MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.265/15.070J Lecture 11 — March 20, SP17

Lecturer: Yury Polyansky — Scribe notes by Justin Holmgren

Markov Chains II: Coupling and Mixing

Content.

1. Coupling Markov Processes
2. Dobrushin Extension
3. Mixing Times
4. The Cycle and the Hypercube

1 Coupling Markov Processes

Today we will write X_t and Y_t to denote two Markov chains with the same transition kernel P, but possibly different initial distributions $P_{X_0} \neq P_{Y_0}$.

We review several equivalent definitions of total variational distance. We have

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &= \frac{1}{2}\sum_x |P(x) - Q(X)| \\ &= \sup_E P(E) - Q(E) \\ &= \inf_{\substack{X\sim P \\ Y\sim Q}} \mathbb{P}[X \neq Y] \end{aligned}$$

The supremum characterization makes it easy to prove lower bounds on variational distance, while the infimum characterization allows for upper bounds. We note two basic properties of variational distance.

(a) $d_{\mathrm{TV}}(P_{A,B}, Q_{A,B}) \geq d_{\mathrm{TV}}(P_A, Q_A)$

(b) If $\{X_t\}$ and $\{Y_t\}$ are Markov chains with the same transition kernel P, then

$$d_{\mathrm{TV}}(P_{X_t^\infty}, P_{Y_t^\infty}) = d_{\mathrm{TV}}(P_{X_t}, P_{Y_t}).$$

Proof. (a) Follows from the sup definition of d_{TV}.

(b) We know from (a) that $d_{\mathrm{TV}}(P_{X_t^\infty}, P_{Y_t^\infty}) \geq d_{\mathrm{TV}}(P_{X_t}, P_{Y_t})$, so we just need to prove $d_{\mathrm{TV}}(P_{X_t^\infty}, P_{Y_t^\infty}) \leq d_{\mathrm{TV}}(P_{X_t}, P_{Y_t})$. We will use the following fact.

Fact Every Markov chain $\{X_t\}$ can be represented as $X_{t+1} = f(X_t, U_{t+1})$ for some function $f : \mathcal{X} \times [0,1] \to \mathcal{X}$ and $\{U_t\} \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$.

Now take a coupling such that $\mathbb{P}[X_t \neq Y_t] = d_{\mathrm{TV}}(P_{X_t}, P_{Y_t})$. We extend this to a coupling of X_t^∞ and Y_t^∞ by defining $X_{j+1} = f(X_j, U_{j+1})$ and $Y_{j+1} = f(Y_j, U_{j+1})$ for each $j \in \{t, \ldots, \infty\}$. Notably we use the same randomness for both X_{j+1} and Y_{j+1}. This is a coupling because it is clear that $X_t^\infty \sim P_{X_t^\infty}$ and $Y_t^\infty \sim P_{Y_t^\infty}$.

But we have

$$1 - d_{\mathrm{TV}}(P_{X_t^\infty}, P_{Y_t \infty}) \geq \mathbb{P}[X_s = Y_s \ \forall s \geq t] = \mathbb{P}[X_t = Y_t],$$

so

□

2 Dobrushin Extension

Proposition 1. *For any probability laws P_{AB}, $P_{A'B'}$, and $P_{AA'}$, suppose that $d_{\mathrm{TV}}(P_{B|A=a}, P_{B'|A'=a'}) \leq r(a, a')$ for some function r. Then there is a coupling of (A, B) to (A', B') such that $\mathbb{P}[B \neq B'] \leq \mathbb{E}[r(A, A')]$.*

Proof. For every Q and Q', let $P_{B,B'|A=a,A'=a'}$ be a coupling such that

$$\mathbb{P}[B \neq B'|A = a, A' = a'] \leq r(a, a').$$

Now we construct the desired coupling by defining

$$P_{A,B,A'B'}(a, b, a', b') = P_{AA'}(a, a') \cdot P_{B,B'|A=a,A'=a'}(b, b')$$

□

Theorem 1. *If P is an aperiodic and irreducible Markov transition kernel (on a finite state space), then*

$$d_{\mathrm{TV}}(P_t(x, \cdot), \pi) \leq C\alpha^t$$

for some $\alpha < 1$ and $C > 0$.

Proof. First consider the special case in which $P(x, y) > \epsilon$ for all x, y. Then $d_{\mathrm{TV}}(P(x, \cdot), P(y, \cdot)) \leq (1 - \epsilon) \cdot \mathbf{1}_{x \neq y}$ for all x, y. Now apply Dobrushin extension. Let $X_0 = a_0$ and $Y_0 \sim \pi$. Consider an independent coupling of X_0 to Y_0. By Dobrushin, there is some coupling P_{X_0, X_1, Y_0, Y_1} under which

$$\mathbb{P}[X_1 \neq Y_1] \leq (1 - \epsilon)\mathbb{P}[X_0 \neq Y_0].$$

Repeating this argument,

In the general case, there is some $t_0 > 0$ for which $P_{t_0}(x, y) > 0$ for all x, y. Since the state space is finite, it is in fact the case that $P_{t_0}(x, y) > \epsilon$ for some $\epsilon > 0$ and all x, y. Thus applying the previous bound, we get

$$d_{\mathrm{TV}}(P_{X_{t_0 \cdot m}}, P_{Y_{t_0 \cdot m}}) \leq (1 - \epsilon)^m,$$

and so $d_{\mathrm{TV}}(P_{X_t}, P_{Y_t}) \leq (1 - \epsilon)^{\lfloor \frac{t}{t_0} \rfloor}$. □

3 Mixing Times

We now study the mixing time of Markov chains. To do so, we introduce several definitions of a distance from the stationary distribution π at time t for a Markov transition kernel P.

- $d(t) \stackrel{\mathsf{def}}{=} \sup_{x \in \mathcal{X}} d_{\mathrm{TV}}(P_t(x, \cdot), \pi)$.
- $\bar{d}(t) \stackrel{\mathsf{def}}{=} \sup_{x, y \in \mathcal{X}} d_{\mathrm{TV}}(P_t(x, \cdot), P_t(y, \cdot))$.
- $t_{mix}(\epsilon) \stackrel{\mathsf{def}}{=} \inf \{t : d(t) \leq \epsilon\}$.
- $t_{mix} \stackrel{\mathsf{def}}{=} t_{mix}(\frac{1}{4})$.
- $d^{(p)}(t) = \sup_{x \in \mathcal{X}} \|q_t(x, \cdot) - 1\|_p$, where

  $$q_t(x, y) \stackrel{\mathsf{def}}{=} \frac{p_t(x, y)}{\pi(y)}$$

  and

  $$\|f\|_p \stackrel{\mathsf{def}}{=} \left(\sum_x \pi(x) |f(x)|^p \right)^{1/p}.$$

  For example, $d^{(2)}(t)$ is just $\sup_{x \in \mathcal{X}} \chi^2(p_t(x, \cdot) \| \pi)$, where χ^2 is the chi-squared divergence.

These metrics are all essentially equivalent and sub-multiplicative, as described in the following proposition.

Proposition 2. *We have*

1. $d(t) \le \bar{d}(t) \le 2d(t)$.
2. $\bar{d}(t+s) \le \bar{d}(t)\bar{d}(s)$.
3. $d(t) = \frac{1}{2}d^{(1)}(t)$.
4. $d^{(p)}(t) \le d^{(q)}(t)$ *for* $p \le q$ *and furthermore,* $d^{(p)}(t+s) \le d^{(1)}(t)d^{(p)}(s) \le d^{(p)}(t)d^{(p)}(s)$.
5. *For a reversible Markov chain,*

$$\begin{aligned} d^{(2)}(t) &= \sup_x (q_{2t}(x,x) - 1)^{1/2} \\ &= (d^{(\infty)}(2t))^{1/2}. \end{aligned}$$

Proof. 1. This is straight-forward; it follows from the convexity of d_{TV} and the triangle inequality.

2. This follows from Dobrushin.

3. This follows from the definition of d_{TV}.

4. The claim that $d^{(p)}(t) \le d^{(q)}(t)$ is just Minkowski's inequality.

   To prove the second part, let (f,g) denote the inner product on the state space of the Markov chain with respect to the stationary distribution π. That is, define

$$(f,g) \stackrel{\mathsf{def}}{=} \sum_x \pi(x) f(x) \bar{g}(x).$$

   Then we have

$$\begin{aligned} \|q_t(x,\cdot) - 1\|_p &= \sup_{\substack{f:\|f\|_{p'} \le 1 \\ p' = \frac{p}{p-1}}} \|P_t f - \mathbb{E}[f]\|_\infty \qquad \text{by Holder} \\ &= \sup_{\substack{f:\|f\|_{p'} = 1 \\ \mathbb{E}[f] = 0}} \|P_t f\|_\infty, \end{aligned}$$

   which implies $\|P_{t+s} f\|_\infty = \|P_t P_s f\|_\infty \le d^{(1)}(t) d^{(p)}(s)$.

5. By definition, $P_t f(x) = (q_t(x, \cdot), f)$, so

$$\begin{aligned} q_{2t}(x,y) &= P_t P_t(x,y)/\pi(y) \\ &= (q_t(x,\cdot), q_t(\cdot, y)) \\ &= (q_t(x,\cdot), q_t(y,\cdot)) \qquad \text{by reversibility} \\ &\le \sup_x \|q_t(x,\cdot)\|_2^2 \end{aligned}$$

with equality iff $x = y$.

Now by 4, it suffices to bound $d^{(2)}(2t)$, and indeed we have

$$\begin{aligned} d^{(2)}(2t) &= \sup_x \|q_t(x,\cdot)\|_2^2 - 1 \qquad \text{because } (q_t(x,\cdot), 1) = 1 \\ &= \sup_x (q_t(x,\cdot), q_t(\cdot, x)) \\ &= \sup_x q_{2t}(x,x) - 1. \end{aligned}$$

To get the second equality, we observe that

$$d^{(\infty)}(2t) = \sup_{x,y} q_{2t}(x,y) - 1 = \sup_x q_{2t}(x,x) - 1.$$

□

4 The Cycle and the Hypercube

Proposition 3. *We have the following general coupling facts.*

I. If $\{X_t, Y_t\}_{t=1}^{\infty}$ is an arbitrary coupling and $\tau_{\text{couple}} \stackrel{\text{def}}{=} \inf\{t \ge 0 : X_t^{\infty} = Y_t^{\infty}\}$ (i.e. the first time at which X_t and Y_t agree forever), then $d_{\text{TV}}(P_{X_t}, P_{Y_t}) \le \mathbb{P}[\tau_{\text{couple}} > t]$.

II. If X_t, Y_t are Markov chains with the same transition kernel, and if X_t is Markov with respect to the filtration $\mathcal{F}_t = \sigma(X_0^t, Y_0^t)$ (i.e. if $X_{t+1} \perp\!\!\!\perp \mathcal{F}_t | X_t$) then

$$d_{\text{TV}}(P_{X_t}, P_{Y_t}) \le \mathbb{P}[\tau_{\text{meet}} > t],$$

where $\tau_{\text{meet}} \stackrel{\text{def}}{=} \inf\{t \ge 0 : X_t = Y_t\}$.

Now we study mixing times for two examples of Markov chains: the lazy random walk on the n-cycle C_n and the boolean hypercube H_n. We will see that even though $|C_n| \ll |H_n|$ and C_n and H_n have roughly the same diameter, C_n has a much larger mixing time than H_n ($\Theta(n^2)$ compared to $\Theta(n \log n)$). We will develop a better understanding of this phenomenon in future lectures.

4.1 The mixing time for C_n

To upper bound the mixing time, we use a coupling of two lazy random walks $\{X_t\}$ and $\{Y_t\}$ on C_n. Because the random walks are lazy, we can define a coupling in which X_t moves with probability $1/2$ and otherwise Y_t moves. If $X_t = Y_t$, then they move together. Now define $Z_t = d(X_t, Y_t)$. Then

$$\mathbb{P}^{n/2}[\tau_0 > t] = \mathbb{P}[\max_{0 \le s \le t} S_{\mathbb{Z}} < n/2] = \mathbb{P}[|S_t| \le n/2] \le \frac{cn}{\sqrt{t}}$$

which implies that $t_{mix} = O(n^2)$.

For a lower bound, fix some point p on the cycle, and define $A_n = \{d(\cdot, p) < n/4\}$. Now clearly under the stationary distribution, $\mathbb{P}[A] = \frac{1}{2}$. But if $X_0 = p$, then by reduction to the symmetric walk, $\mathbb{P}[d(X_t, p) < n/4] > 1 - Ct/n^2$, which implies that the mixing time is $\Omega(n^2)$.

4.2 The mixing time for H_n

There is also a natural coupling for two hypercube lazy random walks $\{X_t\}$ and $\{Y_t\}$. At every time step, we pick a uniformly random index $i \in \{1, \ldots, n\}$, and a uniform bit $b \in \{0, 1\}$. Define X_{t+1} and Y_{t+1} by changing both of their i^{th} bits to b. Now we can observe that $Z_t \stackrel{\text{def}}{=} d_H(X_t, Y_t)$ is also a Markov chain (corresponding to the coupon collector problem).

So

$$\mathbb{P}^n[Z_t \neq 0] \le n(1 - \frac{1}{n})^t \le ne^{-t/n} \le e^{-C}$$

if $t > n \ln n + nC$.

To lower bound the mixing time, define $f(X_t) \stackrel{\text{def}}{=} \|X_t\|_H$. Now $\pi(\{f > n/2\}) = 1/2$. But if $X_0 = 0^n$, we can use concentration of measure to show that

$$\mathbb{P}\left[f(X_t) > n/2\right]$$

is much smaller than $1/2$ if $t = o(n \ln n)$.