6.896 Probability and Computation February 9, 2011

Lecture 3

Lecturer: Constantinos Daskalakis *Scribe: Alessandro Chiesa & Zeyuan Zhu*

NOTE: The content of these notes has not been formally reviewed by the lecturer. It is recommended that they are read critically.

1 Recap

Last time we defined Markov chains as time-independent memoryless stochastic processes over a finite set Ω; we saw how a Markov chain $\mathcal{X}$ can be naturally represented as a weighted directed graph $G(\mathcal{X})$; we introduced fundamental properties of a Markov chain such as irreducibility and aperiodicity, and then used these properties to give some basic characterizations of Markov chains.

Finally, last time we stated (and proved the "reversible" weaker version of) the Fundamental Theorem of Markov Chains:

Theorem 1. *If a Markov chain $\mathcal{X}$ is irreducible and aperiodic, then:*

1. *it has a unique stationary distribution π, and*
2. *for every element x in Ω, $\lim_{t\to+\infty} p_x^{(t)} = \pi$.*

We give three example applications of the theorem:

Example 1 (Card Shuffling)**.** *Last time, for the methods of top-in-at-random and riffle shuffle, we observed that the corresponding transition matrices are doubly stochastic, and thus both Markov chains are reversible with respect to the uniform distribution. Combined with the observations that both Markov chains are irreducible and aperiodic, we deduced that the two stochastic processes converge to the uniform distribution.*

Example 2 (Random Walk on Undirected Graph)**.** *Consider any undirected graph $G = (V, E)$, and define a Markov chain on it where the transition probability is as follows:*

$$P(x,y) \overset{\text{def}}{=} \begin{cases} \dfrac{1}{\deg(x)} & \textit{if } (x,y) \in E \\ 0 & \textit{otherwise} \end{cases} .$$

Let $\mathcal{X}$ the Markov chain whose transition matrix is P. Observe that:

- *$\mathcal{X}$ is irreducible if and only if G is connected;*
- *$\mathcal{X}$ is aperiodic if and only if G non-bipartite;*
- *$\mathcal{X}$ is reversible with respect to π where $\pi(x) \overset{\text{def}}{=} \frac{\deg(x)}{2|E|}$ for every $x \in V$.*

In particular, if $\mathcal{X}$ is both irreducible and aperiodic, then the random walk converges to π from any starting point.

Example 3 (Designing Markov Chains)**.** *A Markov chain is usually not given to us by an angel with the command to analyze it. Instead, we have in mind a finite set Ω and a positive weight function $w\colon \Omega \to \mathbb{R}^+$, and the goal is to* design *a Markov chain $\mathcal{X}$ such that, for every element x in Ω, $\lim_{t\to+\infty} p_x^{(t)} = \pi$, where $\pi \overset{\text{def}}{=} w/Z_w$. How to do that?*

We briefly discuss a very generic solution to the problem, known as the Metropolis *approach. Consider an arbitrary graph $\Gamma = (\Omega, E)$ and an arbitrary function $\kappa\colon \Omega \times \Omega \to \mathbb{R}^+$ such that:*

1. *for all distinct elements x and y in Ω, if $(x,y) \in E$ then $\kappa(x,y) = \kappa(y,x) > 0$, and*
2. *for all elements x and y in Ω, if $(x,y) \notin E$ then $\kappa(x,y) = 0$.*

Now consider the Markov chain $\mathcal{X} = \mathcal{X}(\Gamma, \kappa)$ that is defined as follows: whenever the chain is in a state $x \in \Omega$, pick a neighbor $y \in \Omega$ with probability $\kappa(x,y)$ and then only accept to move to y with probability $\min\left\{1, \frac{w(y)}{w(x)}\right\}$. *(And hence the function κ is known as a* proposal function.*) In other words, the transition matrix $P_{\mathcal{X}}$ is such that $P_{\mathcal{X}}(x,y) = \kappa(x,y) \cdot \min\left\{1, \frac{w(y)}{w(x)}\right\}$. It is easy to check that $\mathcal{X}$ is reversible with respect to our goal distribution π. Hence, whenever $\mathcal{X}$ is irreducible and aperiodic, the Metropolis random walk converges to π from any starting point.*

Exericse (1pt): Prove that the Metropolis Markov chain $\mathcal{X}(\Gamma, \kappa)$ is reversible with respect to the distribution $\pi \stackrel{\text{def}}{=} w/Z_w$. (Hint: $w(y)/w(x) = \pi(y)/\pi(x)$.)

Today we prove Theorem 1 in its full generality.

2 Total Variation Distance and Couplings

First, we need some definitions and basic facts about statistical distance and couplings between probability measures.

Definition 1. *Let μ and η be two probability measures over a finite set Ω. The* total variation distance *between μ and η (also called statistical distance) is the normalized ℓ_1-distance between the two probability measures:*

$$\|\mu - \eta\|_{\text{tv}} \stackrel{\text{def}}{=} \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \eta(x)| \ .$$

Moreover, if X and Y are two random variables drawn from μ and η respectively, we will write $\|X - Y\|_{\text{tv}}$ to denote $\|\mu - \eta\|_{\text{tv}}$.

The total variation distance denotes the "area in between" the two curves $C_\mu \stackrel{\text{def}}{=} \{(x, \mu(x))\}_{x \in \Omega}$ and $C_\eta \stackrel{\text{def}}{=} \{(x, \eta(x))\}_{x \in \Omega}$. Next, we prove a simple relation that shows that the total variation distance is exactly the largest different in probability, taken over all possible events:

Lemma 1. *Let μ and η be two probability measures over a finite set Ω. Then,*

$$\|\mu - \eta\|_{\text{tv}} = \max_{E \subseteq \Omega} |\mu(E) - \eta(E)| \ .$$

Proof: Define A to be the subset of those elements x in Ω for which $\mu(x) \geq \eta(x)$. Then,

$$\begin{aligned}
\|\mu - \eta\|_{\text{tv}} &= \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \eta(x)| \\
&= \frac{1}{2} \left(\sum_{x \in A} \big(\mu(x) - \eta(x)\big) + \sum_{x \in \overline{A}} \big(\eta(x) - \mu(x)\big) \right) \\
&= \frac{1}{2} \left(\mu(A) - \eta(A) + \eta(\overline{A}) - \mu(\overline{A}) \right) \\
&= \mu(A) - \eta(A) \\
&= \max_{E \subseteq \Omega} |\mu(E) - \eta(E)| \ ,
\end{aligned}$$

as desired. □

Now we introduce the notion of a coupling:

Definition 2. *Let μ and η be two probability measures over a finite set Ω. A probability measure ω over $\Omega \times \Omega$ is a* coupling *of (μ, η) if its two marginals are μ and η respectively, that is,*

- *for every $x \in \Omega$, $\sum_{y\in\Omega} \omega(x, y) = \mu(x)$, and*
- *for every $y \in \Omega$, $\sum_{x\in\Omega} \omega(x, y) = \eta(y)$.*

Next, we prove a basic lemma that says that one cannot jointly sample from two probability measures over the same finite set in a way that will make the two samples be different less often than the total variation distance between the probability measures (and, moreover, there is a sampling technique that matches this bound):

Lemma 2 (Coupling Lemma)**.** *Let μ and η be two probability measures over a finite set Ω. Then:*

1. *For any coupling ω of (μ, η), if the random variable (X, Y) is distributed according to ω, then*

$$\Pr\big[X \neq Y\big] \geq \|\mu - \eta\|_{\mathrm{tv}} \ .$$

2. *There exists an* optimal *coupling ω^* of (μ, η) for which* $\Pr\big[X \neq Y\big] = \|\mu - \eta\|_{\mathrm{tv}}$.

Proof: Fix any coupling ω of (μ, η), and let (X, Y) be distributed according to ω. Then, for any $z \in \Omega$,

$$\begin{aligned}
\mu(z) &= \Pr\big[X = z\big] \\
&= \Pr\big[X = z \wedge Y = X\big] + \Pr\big[X = z \wedge Y \neq X\big] \\
&\leq \Pr[Y = z] + \Pr\big[X = z \wedge Y \neq X\big] \\
&= \eta(z) + \Pr\big[X = z \wedge Y \neq X\big] \ ,
\end{aligned}$$

so that we deduce $\mu(z) - \eta(z) \leq \Pr\big[X = z \wedge Y \neq X\big]$. Moreover, by symmetry, for any $z \in \Omega$, $\eta(z) - \mu(z) \leq \Pr\big[Y = z \wedge X \neq Y\big]$. Therefore,

$$\begin{aligned}
2 \cdot \|\mu - \eta\|_{\mathrm{tv}} &= \sum_{z\in\Omega} |\mu(z) - \eta(z)| \\
&= \sum_{\substack{z\in\Omega \\ \mu(z)\geq\eta(z)}} \big(\mu(z) - \eta(z)\big) + \sum_{\substack{z\in\Omega \\ \mu(z)<\eta(z)}} \big(\eta(z) - \mu(z)\big) \\
&\leq \sum_{\substack{z\in\Omega \\ \mu(z)\geq\eta(z)}} \Pr\big[X = z \wedge Y \neq X\big] + \sum_{\substack{z\in\Omega \\ \mu(z)<\eta(z)}} \Pr\big[Y = z \wedge X \neq Y\big] \\
&\leq \Pr\big[X \neq Y\big] + \Pr\big[X \neq Y\big] \ .
\end{aligned}$$

Alternatively, we could have used Lemma 1 to prove the bound.[1] Next, we construct an optimal coupling ω^* of (μ, η). For each $(x, y) \in \Omega \times \Omega$, define

$$\omega^*(x, y) \stackrel{\text{def}}{=} \begin{cases} \min\{\mu(x), \eta(y)\} & \text{if } x = y \\ \dfrac{\max\big\{\mu(x) - \eta(x), 0\big\} \cdot \max\big\{\eta(y) - \mu(y), 0\big\}}{\|\mu - \eta\|_{\mathrm{tv}}} & \text{otherwise} \end{cases} \ .$$

[1] First observe that $\Pr\big[X = Y\big] = \sum_{z\in\Omega} \omega(z, z) \leq \sum_{z\in\Omega} \min\{\mu(z), \eta(z)\}$, and then deduce that

$$\begin{aligned}
\Pr\big[X \neq Y\big] \geq 1 - \sum_{z\in\Omega} \min\{\mu(z), \eta(z)\} &= \sum_{z\in\Omega} \big(\mu(z) - \min\{\mu(z), \eta(z)\}\big) \\
&= \sum_{\substack{z\in\Omega \\ \mu(z)\geq\eta(z)}} \big(\mu(z) - \eta(z)\big) = \max_{E\subseteq\Omega} |\mu(E) - \eta(E)| = \|\mu - \eta\|_{\mathrm{tv}} \ .
\end{aligned}$$

Define A to be the subset of those elements x in Ω for which $\mu(x) \geq \eta(x)$. We first argue that ω^* is a valid coupling. Indeed, for each $x \in A$,

$$\begin{aligned}
\sum_{y\in\Omega} \omega^*(x,y) &= \min\{\mu(x),\eta(x)\} + \sum_{\substack{y\in\Omega \\ y\neq x}} \omega^*(x,y) \\
&= \min\{\mu(x),\eta(x)\} + \max\big\{\mu(x)-\eta(x),0\big\} \cdot \frac{\sum_{\substack{y\in\Omega \\ y\neq x}} \max\big\{\eta(y)-\mu(y),0\big\}}{\|\mu-\eta\|_{\mathrm{tv}}} \\
&= \eta(x) + \big(\mu(x)-\eta(x)\big) \cdot \frac{\sum_{y\in\overline{A}} \big(\eta(y)-\mu(y)\big)}{\|\mu-\eta\|_{\mathrm{tv}}} \\
&= \eta(x) + \big(\mu(x)-\eta(x)\big) \cdot 1 \\
&= \mu(x) \ ,
\end{aligned}$$

and, for each $x \in \overline{A}$,

$$\begin{aligned}
\sum_{y\in\Omega} \omega^*(x,y) &= \min\{\mu(x),\eta(x)\} + \sum_{\substack{y\in\Omega \\ y\neq x}} \omega^*(x,y) \\
&= \min\{\mu(x),\eta(x)\} + \max\big\{\mu(x)-\eta(x),0\big\} \cdot \frac{\sum_{\substack{y\in\Omega \\ y\neq x}} \max\big\{\eta(y)-\mu(y),0\big\}}{\|\mu-\eta\|_{\mathrm{tv}}} \\
&= \mu(x) + 0 \cdot \frac{\sum_{\substack{y\in\Omega \\ y\neq x}} \max\big\{\eta(y)-\mu(y),0\big\}}{\|\mu-\eta\|_{\mathrm{tv}}} \\
&= \mu(x) \ ,
\end{aligned}$$

so that the marginal for X is indeed μ. An analogous argument shows that the marginal for Y is η. Finally,

$$\begin{aligned}
\Pr\big[X = Y\big] &= \sum_{x\in\Omega} \omega^*(x,x) \\
&= \sum_{x\in\Omega} \min\{\mu(x),\eta(x)\} \\
&= \sum_{x\in A} \eta(x) + \sum_{x\in\overline{A}} \mu(x) \\
&= \eta(A) + \mu(\overline{A}) \\
&= 1 - \big(\mu(A) - \eta(A)\big) \\
&= 1 - \|\mu-\eta\|_{\mathrm{tv}} \ ,
\end{aligned}$$

as desired. □

3 The Proof of the Fundamental Theorem of Markov Chains

Now we prove Theorem 1, in three main steps:

Step I: we show that there exists some (but not necessarily unique) stationary distribution π;

Step II: we use Step I and aperiodicity to prove that $\lim_{t\to+\infty} p_x^{(t)} = \pi$ for every $x \in \Omega$; and

Step III: we use Step I and Step II to prove the uniqueness of the stationary distribution.

Proof: [Proof of Theorem 1] We prove the three steps in reverse order:

Step III. Assume by way of contradiction that there exists some other stationary distribution π', i.e., for which $\pi' = \pi' P_{\mathcal{X}}$. We can write π' in the standard basis as $\pi' = \sum_{x\in\Omega} \alpha_x e_x$ with $\sum_{x\in\Omega} \alpha_x = 1$. Then, we get

$$\pi' = \pi' P^t_{\mathcal{X}} = \left(\sum_{x\in\Omega} \alpha_x e_x\right) P^t_{\mathcal{X}} = \sum_{x\in\Omega} \alpha_x (e_x P^t_{\mathcal{X}}) \ .$$

However, the above equation implies that

$$\pi' = \lim_{t\to+\infty} \sum_{x\in\Omega} \alpha_x (e_x P^t_{\mathcal{X}}) = \sum_{x\in\Omega} \alpha_x \pi = \pi \ ,$$

which is a contradiction to the assumption that $\pi' \neq \pi$. (Note that we have used Step I to guarantee the existence of π and Step II to claim that $e_x P^t_{\mathcal{X}} \to \pi$ as $t \to +\infty$ for every $x \in \Omega$.)

Step II. This is the interesting step, and it is here that we will make use of the definitions and lemmas on total variation distance and coupling from the previous section. We need to prove that, for every $x \in \Omega$, $\lim_{t\to+\infty} p^{(t)}_x = \pi$.

Define $\Delta(t) \overset{\text{def}}{=} \max_{x\in\Omega} \|p^{(t)}_x - \pi\|_{\text{tv}}$. We will show that $\lim_{t\to+\infty} \Delta(t) = 0$.

First, fix two arbitrary elements x and y in Ω. We argue that $p^{(t)}_x$ and $p^{(t)}_y$ converge to the *same* distribution, which is the stationary distribution. Consider two copies $(X_t)_t$ and $(Y_t)_t$ of the same Markov chain $\mathcal{X}$, the first starting at $X_0 = x$ and the second at $Y_0 = y$. Let (X_t, Y_t) be an "independent but sticky" coupling of the two chains: they move independently, but, if at some time t, $X_t = Y_t$ then $X_s = Y_s$ for all $s \geq t$. Intuitively, the marginal distribution of this coupling is exactly the same as the original chain $\mathcal{X}$. Formally, $(X_t, Y_t)_t$ is a Markov chain on $\Omega \times \Omega$ with a transition probability matrix Q given by:

$$Q\big((x_1,y_1),(x_2,y_2)\big) = \begin{cases} P(x_1,x_2)P(y_1,y_2) & \text{if } x_1 \neq y_1 \\ P(x_1,x_2) & \text{if } x_1 = y_1 \text{ and } x_2 = y_2 \\ 0 & \text{if } x_1 = y_1 \text{ and } x_2 \neq y_2 \end{cases} \ .$$

Let $T \overset{\text{def}}{=} \min\{t : X_t = Y_t\}$ be a random variable indicating the earliest time the two chains meet. By the coupling lemma, for all t,

$$\|p^{(t)}_x - p^{(t)}_y\|_{\text{tv}} = \frac{1}{2}\sum_{z\in\Omega} \Big| \Pr\big[X_t = z\big] - \Pr\big[Y_t = z\big] \Big| \leq \Pr\big[X_t \neq Y_t\big] = \Pr\big[T > t\big] \ . \tag{1}$$

We now show that the right hand size tends to zero as $t \to +\infty$. Invoking the aperiodicity and irreducibility of $\mathcal{X}$, there exists some common critical time τ such that, for every two elements z_1 and z_2 in Ω, $P^\tau_{\mathcal{X}}(z_1,z_2) > 0$. Then, letting $C \overset{\text{def}}{=} \min_{z_1,z_2} P^\tau_{\mathcal{X}}(z_1,z_2) > 0$, we get that $P^\tau_{\mathcal{X}}(x_1,z) \cdot P^\tau_{\mathcal{X}}(y_1,z) \geq C^2$ for every tuple (x_1,y_1,z). This actually tells us that after τ time steps, X_τ and Y_τ are going to meet with probability at least C^2. (Note that the chains $(X_t)_t$ and $(Y_t)_t$ are not completely independent: they become correlated when they first meet. Therefore, it suffices to give a lower bound on the probability that they do meet, pretending that the two chains are independent.) Hence, considering integer multiples of τ, we get

$$\Pr[X_{k\tau} \neq Y_{k\tau}] \leq (1 - C^2)^k \to 0 \quad \text{as} \quad k \to +\infty \ .$$

Going back to Equation 1, we conclude that $\|p^{(t)}_x - p^{(t)}_y\|_{\text{tv}} \to 0$ as $t \to \infty$.

Next, define $D(t) \overset{\text{def}}{=} \max_{x,y\in\Omega} \|p^{(t)}_x - p^{(t)}_y\|_{\text{tv}}$. We claim that

$$\Delta(t) \leq D(t) \leq 2\Delta(t) \ .$$

The second inequality follows easily from the triangle inequality. The first one follows from linearity and, again, the triangle inequality: by Step I, we are guaranteed to have a stationary distribution, so we can write

$$\pi = \pi P^t_{\mathcal{X}} = \sum_{y\in\Omega} \alpha_y e_y P^t_{\mathcal{X}} = \sum_{y\in\Omega} \alpha_y p^{(t)}_y \ ,$$

then

$$\|p_x^{(t)} - \pi\|_{\mathrm{tv}} = \|p_x^{(t)} - \sum_{y\in\Omega} \alpha_y p_y^{(t)}\|_{\mathrm{tv}} = \|\sum_{y\in\Omega} \alpha_y p_x^{(t)} - \sum_{y\in\Omega} \alpha_y p_y^{(t)}\|_{\mathrm{tv}}$$
$$\leq \sum_{y\in\Omega} \alpha_y \|p_x^{(t)} - p_y^{(t)}\|_{\mathrm{tv}} \leq \sum_{y\in\Omega} \alpha_y D(t) = D(t) \ .$$

Therefore, we conclude that $\Delta(t)$ tends to zero as t goes to infinity, therefore completing Step II.

Step I. In this step we argue that there exists a stationary distribution for any irreducible Markov chain (and do not invoke its aperiodicity). For every $x \in \Omega$, define a distribution q_x over Ω as follows: set $q_x(x) = 1$ and, for $y \neq x$, $q_x(y)$ is defined as the expected number of times that the Markov chain starting at x visits y before coming back to x for the first time.

Exericse (1pt): Prove that, for any element $x \in \Omega$, the distribution $\pi \stackrel{\text{def}}{=} \frac{q_x}{\sum_z q_x(z)}$ is stationary for the Markov chain $\mathcal{X}$.

The above exercise completes the proof of Step I. □