STATS 217: Introduction to Stochastic Processes I

Lecture 15

Period of a state

- Let P be the transition matrix of a DTMC on S.
- For a state $x \in S$, let

$$\mathcal{T}(x) := \{t \geq 1 : P^t_{x,x} > 0\}$$

  denote the set of times when it is possible for the chain to return to its starting position x.

Period of a state

- Let P be the transition matrix of a DTMC on S.
- For a state $x \in S$, let

$$\mathcal{T}(x) := \{t \geq 1 : P_{x,x}^t > 0\}$$

  denote the set of times when it is possible for the chain to return to its starting position x.
- The **period** of $x \in S$ is defined to be the greatest common divisor (gcd) of $\mathcal{T}(x)$.

Example

$$\begin{tikzcd} A \arrow[r, bend left, "\text{w.p. } 1"] & B \arrow[l, bend left, "\text{w.p. } 1"] \end{tikzcd}$$

Two-state Markov chain with the transition matrix

$$P = \begin{array}{c c} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

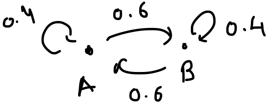
Example

Two-state Markov chain with the transition matrix

$$P = \begin{array}{c c} & \begin{array}{cc} A & B \end{array} \\ \begin{array}{c} A \\ B \end{array} & \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

- $\mathcal{T}(A) = \{2, 4, 6, 8, \dots\}$ and $\mathcal{T}(B) = \{2, 4, 6, 8, \dots\}$.
- Hence, $\gcd(\mathcal{T}(A)) = 2 = \gcd(\mathcal{T}(B))$.

On the other hand



A & B have period 1

Periodicity is a class property

$x \longleftrightarrow y$

Recall this means that

$p^t_{x,y} > 0; \quad p^r_{y,x} > 0$

- In the previous example, the chain is irreducible and both states have the same period.
- This is true in general i.e. if P is irreducible, then $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$ for all $x, y \in S$.

consequence: for P irreducible, makes sense to talk about the period of P.

Periodicity is a class property

- In the previous example, the chain is irreducible and both states have the same period.
- This is true in general i.e. if P is irreducible, then $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$ for all $x, y \in S$.
- To see this, fix $x, y \in S$. By irreducibility, we can find $r, \ell \geq 0$ such that $P^r_{x,y} > 0$ and $P^\ell_{y,x} > 0$.

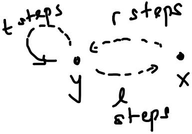
Periodicity is a class property

- In the previous example, the chain is irreducible and both states have the same period.
- This is true in general i.e. if P is irreducible, then $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$ for all $x, y \in S$.
- To see this, fix $x, y \in S$. By irreducibility, we can find $r, \ell \geq 0$ such that $P^r_{x,y} > 0$ and $P^\ell_{y,x} > 0$.
- We will show that $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$.

Periodicity is a class property

- In the previous example, the chain is irreducible and both states have the same period.
- This is true in general i.e. if P is irreducible, then $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$ for all $x, y \in S$.
- To see this, fix $x, y \in S$. By irreducibility, we can find $r, \ell \geq 0$ such that $P^r_{x,y} > 0$ and $P^\ell_{y,x} > 0$.
- We will show that $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$.
- For this, note that if $t \in \mathcal{T}(y)$, then we must have that $t + (r + \ell) \in \mathcal{T}(x)$.
- Therefore,

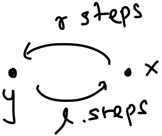
$$\mathcal{T}(y) \subseteq \mathcal{T}(x) - (r + \ell).$$



Periodicity is a class property

- In the previous example, the chain is irreducible and both states have the same period.
- This is true in general i.e. if P is irreducible, then $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$ for all $x, y \in S$.
- To see this, fix $x, y \in S$. By irreducibility, we can find $r, \ell \geq 0$ such that $P^r_{x,y} > 0$ and $P^\ell_{y,x} > 0$.
- We will show that $\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y))$.
- For this, note that if $t \in \mathcal{T}(y)$, then we must have that $t + (r + \ell) \in \mathcal{T}(x)$.
- Therefore,

$$\mathcal{T}(y) \subseteq \mathcal{T}(x) - (r + \ell).$$



- Moreover, we have $(r + \ell) \subseteq \mathcal{T}(x)$.

$\cap \mathcal{T}(y)$

Periodicity is a class property

- $\gcd(A)$ divides all $a \in A$
- it is the largest such integer.

- Therefore, every element of $\mathcal{T}(x) - (r + \ell)$ is divisible by $\gcd(\mathcal{T}(x))$.

$$\mathcal{T}(x) - (r+\ell) = \{ t - (r+\ell) : t \in \mathcal{T}(x) \}$$

Periodicity is a class property

- Therefore, every element of $\mathcal{T}(x) - (r+\ell)$ is divisible by $\gcd(\mathcal{T}(x))$.
- Hence, every element of $\mathcal{T}(y)$ is divisible by $\gcd(\mathcal{T}(x))$, so that, by definition of the gcd, we have

$$\gcd(\mathcal{T}(x)) \leq \gcd(\mathcal{T}(y)).$$

$$\mathcal{T}(y) \subseteq \mathcal{T}(x) - (r+\ell)$$

Periodicity is a class property

- Therefore, every element of $\mathcal{T}(x) - (r + \ell)$ is divisible by $\gcd(\mathcal{T}(x))$.
- Hence, every element of $\mathcal{T}(y)$ is divisible by $\gcd(\mathcal{T}(x))$, so that, by definition of the gcd, we have
$$\gcd(\mathcal{T}(x)) \leq \gcd(\mathcal{T}(y)).$$
- Interchanging the roles of x, y, we see that $\gcd(\mathcal{T}(y)) \leq \gcd(\mathcal{T}(x))$ as well, which shows that x and y have the same period.

Periodicity is a class property

- Therefore, every element of $\mathcal{T}(x) - (r + \ell)$ is divisible by $\gcd(\mathcal{T}(x))$.
- Hence, every element of $\mathcal{T}(y)$ is divisible by $\gcd(\mathcal{T}(x))$, so that, by definition of the gcd, we have
$$\gcd(\mathcal{T}(x)) \leq \gcd(\mathcal{T}(y)).$$
- Interchanging the roles of x, y, we see that $\gcd(\mathcal{T}(y)) \leq \gcd(\mathcal{T}(x))$ as well, which shows that x and y have the same period.
- In fact, the same argument as above shows that if $x \leftrightarrow y$ are two communicating states in S, then
$$\gcd(\mathcal{T}(x)) = \gcd(\mathcal{T}(y)).$$

Aperiodicity

- Let P be the transition matrix of an irreducible DTMC on S.
- We say that P is **aperiodic** if the period of some state (and hence, all states) is 1.

Aperiodicity

- Let P be the transition matrix of an irreducible DTMC on S.
- We say that P is **aperiodic** if the period of some state (and hence, all states) is 1.
- In practice, aperiodicity is not a serious restriction. For instance, if P is an irreducible transition matrix with the unique stationary distribution π, then

$$P' = \frac{1}{2}P + \frac{1}{2}I \rightarrow \left(|S| \times |S| \text{ identity matrix} \right)$$

  is clearly an irreducible, aperiodic, transition matrix with the unique stationary distribution π.
- P' is called the **lazy version** of P.

$$\pi P' = \frac{1}{2}\pi P + \frac{1}{2}\pi I$$
$$= \frac{1}{2}\pi + \frac{1}{2}\pi$$

Convergence theorem

Next week, we will prove the following theorem, which is often called the **Fundamental ~~theorm~~ of Markov Chains**.
theorem

Let P be an irreducible and aperiodic transition matrix on a finite state space S. Then, P has a unique stationary distribution π and moreover, for any $x \in S$,

$$P^t_{x,y} \to \pi(y) \quad \text{as } t \to \infty.$$

$$\| $$

$$\mathbb{P}[X_t = y \mid X_0 = x]$$

$$X_0, X_1, X_2, X_3 \; \cdots \; \underset{\approx}{X_t}, X_{t+1}, \cdots$$

Convergence theorem

Next week, we will prove the following theorem, which is often called the **Fundamental theorm of Markov Chains**.

Let P be an irreducible and aperiodic transition matrix on a finite state space S. Then, P has a unique stationary distribution π and moreover, for any $x \in S$,

$$P^t_{x,y} \to \pi(y) \quad \text{as } t \to \infty.$$

For the rest of this week, we will explore some applications of this theorem.

Markov chain Monte Carlo (MCMC)

- A fundamental computational task in many applications is to sample from a given distribution π on a finite set S.

right now, this has nothing to do w/ Markov chains.

Markov chain Monte Carlo (MCMC)

- A fundamental computational task in many applications is to sample from a given distribution π on a finite set S.
- Given the convergence theorem, the following is a natural approach: construct an irreducible, aperiodic transition matrix on S with stationary distribution π. Simulate a DTMC $(X_n)_{n\geq 0}$ with transition matrix P and starting from $X_0 = x$ (for some $x \in S$). Output X_t for 'sufficiently large' t.

algorithmically (rigorous sense)

* find P

* you have some error tolerance ε.

hard part $\longrightarrow$ * prove a thm upper bounding t in terms of ε.

* run the chain for t steps; return X_t

Markov chain Monte Carlo (MCMC)

- A fundamental computational task in many applications is to sample from a given distribution π on a finite set S.
- Given the convergence theorem, the following is a natural approach: construct an irreducible, aperiodic transition matrix on S with stationary distribution π. Simulate a DTMC $(X_n)_{n\geq 0}$ with transition matrix P and starting from $X_0 = x$ (for some $x \in S$). Output X_t for 'sufficiently large' t.
- By the convergence theorem,

$$\mathbb{P}[X_t = y \mid X_0 = x] = P^t_{x,y} \to \pi(y),$$

so that X_t has distribution approximately equal to π when t is sufficiently large.

Markov chain Monte Carlo (MCMC)

- Given a probability distribution π on S, how can we construct an irreducible and aperiodic transition matrix with stationary distribution π?

main criterion: convergence happens quickly

Markov chain Monte Carlo (MCMC)

- Given a probability distribution π on S, how can we construct an irreducible and aperiodic transition matrix with stationary distribution π?
- In fact, in many applications, we are not given $\pi(x)$, but only $\tilde{\pi}(x) = \pi(x) \cdot Z$ for an unknown (and computationally intractable) constant Z.

$$\pi(x) = \frac{\tilde{\pi}(x)}{Z}$$

since $\sum_x \pi(x) = 1 \Rightarrow Z = \sum_x \tilde{\pi}(x)$

Markov chain Monte Carlo (MCMC)

- Given a probability distribution π on S, how can we construct an irreducible and aperiodic transition matrix with stationary distribution π?
- In fact, in many applications, we are not given $\pi(x)$, but only $\tilde{\pi}(x) = \pi(x) \cdot Z$ for an unknown (and computationally intractable) constant Z.
- As an example, consider Markov random fields (undirected graphical models). Here, we are given an undirected graph $G = (V, E)$ and the state space S is (for instance)

$$S = \{-1, 1\}^V$$

Markov chain Monte Carlo (MCMC)

- Given a probability distribution π on S, how can we construct an irreducible and aperiodic transition matrix with stationary distribution π?
- In fact, in many applications, we are not given $\pi(x)$, but only $\tilde{\pi}(x) = \pi(x) \cdot Z$ for an unknown (and computationally intractable) constant Z.
- As an example, consider Markov random fields (undirected graphical models). Here, we are given an undirected graph $G = (V, E)$ and the state space S is (for instance)

$$S = \{-1, 1\}^V$$

  i.e. there is a variable assigned to each vertex of the graph, which can take on the values ± 1.
- We will denote the number of vertices $|V|$ by n.

The Ising model

- For each element of S (i.e. each configuration of assignments to the variables), there is an associated **Hamiltonian**, which is typically easy to compute.

The Ising model

- For each element of S (i.e. each configuration of assignments to the variables), there is an associated **Hamiltonian**, which is typically easy to compute.
- For instance, for the so-called **ferromagnetic Ising model**, this is given by the function $H : \{\pm 1\}^n \to \mathbb{R}$, where

$$H(x_1, \ldots, x_n) = -\sum_{uv \in E} x_u x_v - h \sum_{v \in V} x_v,$$

where h is a parameter known as the external field.

"lower energy" is "more stable"

The Ising model

- For each element of S (i.e. each configuration of assignments to the variables), there is an associated **Hamiltonian**, which is typically easy to compute.
- For instance, for the so-called **ferromagnetic Ising model**, this is given by the function $H : \{\pm 1\}^n \to \mathbb{R}$, where

$$H(x_1, \ldots, x_n) = -\sum_{uv \in E} x_u x_v - h \sum_{v \in V} x_v,$$

this is a modeling thing.

where h is a parameter known as the external field.

- So, the energy will be lower if neighboring vertices have the same value and if vertices have the same sign as the external field.

The Ising model

- The corresponding **Gibbs distribution/Boltzmann distribution**, whose form is motivated by the principle of maximum entropy, is given by

$$\pi(x) := \exp(-\beta H(x))/Z,$$

where $\beta \geq 0$ is called the **inverse temperature** and Z is a normalizing constant called the **partition function**.

more stable state $\longrightarrow$ $H(x)$ is smaller

$\longrightarrow$ $-H(x)$ is larger

$\longrightarrow$ $\pi(x)$ is larger.

The Ising model

- The corresponding **Gibbs distribution/Boltzmann distribution**, whose form is motivated by the principle of maximum entropy, is given by

$$\pi(x) = \exp(-\beta H(x))/Z,$$

where $\beta \geq 0$ is called the **inverse temperature** and Z is a normalizing constant called the **partition function**.

- Explicitly,

$$Z = \sum_{x \in \{-1,1\}^n} \exp(-\beta H(x)),$$

which is a sum of exponentially many terms.

The Ising model

- The corresponding **Gibbs distribution/Boltzmann distribution**, whose form is motivated by the principle of maximum entropy, is given by

$$\pi(x) = \exp(-\beta H(x))/Z,$$

  where $\beta \geq 0$ is called the **inverse temperature** and Z is a normalizing constant called the **partition function**.
- Explicitly,

$$Z = \sum_{x \in \{-1,1\}^n} \exp(-\beta H(x)),$$

  which is a sum of exponentially many terms.
- In general, Z is computationally intractable (under standard assumptions in computational complexity theory).

The Ising model

- Since Z is computationally intractable, we essentially have access to the function $\tilde{\pi} : \{-1, 1\}^n \to \mathbb{R}^{\geq 0}$ given by

$$\tilde{\pi}(x) = \exp(-\beta H(x)) = \pi(x) \cdot Z.$$

The Ising model

- Since Z is computationally intractable, we essentially have access to the function $\tilde{\pi} : \{-1, 1\}^n \to \mathbb{R}^{\geq 0}$ given by

$$\tilde{\pi}(x) = \exp(-\beta H(x)) = \pi(x) \cdot Z.$$

- The reason for the negative sign is the exponent is to ensure that states with a lower Hamiltonian (energy) have a higher probability.

The Ising model

- Since Z is computationally intractable, we essentially have access to the function $\tilde{\pi} : \{-1, 1\}^n \to \mathbb{R}^{\geq 0}$ given by

$$\tilde{\pi}(x) = \exp(-\beta H(x)) = \pi(x) \cdot Z.$$

$H(x) = -\sum_{u \sim v} x_u x_v$ the smallest I can make this is to make $x_u x_v = 1$

- The reason for the negative sign is the exponent is to ensure that states with a lower Hamiltonian (energy) have a higher probability.
- In particular, for the ferromagnetic Ising model with zero external field $h = 0$, the states with the highest probability are $(1, \ldots, 1)$ and $(-1, \ldots, -1)$.
- As $\beta \to \infty$, π converges to the uniform distribution on $(1, \ldots, 1) \cup (-1, \ldots, -1)$.

$$\pi(x) = \frac{\exp(-\beta H(x))}{\sum \exp(-\beta H(x))}$$

H_{max}

$H_{max} - 0.01$

The Ising model

- Since Z is computationally intractable, we essentially have access to the function $\tilde{\pi} : \{-1, 1\}^n \to \mathbb{R}^{\geq 0}$ given by

$$\tilde{\pi}(x) = \exp(-\beta H(x)) = \pi(x) \cdot Z.$$

- The reason for the negative sign is the exponent is to ensure that states with a lower Hamiltonian (energy) have a higher probability.
- In particular, for the ferromagnetic Ising model with zero external field $h = 0$, the states with the highest probability are $(1, \ldots, 1)$ and $(-1, \ldots, -1)$.
- As $\beta \to \infty$, π converges to the uniform distribution on $(1, \ldots, 1) \cup (-1, \ldots, -1)$. i.e. $\mathbb{P}[(1,\ldots,1)] = \mathbb{P}[(-1,\ldots,-1)] = 1/2$.
- On the other hand, for $\beta = 0$, π is simply the uniform distribution on the entire discrete hypercube $\{-1, 1\}^n$. $\beta H(x) = 0$

The Metropolis chain

- Now, suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.

The Metropolis chain

- Now, suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.
- Next time, we will see the **Metropolis chain**, which provides a very general way to construct a transition matrix P with stationary distribution π.

The Metropolis chain

- Now, suppose that we are given a probability distribution π on S with $\pi(x) > 0$ for all $x \in S$. Possibly, we are not given π, but rather $\tilde{\pi}$, with $\tilde{\pi} = \pi \cdot Z$ for some unknown constant Z.
- Next time, we will see the **Metropolis chain**, which provides a very general way to construct a transition matrix P with stationary distribution π.
- Moreover, the transition matrix only depends on $\tilde{\pi}$ and not π, which as we have seen, is a very important consideration.