential convergence:

$$d(\rho_k, \nu) \leq \alpha^k, \text{ for } \alpha < 1$$

so that dependence on ϵ is $\log(1/\epsilon)$. We want to study the dependence of mixing time on other parameters, such as problem size/dimension, target distribution and initial distribution.

4.1 Example: Card Shuffles

Let $\mathcal{X} = S_d = \{\text{permutations on d cards}\}$, then $|\mathcal{X}| = d! \sim d^d$, and the target distribution ν is uniform distribution. Then we have the mixing time results (in Total Variation distance) for the following two Markov chains that can be described by two famous card shuffle strategies:

- Top-to-random shuffle $\sim d \log d$
- Riffle shuffle $\sim \frac{3}{2} \log_2 d$

More details of these results can be found in Reference [8].

Reference

[1] Aldous and Fill, *Reversible Markov Chains and Random Walks on Graphs*, 2002

[2] Levin, Peres, and Wilmer, *Markov Chains and Mixing Times*, AMS, 2009

[3] Lov´asz and Simonovits, *Random Walks in a Convex Body and an Improved Volume Algorithm*, Random Structures and Algorithms, 1993

[4] Vempala, *Geometric Random Walk: A Survey, Combinatorial and Computational Geometry*, 2005

[5] Diaconis and Freedman, *On Markov Chains with Continuous State Space*, 1997

[6] Alon et al., *Non-backtracking random walks mix faster*, Communications in Contemporary Mathematics, 2007

[7] Diaconis and Miclo, *On the spectral analysis of second-order Markov chains*, 2013

[8] Bayes and Diaconis, *Trailing the Dovetail Shuffle to Its Lair*, Annals of Applied Probability, 1992