AI2613 随机过程 2020–2021 春季学期

Lecture 5 – Examples and Applications of Markov Chains (II), Coupling

2021 年 3 月 22 日

Lecturer: 张驰豪 *Scribe*: 杨宽

1 2-SAT

Review of the last lecture:

2-SAT is the problem of determining whether a CNF formula

$$\phi = C_1 \wedge C_2 \wedge \cdots \wedge C_m$$

on variables V has satisfying assignments, where each clause C_i consists of exact 2 literals. Our goal is to find an assignment $\sigma \in \{0,1\}^V$ such that $\sigma \models \phi$.

Here we introduce a much simpler randomized algorithm that can solve this problem with high probability.

1. Let $V = \{v_1, v_2, \dots, v_n\}$ be the set of variables. Pick an arbitrary assignment $V \to \{\text{true}, \text{false}\}$.
2. If there exists a clause c that has not been satisfied yet, then pick one of two variables incident to c uniformly at random and flip its value. Repeat this step $100n^2$ times, or until there does not exist an unsatisfied clause.
3. The algorithm outputs no solution if there still exists an unsatisfied clause after running $100n^2$ times.

We claim that our algorithm outputs the correct answer with probability at least $1 - 1/100$.

Proof (cont'd). Let $\sigma : V \to \{\text{true}, \text{false}\}$ be a satisfying assignment and our algorithm produces $100n^2$ assignments $\sigma_0, \sigma_1, \dots, \sigma_{100n^2}$. We now show that the probability that there is no such k that $\sigma_k = \sigma$ is at most $1/100$.

Let X_i be a random variable that

$$X_i = \sum_{j=1}^{n} \mathbb{1}_{[\sigma_i(j) = \sigma(j)]}\,.$$

The algorithm starts with $X_0 \geq 0$ and ends as long as $X_m = n$ for some m. Note that

$$\mathbf{Pr}[X_{i+1} = X_i + 1] \geq 1/2\,, \text{ and}$$

$$\mathbf{Pr}[X_{i+1} = X_i - 1] \le 1/2 .$$

Here $\{X_i\}$ is not a Markov chain, but let's consider the following Markov chain $\{Y_t\}_{t\ge 0}$: $\{Y_t\}$ is a random walk on $\mathbb{N}$, and

$$\mathbf{Pr}[Y_{t+1} = Y_t + 1] = \mathbf{Pr}[Y_{t+1} = Y_t - 1] = 1/2 .$$

As we already know,

$$\mathbb{E}_Y\left[\text{first hitting time of } n \mid Y_0\right] = n^2 - Y_0^2 .$$

Intuitively, $\{X_i\}$ favors to be right (compared to $\{Y_t\}$), so the first hitting time of n in $\{X_t\}$ should be no later than the first hitting time of n in Markov chain $\{Y_t\}$ with $Y_0 = X_0$.
We will formalize the intuition in the next section, but now we just use it as a conclusion. So $\mathbb{E}_X\left[\text{first hitting time of } n \mid X_0\right] \le n^2 - X_0^2 \le n^2$. Thus, applying the Markov inequality,

$$\mathbf{Pr}_X\left[\text{first hitting time of } n \ge 100n^2\right] \le 1/100 . \qquad \square$$

2 Stochastic Dominance

Now we formalize the intuition used in the last section, which we call *stochastic dominance*:

Definition 1 (Stochastic Dominance). Given two distributions μ and ν over $\mathbb{R}$, we say that μ is *stochastically dominant* ν, denoted by $\mu \succcurlyeq \nu$, if for all $a \in \mathbb{R}$,

$$\mu\big([a,\infty)\big) \ge \nu\big([a,\infty)\big) ,$$

namely,

$$\mathbf{Pr}_{X\sim\mu}[X \ge a] \ge \mathbf{Pr}_{X\sim\nu}[X \ge a] .$$

Remark. Let F_μ and F_ν be the corresponding cumulative distribution functions of μ and ν respectively. Then Figure 1 shows the relation between F_μ and F_ν.

An interesting question is how we could prove stochastic dominance. But before introducing a method, let us see some examples of stochastic dominance first.
The first one is the analysis of our algorithm for the 2-SAT problem.

Example 2 (2-SAT). Suppose that $X_t \sim \mu_t$ and $Y_t \sim \nu_t$. Then $\mu_t \succcurlyeq \nu_t$.

The second example is also a simple one.

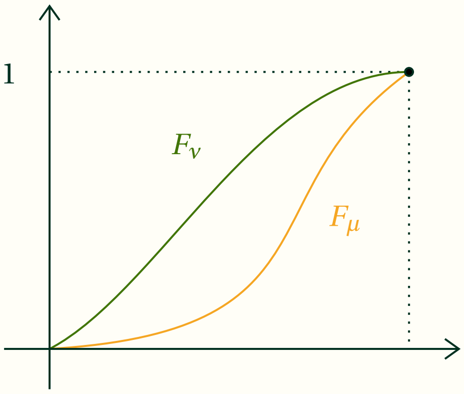

Figure 1: $\mu \succcurlyeq \nu$

Example 3 (Binomial distribution). Consider the binomial distribution $\mathsf{Bin}(n, p)$ of the number of success over n Bernoulli trials with success probability p. Then $\mathsf{Bin}(n, p) \succcurlyeq \mathsf{Bin}(n, q)$ if $p \geq q$.

Another example is an important random graph model.

Example 4 (Erdős-Rényi Random Graph). An Erdős-Rényi random graph, denoted by $\mathcal{G}(n, p)$, is a graph on n vertices and each pair of vertices are independently connected by an edge with probability p.

Let X_p be a random variable that $X_p = \mathbb{1}_{[G\sim\mathcal{G}(n,p) \text{ is connected}]}$ and μ_p be the distribution of X_p. Then $\mu_p \succcurlyeq \mu_q$ if $p \geq q$.

Intuitively, we claim the existence of stochastic dominance in the three examples above, but we haven't proved them.

To prove stochastic dominance, we now introduce a powerful tool.

Definition 5 (Coupling (耦合)). Let μ, ν be two distribution. A *coupling* $\mathcal{C}$ of μ, ν is a joint distribution of μ and ν.

Remark. Let $(X, Y) \sim \mathcal{C}$. Then we have

$$\begin{aligned} \forall\, x, \qquad & \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X = x] = \mu(x)\,; \\ \forall\, y, \qquad & \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[Y = y] = \mu(y)\,. \end{aligned}$$

Remark. Intuitively we can view a coupling as a way to fill in a table with nonnegative reals. For example, let $\Omega = \{1, 2, 3\}$, $\mu = (1/3, 1/3, 1/3)^\top$ and $\nu = (1/2, 1/4, 1/4)^\top$. A coupling $\mathcal{C}$ of μ and ν is a way to fill in the following table with nonnegative reals such that the summations of rows and columns are equal to the corresponding *marginal* probabilities.

ν μ	1/2	1/4	1/4
1/3			
1/3			
1/3			

Note that any joint distribution is a coupling, so there are infinite many couplings. We now introduce a special one.

Definition 6 (Monotone Coupling). $\mathcal{C}$ is a *monotone coupling* of μ and ν if

$$\mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \geq Y] = 1\,.$$

Remark. A monotone coupling corresponds a way to fill in the table where all positive numbers are only in the bottom left part (on or below the diagonal) of the table.

Question. Does a monotone couplings always exist?

Theorem 1. *There is a monotone coupling of μ and ν if and only if $\mu \succcurlyeq \nu$.*

Proof of "$\Longrightarrow$". Suppose $\mathcal{C}$ is a monotone coupling of μ and ν. Then

$$\begin{aligned}
\mathbf{Pr}_{Y\sim\nu}[Y \geq a] &= \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[Y \geq a] \\
&= \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \geq Y \wedge Y \geq a] + \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X < Y \wedge Y \geq a] \\
&= \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \geq Y \geq a] \\
&\leq \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \geq a] = \mathbf{Pr}_{X\sim\mu}[X \geq a]\,.
\end{aligned}$$

□

Remark. Intuitively, for the other direction, we can always fill in the table greedily. The rigorous proof is left as an exercise.

Then the theorem tells us if we can construct a monotone coupling, then we can prove stochastic dominance.

Now we construct monotone couplings for our Examples 2, 3 and 4.

2 Construct by induction. Let $X_0 = Y_0$. Assume that there exists a coupling $\mathcal{C}_t$ s.t. if $(X_t, Y_t) \sim \mathcal{C}_t$ then $\mathbf{Pr}[X_t \geq Y_t] = 1$. Construct $\mathcal{C}_{t+1}$ as follows. Note that $\mathbf{Pr}[X_{t+1} = X_t + 1] = u$ for some $u \geq 1/2$, and $\mathbf{Pr}[Y_{t+1} = Y_t + 1] = \mathbf{Pr}[Y_{t+1} = Y_t - 1] = 1/2$. So we pick a real r in $[0,1]$ uniformly at random. Let $X_{t+1} = X_t + 1$ if $r \leq u$ and $X_{t+1} = X_t - 1$ otherwise. Similarly, let

$Y_{t+1} = Y_t + 1$ if $r \le 1/2$ and $Y_{t+1} = Y_t - 1$ otherwise. Since $u \ge 1/2$, $X_{t+1} - X_t \ge Y_{t+1} - Y_t$, and thus $X_{t+1} \ge Y_{t+1}$.

3 $\mathrm{Bin}(n,p)$ is the distribution of the number of success over n independent trials, so we can consider every trial independently. Suppose that $X \sim \mathrm{Bin}(1,p)$ and $Y \sim \mathrm{Bin}(1,q)$. Again we pick a real r in $[0,1]$ uniformly at random. Then let $X = 1$ iff $r \le p$ and let $Y = 1$ iff $r \le q$. So $X \ge Y$.

4 Let $G_p \sim \mathcal{G}(n,p)$ and $G_q \sim \mathcal{G}(n,q)$. We generate G_p and G_q simultaneously. For each pair of vertices (u,v) we independently pick a real r in $[0,1]$ uniformly at random. Then G_p has edge (u,v) iff $r \le p$ and G_q has edge (u,v) iff $r \le q$. Thus G_q is a subgraph of G_p as long as $p \ge q$. So $X_p \ge X_q$. Moreover applying the proof of Theorem 1 we obtain that

$$\mathbf{Pr}_{G\sim\mathcal{G}(n,p)}[G \text{ is connected}] \ge \mathbf{Pr}_{G\sim\mathcal{G}(n,q)}[G \text{ is connected}]\,.$$

3 Coupling Lemma

Definition 7 (Total Variation Distance). Let Ω be a sample space and $\mu, \nu \in [0,1]^\Omega$ be two distributions. Then the *total variation distance* of μ and ν is given by the half of the L^1-norm of $\mu - \nu$:

$$\left\|\mu-\nu\right\|_{\mathrm{TV}} \triangleq \frac{1}{2}\sum_{x\in\Omega}\left|\mu(x)-\nu(x)\right|.$$

Equivalently, we can also define the total variation distance as

$$\left\|\mu-\nu\right\|_{\mathrm{TV}} \triangleq \max_{A\subseteq\Omega} \mu(A)-\nu(A)\,.$$

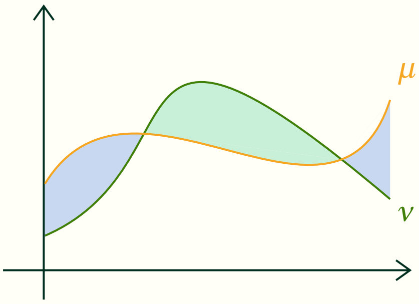

Figure 2: The total variation distance between μ and ν

Remark. Note that $\int \mu \,\mathrm{d}x = \int \nu \,\mathrm{d}x$. So $\int_{\mu(x)\geq\nu(x)} \mu(x) - \nu(x) \,\mathrm{d}x = \int_{\mu(x)\leq\nu(x)} \nu(x) - \mu(x) \,\mathrm{d}x$. See, for example, Figure 2. The area of blue part is equal to the area of green part, and that is why the first definition of the total variation distance has coefficient $1/2$.

The following theorem reveals the connection between coupling and the total variation distance, and thus is a powerful tool to compute the total variation distance.

Theorem 2 (Coupling Lemma). *$\forall$ coupling $\mathcal{C}$ of distributions μ and ν. Then,*

$$\mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \neq Y] \geq \left\|\mu - \nu\right\|_{\mathrm{TV}}.$$

Moreover, there exists an (optimal) coupling $\mathcal{C}^*$ that achieves the equality.

Proof. Let's view the coupling as a way to fill in the table. Then the probability $\mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X = Y]$ is the summation of numbers on diagonal. Intuitively, the number on i-th row and j-th column is upper bounded by $\mu(i)$ and $\nu(j)$, so the summation of numbers on diagonal should be upper bounded by $\sum_x \min\{\mu(x), \nu(x)\}$. This intuition gives a proof of the coupling lemma directly:

$$\begin{aligned}\mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \neq Y] &= 1 - \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X = Y] \\ &= 1 - \sum_{z\in\Omega} \mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X = Y = z] \\ &\geq 1 - \sum_{z\in\Omega} \min\{\mu(z), \nu(z)\} \\ &= \sum_{z\in\Omega} \big(\mu(z) - \min\{\mu(z), \nu(z)\}\big) = \left\|\mu - \nu\right\|_{\mathrm{TV}}. \end{aligned}$$

□

Note that we only prove the lower bound of $\mathbf{Pr}_{(X,Y)\sim\mathcal{C}}[X \neq Y]$ here. The proof for the existence of the optimal coupling is left as an exercise.

Finally, let's prove the Fundamental Theorem for Markov Chains.

Theorem 3 (Fundamental Theorem for Markov Chains). $(\mathrm{I}) + (\mathrm{A}) + (\mathrm{PR}) \implies (\mathrm{S}) + (\mathrm{U}) + (\mathrm{C})$.

Proof. We already know that $(\mathrm{I}) + (\mathrm{PR}) \implies (\mathrm{S}) + (\mathrm{U})$. So we only need to prove convergence here.

First we should characterize the convergence. We already know that there exists a unique stationary distribution π. What we would like to show is that for all starting distribution μ_0, it holds that

$$\lim_{t\to\infty} \left\|\mu_t - \pi\right\|_{\mathrm{TV}} = 0,$$

where ${\mu_t}^\top = {\mu_0}^\top \cdot \mathbf{P}^t$.

Suppose that $\{X_t\}$ and $\{Y_t\}$ are two identical Markov chains starting from different distribution, where $Y_0 \sim \pi$ while X_0 is generated from an arbitrary distribution μ_0.

Now we have two sequence of random variables:

$$\begin{tikzcd}
\mu_0 & \mu_1 & & & \mu_t & & \\
\wr & \wr & & & \wr & & \\
X_0 \arrow[r] & X_1 \arrow[r] & X_2 \arrow[r] & \cdots \arrow[r] & X_t \arrow[r] & X_{t+1} \arrow[r] & \cdots \\
Y_0 \arrow[r] & Y_1 \arrow[r] & Y_2 \arrow[r] & \cdots \arrow[r] & Y_t \arrow[r] & Y_{t+1} \arrow[r] & \cdots \\
\wr & \wr & & & \wr & & \\
\pi & \pi & & & \pi & &
\end{tikzcd}$$

The coupling lemma establishes the connection between the distance of distributions and the discrepancy of random variables. To show that $\|\mu_t - \pi\|_{\mathrm{TV}} \to 0$, it is sufficient to construct a coupling $\mathcal{C}_t$ of μ_t and π and then compute $\mathbf{Pr}[X_t \neq Y_t]$.

Here we give a simple coupling. Let $(X_t, Y_t) \sim \mathcal{C}_t$ and we construct $\mathcal{C}_{t+1}$. If $X_t = Y_t$ for some $t \geq 0$, then let $X_{t'} = Y_{t'}$ for all $t > t'$, otherwise X_{t+1} and Y_{t+1} are independent. Namely, $\{X_t\}$ and $\{Y_t\}$ are two independent Markov chains until X_t and Y_t reach the same state for some $t \geq 0$, and once they meet together then they move together forever. The coupling lemma tells us that $\|\mu_t - \pi\|_{\mathrm{TV}} \leq \mathbf{Pr}[X_t \neq Y_t]$.

We now prove the theorem for the finite case, i.e., the state space of the Markov chain is a finite set. Let's review the property of (I) and (A). The property (I) implies that

$$\forall\, i, j, \quad \exists\, n \quad \text{s.t.} \quad \mathbf{P}^n(i,j) > 0\,.$$

We claim that combining with property (A),

$$\exists\, n \quad \text{s.t.} \quad \forall\, i, j, \quad \mathbf{P}^n(i,j) > 0\,.$$

Since the state space $\mathcal{S}$ is finite, it is sufficient to show that

$$\forall\, i, j, \quad \exists\, t_{i,j} \quad \text{s.t.} \quad \forall\, n > t_{i,j}, \quad \mathbf{P}^n(i,j) > 0\,.$$

Suppose that there are s loops of length $c_1, c_2, \ldots, c_s$ starting from and ending at state i. Then property (A) implies that

$$\gcd(c_1, c_2, \ldots, c_s) = 1\,.$$

Thus, by Bézout's theorem there exists $x_1, x_2, \ldots, x_s \in \mathbb{Z}$ such that

$$c_1 x_1 + c_2 x_2 + \cdots c_s x_s = 1\,.$$

It yields straightforwardly that there exists $y_1, y_2, \ldots, y_s \in \mathbb{N}$ such that

$$c_1 y_1 + c_2 y_2 + \cdots c_s y_s = b$$

for all sufficiently large b. (For example, $b \geq (|x_1| + |x_2| + \cdots + |x_s|)(c_1 + c_2 + \cdots + c_s)^2$ suffices.) Combining with property (I), we complete the proof for our claim.

Now we know that $\exists\, n$ s.t. $\forall\, i, j$, $\mathbf{P}^n(i,j) > 0$. Then we define

$$\theta \triangleq \min_{x_0, y_0 \in \mathcal{S}} \mathbf{Pr}\left[X_n = Y_n \mid X_0 = x_0, Y_0 = y_0\right].$$

For simplicity, we use $\mathbf{Pr}_{x_0,y_0}[\cdot]$ to denote the conditional probability $\mathbf{Pr}\left[\cdot \mid X_0 = x_0, Y_0 = y_0\right]$ from now on.

Fix $z \in \mathcal{S}$. Let

$$\alpha = \min_{w \in \mathcal{S}} \mathbf{P}^n(w, z) > 0,$$

and for any $t \geq 0$ and $z' \in \mathcal{S}$,

$$\beta_{t,z'} = \mathbf{Pr}_{x_0,y_0}\left[X_t = Y_t = z' \wedge X_{t'} \neq Y_{t'} \text{ for all } t' < t\right].$$

By the Markov property and the independence of $\{X_t\}$ and $\{Y_t\}$ before $X_t = Y_t$, we obtain that

$$\begin{aligned}
\mathbf{Pr}_{x_0,y_0}[X_n = Y_n] &\geq \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z] \\
&= \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z \wedge \forall\, t < n, X_t \neq Y_t] + \mathbf{Pr}_{x_0,y_0}[X_n = Y_n = z \wedge \exists\, t < n, X_t = Y_t] \\
&= \left(\mathbf{P}^n(x_0, z) \cdot \mathbf{P}^n(y_0, z) - \sum_{t=0}^{n-1} \sum_{z'} \beta_{t,z'} \cdot \left(\mathbf{P}^{n-t}(z', z)\right)^2\right) + \sum_{t=0}^{n-1} \sum_{z'} \beta_{t,z'} \cdot \mathbf{P}^{n-t}(z', z) \\
&\geq \mathbf{P}^n(x_0, z) \cdot \mathbf{P}^n(y_0, z) \ \geq \alpha^2.
\end{aligned}$$

Hence $\theta > 0$. By the coupling and the Markov property, we have

$$\begin{aligned}
\mathbf{Pr}_{x_0,y_0}[X_{2n} \neq Y_{2n}] &= \sum_{x_n \neq y_n} \mathbf{Pr}_{x_0,y_0}\left[X_{2n} \neq Y_{2n}, X_n = x_n, Y_n = y_n\right] \\
&= \sum_{x_n \neq y_n} \mathbf{Pr}_{x_n,y_n}[X_n \neq Y_n] \cdot \mathbf{Pr}_{x_0,y_0}\left[X_n = x_n, Y_n = y_n\right] \\
&\leq (1-\theta) \sum_{x_n \neq y_n} \mathbf{Pr}_{x_0,y_0}\left[X_n = x_n, Y_n = y_n\right] \ \leq (1-\theta)^2,
\end{aligned}$$

and so on ($\mathbf{Pr}_{x_0,y_0}[X_{kn} \neq Y_{kn}] \leq (1-\theta)^k$). It yields directly that

$$\mathbf{Pr}[X_t \neq Y_t] = \sum_{x_0, y_0} \mu_0(x_0) \cdot \pi(y_0) \cdot \mathbf{Pr}_{x_0,y_0}[X_t \neq Y_t] \to 0$$

as $t \to \infty$. So we conclude that $\lim_{t\to\infty} \left\|\mu_t - \pi\right\|_{\mathrm{TV}} = 0$ for the finite case. □

Remark. In our proof for the finite case, we use the following lemma.

Theorem 4 (Bézout's Theorem). *Let $a, b \in \mathbb{Z}$ be any two integers, then there exist $u, v \in \mathbb{Z}$ such that*

$$au + bv = \gcd(a, b)\,.$$

However, for the infinite case, the proof is a bit more complicated. The complete proof can be found in the reference materials. Here we only give a brief summary.

Proof for the infinite case. Suppose the state space $\mathcal{S}$ is countably infinite. Let $P : \mathcal{S} \times \mathcal{S} \to [0, 1]$ be the transition function. Assume that P is $(\mathtt{I})$ (irreducible), (A) (aperiodic) and (PR) (positive-recurrent).

Similar to the finite case, we run two chains $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 0}$ independently with $X_0 \sim \mu_0$ and $Y_0 \sim \pi$, and couple them once they meet. We consider the transition function $Q\big((\cdot,\cdot),(\cdot,\cdot)\big)$ of the product chain $Z_t = (X_t, Y_t)$ before $X_t = Y_t$ (so that both chain run independently).

Notice that if Q is $(\mathtt{I})$ and (PR), then we are done. The theorem follows directly from $P_{i,j}(T_{k,k} < \infty) = 1$ for any $i, j, k \in \mathcal{S}$, where $T_{k,k}$ is the first hitting time of (k, k) in Q and $P_{i,j}$ is the probability conditioned on $(X_0, Y_0) = (i, j)$.

First we prove that Q is $(\mathtt{I})$. For any i, j, k, ℓ, we would like to find a certain n such that

$$Q^n\big((i,j),(k,\ell)\big) = P^n(i,k) \cdot P^n(j,\ell) > 0\,.$$

Similar to the finite case, (A) and $(\mathtt{I})$ for P imply that for any j, k, and sufficiently large n, it holds that $P^n(j,k) > 0$ and therefore concludes the proof.

Next we prove that Q is (PR). This is trivial given $(\mathtt{I})$ since Q has a stationary distribution. □