STATS 217: Introduction to Stochastic Processes I

Lecture 20

Convergence theorem

- Last time, we proved the convergence theorem for irreducible, aperiodic, finite-state Markov chains.
- Let $(X_n)_{n\geq 0}$ be a DTMC on S with transition matrix P. Suppose that P is irreducible and aperiodic with unique stationary distribution π.
- Let
$$\Delta(n) = \max_{x\in S} \Delta_x(n) = \max_{x\in S} \mathrm{TV}(X_n \mid X_0 = x, \pi).$$
- There exist constants $\epsilon > 0$ and $C > 0$ (depending on P) such that
$$\Delta(n) \leq C \cdot (1-\epsilon)^n.$$

Sub-multiplicativity

- In fact, we worked with the quantities

$$D_{x,y}(n) = \mathrm{TV}(X_n \mid X_0 = x, X_n \mid X_0 = y)$$

  and

$$D(n) = \max_{x,y \in S} D_{x,y}(n).$$

- We showed that

$$\Delta(n) \leq D(n) \leq 2\Delta(n)$$

  for all integers $n \geq 0$ and that for any integers $s, t \geq 0$,

$$D(s+t) \leq D(s)D(t).$$

Mixing time

- For $\varepsilon \in [0,1]$, define the ε-**mixing time** of the chain to be

$$\tau_{\mathsf{mix}}(\varepsilon) := \min\{t : \Delta(t) \leq \varepsilon\}.$$

Mixing time

- For $\varepsilon \in [0, 1]$, define the ε-**mixing time** of the chain to be
$$\tau_{\text{mix}}(\varepsilon) := \min\{t : \Delta(t) \leq \varepsilon\}.$$

- Since $\Delta(n+1) \leq \Delta(n)$ for all $n \geq 0$, it follows that for any $t \geq \tau_{\text{mix}}(\varepsilon)$ and for any $x \in S$,
$$\text{TV}(X_t \mid X_0 = x, \pi) \leq \varepsilon.$$

Mixing time

- For $\varepsilon \in [0,1]$, define the ε-**mixing time** of the chain to be
$$\tau_{\text{mix}}(\varepsilon) := \min\{t : \Delta(t) \leq \varepsilon\}.$$
- Since $\Delta(n+1) \leq \Delta(n)$ for all $n \geq 0$, it follows that for any $t \geq \tau_{\text{mix}}(\varepsilon)$ and for any $x \in S$,
$$\mathrm{TV}(X_t \mid X_0 = x, \pi) \leq \varepsilon.$$
- It is convenient to define
$$\tau_{\text{mix}} := \tau_{\text{mix}}(1/4).$$

Mixing time

- For $\varepsilon \in [0,1]$, define the ε-**mixing time** of the chain to be
$$\tau_{\mathsf{mix}}(\varepsilon) := \min\{t : \Delta(t) \leq \varepsilon\}.$$

- Since $\Delta(n+1) \leq \Delta(n)$ for all $n \geq 0$, it follows that for any $t \geq \tau_{\mathsf{mix}}(\varepsilon)$ and for any $x \in S$,
$$\mathsf{TV}(X_t \mid X_0 = x, \pi) \leq \varepsilon.$$

- It is convenient to define
$$\tau_{\mathsf{mix}} := \tau_{\mathsf{mix}}(1/4).$$

- The choice of the constant $1/4$ is not important and can be replaced by another constant which is strictly smaller than $1/2$.

Mixing time

- The reason that it's often enough to look only at τ_{mix} is because for any $\varepsilon \in (0,1)$,

$$\tau_{\mathsf{mix}}(\varepsilon) \leq \lceil \log_2 \varepsilon^{-1} \rceil \tau_{\mathsf{mix}}.$$

e.g. $\varepsilon = 2^{-100}$

$$\tau_{mix}(2^{-100}) \leq 100\, \tau_{mix}(1/4)$$

Mixing time

- The reason that it's often enough to look only at τ_{mix} is because for any $\varepsilon \in (0,1)$,

$$\tau_{\text{mix}}(\varepsilon) \leq \lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}.$$

$$\Delta(n) \leq D(n) \leq 2\Delta(n)$$

- Indeed,

$$\begin{aligned} \Delta(\lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}) &\leq D(\lceil \log_2 \varepsilon^{-1} \rceil \cdot \tau_{\text{mix}}) \\ &\leq D(\tau_{\text{mix}})^{\lceil \log_2 \varepsilon^{-1} \rceil} \end{aligned}$$

$$D(s+t) \leq D(s)\, D(t)$$

$$D(ks) \leq D(s)^k$$

Mixing time

- The reason that it's often enough to look only at τ_{mix} is because for any $\varepsilon \in (0,1)$,
$$\tau_{\text{mix}}(\varepsilon) \leq \lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}.$$

- Indeed,
$$\begin{aligned}\Delta(\lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}) &\leq D(\lceil \log_2 \varepsilon^{-1} \rceil \cdot \tau_{\text{mix}}) \\ &\leq D(\tau_{\text{mix}})^{\lceil \log_2 \varepsilon^{-1} \rceil} \\ &\leq (2\Delta(\tau_{\text{mix}}))^{\lceil \log_2 \varepsilon^{-1} \rceil}\end{aligned}$$

$$\Delta(\tau_{mix}) \leq 1/4$$

$$\Rightarrow 2\Delta(\tau_{mix}) \leq 1/2$$

Mixing time

- The reason that it's often enough to look only at τ_{mix} is because for any $\varepsilon \in (0,1)$,

$$\tau_{\text{mix}}(\varepsilon) \leq \underbrace{\lceil \log_2 \varepsilon^{-1} \rceil}_{} \tau_{\text{mix}}.$$

this is an upper bound, but in specific examples, this could be smaller.

- Indeed,

$$\begin{aligned}\Delta(\lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}) &\leq D(\lceil \log_2 \varepsilon^{-1} \rceil \cdot \tau_{\text{mix}}) \\ &\leq D(\tau_{\text{mix}})^{\lceil \log_2 \varepsilon^{-1} \rceil} \\ &\leq (2\Delta(\tau_{\text{mix}}))^{\lceil \log_2 \varepsilon^{-1} \rceil} \\ &\leq 2^{-\lceil \log_2 \varepsilon^{-1} \rceil}\end{aligned}$$

"cut-off phenomenon".

Mixing time

- The reason that it's often enough to look only at τ_{mix} is because for any $\varepsilon \in (0, 1)$,
$$\tau_{\text{mix}}(\varepsilon) \le \lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}.$$

- Indeed,

$$\begin{aligned}
\Delta(\lceil \log_2 \varepsilon^{-1} \rceil \tau_{\text{mix}}) &\le D(\lceil \log_2 \varepsilon^{-1} \rceil \cdot \tau_{\text{mix}}) \\
&\le D(\tau_{\text{mix}})^{\lceil \log_2 \varepsilon^{-1} \rceil} \\
&\le (2\Delta(\tau_{\text{mix}}))^{\lceil \log_2 \varepsilon^{-1} \rceil} \\
&\le 2^{-\lceil \log_2 \varepsilon^{-1} \rceil} \\
&\le \varepsilon.
\end{aligned}$$

Coupling of Markov chains

- Consider a transition matrix P on a finite state space S.
- A **coupling of Markov chains** with transition matrix P and initial distributions μ and ν is a process

for concreteness think of
$\left\{ \begin{array}{l} \mu \equiv x \quad \text{det.} \\ \nu \equiv y \end{array} \right\}$

$$(\widehat{X}_t, \widehat{Y}_t)_{t=0}^{\infty}$$

such that for all $t \geq 0$,

$$\widehat{X}_t \sim (X_t \mid X_0 \sim \mu)$$
$$\widehat{Y}_t \sim (X_t \mid X_0 \sim \nu),$$

and such that

$$\left.\widehat{X}_t = \widehat{Y}_t \implies \widehat{X}_{t+1} = \widehat{Y}_{t+1}.\right\}$$

going to be satisfied for most "Reasonable" couplings.

Coupling of Markov chains

- Consider a transition matrix P on a finite state space S.
- A **coupling of Markov chains** with transition matrix P and initial distributions μ and ν is a process

$$(\widehat{X}_t, \widehat{Y}_t)_{t=0}^{\infty}$$

such that for all $t \geq 0$,

$$\widehat{X}_t \sim (X_t \mid X_0 \sim \mu)$$
$$\widehat{Y}_t \sim (X_t \mid X_0 \sim \nu),$$

and such that

$$\widehat{X}_t = \widehat{Y}_t \implies \widehat{X}_{t+1} = \widehat{Y}_{t+1}.$$

- We have already seen couplings of Markov chains in our proof of the convergence theorem

$\rightarrow$ in this case, we used the coupling lemma

Coupling of Markov chains

to guarantee the existence of couplings w/ certain properties.

- As we will soon see, couplings of Markov chains are a useful tool to bound the mixing time in applications.

Coupling of Markov chains

- As we will soon see, couplings of Markov chains are a useful tool to bound the mixing time in applications.
- This is due to the following: Let $(\widehat{X}_t, \widehat{Y}_t)$ be a coupling of two Markov chains with transition matrix P and with $\widehat{X}_0 = x$, $\widehat{Y}_0 = y$. Let

$$\tau_{\text{couple}} := \min\{t : \widehat{X}_t = \widehat{Y}_t\}.$$

$$\hat{X}_t = \hat{Y}_t \Rightarrow \hat{X}_{t+1} = \hat{Y}_{t+1}$$

we know that

$$t \geq \tau_{\text{couple}} \Rightarrow \hat{X}_t = \hat{Y}_t.$$

Coupling of Markov chains

- As we will soon see, couplings of Markov chains are a useful tool to bound the mixing time in applications.
- This is due to the following: Let $(\widehat{X}_t, \widehat{Y}_t)$ be a coupling of two Markov chains with transition matrix P and with $\widehat{X}_0 = x$, $\widehat{Y}_0 = y$. Let

$$\tau_{\text{couple}} := \min\{t : \widehat{X}_t = \widehat{Y}_t\}.$$

idea: we will construct a not-too-hard to analyse coupling for which the dis. of τ_{couple} can be studied.

Recall that

$$D_{x,y}(n) = \text{TV}(X_n \mid X_0 = x, X_n \mid X_0 = y).$$

Then,

$$D_{x,y}(n) \leq \mathbb{P}[\tau_{\text{couple}} \geq n].$$

& then use $\Delta(n) \leq D(n) = \max_{x,y} D_{x,y}(n)$.

Coupling of Markov chains

- The proof is a direct application of the coupling lemma.
- Indeed, since $\widehat{X}_n \sim X_n \mid X_0 = x$ and $\widehat{Y}_n \sim X_n \mid X_0 = y$, we have

$$D_{x,y}(n) \leq \mathbb{P}[\widehat{X}_n \neq \widehat{Y}_n] \leq \mathbb{P}[\tau_{\text{couple}} \geq n].$$

$$\text{TV}(X_n \mid X_0 = x, X_n \mid X_0 = y)$$

$$= \text{TV}(\hat{X}_n, \hat{Y}_n) \leq \mathbb{P}[\hat{X}_n \neq \hat{Y}_n]$$

Coupling of Markov chains

- The proof is a direct application of the coupling lemma.
- Indeed, since $\widehat{X}_n \sim X_n \mid X_0 = x$ and $\widehat{Y}_n \sim X_n \mid X_0 = y$, we have

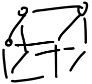
$$D_{x,y}(n) \le \mathbb{P}[\widehat{X}_n \neq \widehat{Y}_n] \le \mathbb{P}[\tau_{\text{couple}} \ge n].$$

- Therefore, by Markov's inequality,

$$D_{x,y}(4 \cdot \mathbb{E}[\tau_{\text{couple}}]) \le \mathbb{P}[\tau_{\text{couple}} \ge 4 \cdot \mathbb{E}[\tau_{\text{couple}}]] \le \frac{1}{4}.$$

plug in $n = 4 \cdot \mathbb{E}[\tau_{\text{couple}}]$

Example: Lazy random walk on the hypercube



$n=3$

000

100 010 001

- $S = \{0,1\}^n$.
- The transitions are given as follows. Suppose the current state is x. With probability $1/2$, the chain remains at x; with probability $1/2$, it moves uniformly to one of the n possible vectors y which differ from x in exactly one coordinate.
- The transition matrix is clearly aperiodic and irreducible, and we have seen that the unique stationary distribution is the uniform distribution on $\{0,1\}^n$.

$\to$ we want to study

$$\tau_{mix}(\varepsilon) \sim n \log n$$

$\to$ we will use the coupling technique.

Example: Lazy random walk on the hypercube

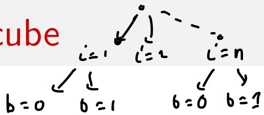
- $S = \{0, 1\}^n$.
- The transitions are given as follows. Suppose the current state is x. With probability $1/2$, the chain remains at x; with probability $1/2$, it moves uniformly to one of the n possible vectors y which differ from x in exactly one coordinate.
- The transition matrix is clearly aperiodic and irreducible, and we have seen that the unique stationary distribution is the uniform distribution on $\{0, 1\}^n$.
- Here is an equivalent description of the transitions: suppose the current state is x. We choose a coordinate $i \in \{1, \ldots, n\}$ uniformly at random and an unbiased bit $b \in \{0, 1\}$, also uniformly at random, and independent of the coordinate i. (i, b)
- Then, we set the value of coordinate i to b and keep all other coordinates unchanged.

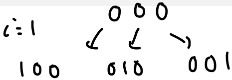
$(X_{t+1})_i = b$

$(X_{t+1})_j = (X_t)_j \quad \forall j \neq i$

the chain stays at X_t $\iff b = (X_t)_i$

Example: Lazy random walk on the hypercube





- Given this alternate description, there is a natural choice of coupling: for the two chains started from x and y, use the same i and b at every step.

$\hat{X}_0 = x \qquad \hat{Y}_0 = y$

first step: generate (i, b)

$$(\hat{X}_1)_i = b \quad = \quad (\hat{Y}_1)_i = b$$

$$(\hat{X}_1)_j = x_j \quad j \neq i \qquad (\hat{Y}_1)_j = y_j \quad j \neq i$$

$$\hat{X}_t = \hat{Y}_t \Rightarrow \hat{X}_{t+1} = \hat{Y}_{t+1}$$

we want to understand τ_{couple}.

Example: Lazy random walk on the hypercube

- Given this alternate description, there is a natural choice of coupling: for the two chains started from x and y, use the same i and b at every step.
- Let τ denote the first time that each coordinate i has been chosen to be updated. Then, clearly, $\widehat{X}_\tau = \widehat{Y}_\tau$.

when we update coord i

we set $(\hat{X}_{t+1})_i = b = (\hat{Y}_{t+1})$

Example: Lazy random walk on the hypercube

- Given this alternate description, there is a natural choice of coupling: for the two chains started from x and y, use the same i and b at every step.
- Let τ denote the first time that each coordinate i has been chosen to be updated. Then, clearly, $\widehat{X}_\tau = \widehat{Y}_\tau$.
- Moreover, τ is exactly the first time to collect all n coupons in the coupon-collector problem and

we have n coupons

$$\mathbb{P}[\tau > t] \leq \boxed{n} \underbrace{\left(1 - \frac{1}{n}\right)^t}_{\text{prob that coupon } i \text{ has not been collected by time } t} \leq ne^{-t/n},$$

union bound

Example: Lazy random walk on the hypercube

- Given this alternate description, there is a natural choice of coupling: for the two chains started from x and y, use the same i and b at every step.
- Let τ denote the first time that each coordinate i has been chosen to be updated. Then, clearly, $\widehat{X}_\tau = \widehat{Y}_\tau$.
- Moreover, τ is exactly the first time to collect all n coupons in the coupon-collector problem and

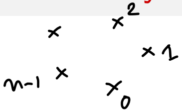
$$\mathbb{P}[\tau > t] \leq n\left(1 - \frac{1}{n}\right)^t \leq ne^{-t/n},$$

which gives $\tau_{\text{mix}} \leq n \log n + n \log(1/4)$.

$\tau_{mix}(\varepsilon) \leq n \log n + n \log(1/\varepsilon)$

$\tau_{mix} \sim \frac{1}{2} n \log n$

Example: Lazy random walk on the cycle



- The states of the n-cycle can be identified with $\mathbb{Z}_n$, the integers modulo n.

Example: Lazy random walk on the cycle

- The states of the n-cycle can be identified with $\mathbb{Z}_n$, the integers modulo n.
- The transitions are given as follows. Suppose that the current state is x. With probability $1/2$, the chain remains at x; with probability $p/2$, it moves to $x+1$; with probability $q/2$, it moves to $x-1$. Here, $p+q=1$.

arithmetic is
mod n

$x+1$
$p/2$
x
$x-1$
$q/2$

0 1

→ mixing time?

→ again, we will find a coupling.

Example: Lazy random walk on the cycle

- The states of the n-cycle can be identified with $\mathbb{Z}_n$, the integers modulo n.
- The transitions are given as follows. Suppose that the current state is x. With probability $1/2$, the chain remains at x; with probability $p/2$, it moves to $x+1$; with probability $q/2$, it moves to $x-1$. Here, $p+q=1$.
- Here is a natural choice of coupling: start the two chains at x and y. At each step, flip an unbiased coin. If the coin lands heads, then the x-chain stays put, and the y chain moves $+1$ with probability p and -1 with probability q. If the coin lands tails, then the y-chain stays put, and the x chain moves $+1$ with probability p and -1 with probability q.

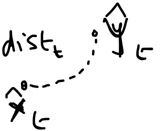
- is this a coupling?

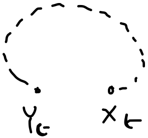
consider the x-chain → staying where you are : $1/2$
→ $+1$: $p/2$
→ -1 : $q/2$

Example: Lazy random walk on the cycle

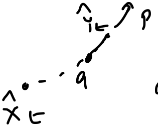
what is the distribution of τ_{couple}?

- Let dist_t denote the (clockwise) distance between the states of the two chains at time t.





dist_{t+1}



if H

$\text{dist}_{t+1} = \text{dist}_t + 1$ w.p. $p \cdot p$

$= \text{dist}_t - 1$ w.p. $p \cdot q$

if T

$\text{dist}_{t+1} = \text{dist}_t - 1$ w.p. p

at each time

$$\text{dist}_{t+1} = \begin{cases} \text{dist}_t + 1 & : \frac{p}{2} + \frac{q}{2} = \frac{1}{2} \\ \text{dist}_t - 1 & : \frac{p}{2} + \frac{q}{2} = \frac{1}{2} \end{cases}$$

Example: Lazy random walk on the cycle

w.p. q
$\text{dist}_t + 1$.

- Let dist_t denote the (clockwise) distance between the states of the two chains at time t.
- Then, $(\text{dist}_t)_{t \geq 0}$ is a simple symmetric random walk on $\{0, \ldots, n\}$ with absorbing states 0 and n.

$0 \quad 1 \quad \cdots \quad k \quad \cdots \quad n$

Example: Lazy random walk on the cycle

- Let dist_t denote the (clockwise) distance between the states of the two chains at time t.
- Then, $(\text{dist}_t)_{t\geq 0}$ is a simple symmetric random walk on $\{0, \ldots, n\}$ with absorbing states 0 and n.
- From Gambler's ruin, we know that if the initial distance is k, then the expected time to absorption is $k(n-k) \leq n^2/4$.
- Hence,

$$\tau_{\text{mix}} \leq 4 \cdot \frac{n^2}{4} = n^2.$$

is this a good bound?

yes, up to a constant