ADVANCED ALGORITHMS (XI)

CHIHAO ZHANG

1. Basic Notations

Let us first review some notations for distributions and Markov chains. Let Ω be a finite state space and μ_1, μ_2 be two distributions on Ω. The total variation distance between them is defined as

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \frac{1}{2} \sum_{x \in \Omega} \left| \mu_1(x) - \mu_2(x) \right|.$$

It is easy to verify that

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \max_{A \subseteq \Omega} \left| \mu_1(A) - \mu_2(A) \right|.$$

Let P be the transition matrix of an irreducible and aperiodic Markov chain with stationary distribution π. For every $\varepsilon > 0$, the mixing time $\tau(\varepsilon)$ of P is defined to be

$$\tau(\varepsilon) = \max_{\mu} \min_{t \geq 0} d_{\mathrm{TV}}(\mu^T P^t, \pi) \leq \varepsilon.$$

In other words, we have that the total variation distance between $\mu^T P^{\tau(\varepsilon)}$ and the stationary distribution π is at most ε for *any* initial distribution μ. We leave as an exercise to show that the distance

$$d_{\mathrm{TV}}(\mu^T P^t, \pi)$$

is non-increasing in t, so the mixing time is well-defined.

2. Coupling and Markovian Coupling

Today we will talk about *coupling*, an important tool to analyze the mixing time of Markov chains. Given two distributions μ_1 and μ_2 over the same space Ω, a coupling μ of μ_1 and μ_2 is a joint distribution μ over Ω^2 such that for every $(X, Y) \sim \mu$, the marginal distributions of $(\cdot, Y)$ and $(X, \cdot)$ are μ_1 and μ_2 respectively. Formally, we require that for every $v \in \Omega$, it holds that

$$\mathbf{Pr}_{(X,Y)\sim\mu}[X = v] = \mu_1(v) \text{ and } \mathbf{Pr}_{(X,Y)\sim\mu}[Y = v] = \mu_2(v).$$

The name "coupling" comes from the fact that in many applications, we prefer those joint distributions μ with large probability of $X = Y$ when $(X, Y) \sim \mu$. We say a coupling μ *optimal* if it maximizes the probability of $X = Y$ when $(X, Y) \sim \mu$.

When Ω is finite, it is natural to view the disjoint distribution μ as a matrix $M_\mu \in [0,1]^{|\Omega| \times |\Omega|}$ such that $M_\mu(i, j) = \mu(i, j)$. The condition for coupling also naturally translates to

- for every $i \in \Omega$, the sum of the probabilities in row i is equal to $\mu_1(i)$, namely
$$\sum_{j \in \Omega} M_\mu(i, j) = \mu_1(i);$$
- for every $j \in \Omega$, the sum of the probabilities in column j is equal to $\mu_2(j)$, namely
$$\sum_{i \in \Omega} M_\mu(i, j) = \mu_2(j).$$

The optimal coupling is therefore the one maximizing the trace $\sum_{i \in \Omega} M_\mu(i, i)$. The coupling lemma says that the optimal coupling captures the total variation distance of two distributions.

Theorem 1 (Coupling Lemma)**.**

$$d_{TV}(\mu_1, \mu_2) = \min_{\text{coupling } \mu} \mathbf{Pr}_{(X,Y)\sim\mu}[X \neq Y].$$

A special family of coupling with respect to Markov chains is called Markovian couplings, or couplings of Markov chains. In our time-homogeneous setting, we view such a coupling as two runs of chains $X_0, X_1, \ldots$ and $Y_0, Y_1, \ldots$ with the same transition matrix P. It is required that, for every $t \geq 0$, the transition from (X_t, Y_t) to (X_{t+1}, Y_{t+1}) is P while being viewed marginally. Formally, we require that for every $t \geq 0$ and $z, z' \in \Omega$,

$$\begin{aligned}
&\mathbf{Pr}\left[(X_{t+1}, Y_{t+1}) = (z', \cdot) \;\middle|\; (X_t, Y_t) = (z, \cdot)\right] = P(z, z'), \text{ and} \\
&\mathbf{Pr}\left[(X_{t+1}, Y_{t+1}) = (\cdot, z') \;\middle|\; (X_t, Y_t) = (\cdot, z)\right] = P(z, z').
\end{aligned}$$

Moreover, we require that once the two chains reach the same state, they will stay the same ever since. Formally, if for some $t' > 0$ we have $X_{t'} = Y_{t'}$, then we require $X_t = Y_t$ for every $t > t'$. This definition justifies the use of the term "coupling".

Assume $\{X_i\}_{i\geq 0}$ and $\{Y_i\}_{i\geq 0}$ are two coupled chains such that $X_0 \sim \mu_X$ and $Y_0 \sim \mu_Y$. Then it is easy to verify that the distribution of (X_t, Y_t) is a coupling of $\mu_X^T P^t$ and $\mu_Y^T P^t$ for every $t \geq 0$.

We can use couplings to prove the convergence theorem of Markov chains, which you already met in Lecture 9.

Theorem 2. *If a finite time-homogeneous chain P is irreducible and aperiodic, then it has a unique stationary distribution π. Moreover, for any initial distribution μ, it holds that*

$$\lim_{t\to\infty} \mu^T P^t = \pi^T.$$

Proof. We know that the conditions of the irreducibility and the aperiodicity imply that for some $t > 0$, the matrix P^t satisfies $P^t(x, y) > 0$ for every pair of states (x, y). We now assume a coupling of two chains $\{X_i\}_{i\geq 0}$ and $\{Y_i\}_{i\geq 0}$ where both chains run independently. Initially, $Y_0 \sim \pi$ and X_0 is arbitrary. The disjoint distribution of these two independent chains is of course a coupling. Then for some $z \in \Omega$, it holds that

$$\mathbf{Pr}\left[X_t = Y_t\right] \geq \mathbf{Pr}\left[X_t = Y_t = z\right] = \mathbf{Pr}\left[X_t = z\right] \cdot \mathbf{Pr}\left[Y_t = z\right] = P^t(X_0, z) \cdot \pi(z) \geq \theta > 0, \tag{1}$$

where θ is some constant larger than zero. This is equivalent to $\mathbf{Pr}\left[X_t \neq Y_t\right] \leq 1-\theta$. By the definition of the coupling, we have

$$\begin{aligned}
\mathbf{Pr}\left[X_{2t} \neq Y_{2t}\right] &= \mathbf{Pr}\left[X_{2t} \neq Y_{2t} \wedge X_t = Y_t\right] + \mathbf{Pr}\left[X_{2t} \neq Y_{2t} \wedge X_t \neq Y_t\right] \\
&= \mathbf{Pr}\left[X_{2t} \neq Y_{2t} \mid X_t \neq Y_t\right] \cdot \mathbf{Pr}\left[X_t \neq Y_{2t}\right] \\
&\leq (1-\theta)^2,
\end{aligned}$$

where the last inequality follows from the same argument in eq. (1). We can then repeat the argument and show that for every $k > 0$,

$$\mathbf{Pr}\left[X_{kt} \neq Y_{kt}\right] \leq (1-\theta)^k.$$

Therefore by the coupling lemma (theorem 1), we have X_i converges to π when i tends to infinity. □

In fact, in addition to the convergence in the limit, the coupling method can be used to bound the mixing time of a chain. By the definition of the mixing time, we have for every initial distribution μ,

$$d_{\mathrm{TV}}(\mu^T P^{\tau(\varepsilon)}, \pi) \leq \varepsilon.$$

Therefore, if we are able to construct a Markovian coupling $(X_i, Y_i)_{i\geq 0}$ with arbitrary initial state (X_0, Y_0) such that

$$\mathbf{Pr}\left[X_t \neq Y_t\right] \leq \varepsilon,$$

then we can conclude that $\tau(\varepsilon) \leq t$.

3. Proof of Mixing

A hypercube of dimension n is a graph $G(V, E)$ whose vertex set is $\{0,1\}^n$ and two vertices $x, y \in V$ are connected iff the ℓ_1 distance $\|x - y\|_1 = 1$. Consider the following random walk on a hypercube: each step when one stands at a state $x \in \{0,1\}^n$,

- with probability $\frac{1}{2}$, do nothing;
- otherwise, choose an index $i \in [n]$ u.a.r. and flip the value $x(i)$.

We can use coupling to bound the mixing time of this random walk. The random walk can be viewed in the following equivalent way, which is more convenient for us to design a coupling: each step when one stands at a state $x \in \{0,1\}^n$,

- choose an index $i \in [n]$ u.a.r. and $b \in \{0,1\}$ u.a.r.;
- change $x(i)$ to b.

We shall analyze the following coupling: Given a pair of states (X_t, Y_t),

- Choose an index $i \in [n]$ u.a.r. and $b \in \{0,1\}$ u.a.r.
- Change $X_t(i)$ to b.
- Change $Y_t(i)$ to b.

In this coupling, as long as we choose some index i at step s, then $X_t(i) = Y_t(i)$ for every $t > s$. This fact implies that the coupling process is equivalent to a *coupon collector* process, and the two chains are coupled if and only if we collect all n coupons. We know that for coupon collector process, if we randomly sample more than $n \log n + cn$ coupons, then the probability that we do not own all coupons is at most e^{-c}. This means that in our coupling,

$$\tau(\varepsilon) \leq n \log n + n \log \varepsilon^{-1}.$$

Now we consider a more sophisticated example, the Glauber dynamics for sampling proper colorings. In this problem, we are given an undirected graph $G = (V, E)$ with maximum degree Δ and q colors. We use the following Markov chain to sample a proper coloring of G: start from any proper coloring $\sigma \in [q]^V$,

- Choose a vertex $v \in V$ u.a.r.
- Sample a color c from all proper colors at v.
- Change the color of v to c.

In step two, the proper colors at v are colors in $[q]$ excluding those used by neighbours of v. The Markov chain is always aperiodic and when $q > \Delta + 1$, it is also irreducible (verify this!). It is also not hard to verify that the uniform distribution on all proper colorings is the stationary distribution and let us denote it by π (by looking at the detailed balance condition). Now we construct a coupling of the chain: Given a pair of colorings (X_t, Y_t),

- Choose a vertex $v \in V$ u.a.r.
- Let $L_X(v), L_Y(v)$ be the set of proper colors at v in X_t and Y_t respectively. Let μ_X and μ_Y be the uniform distribution on $L_X(v)$ and $L_Y(v)$ respectively.
- Color $X_t(v)$ and $Y_t(v)$ with a pair of colorings (c_X, c_Y) sampled from the optimal coupling of μ_X, μ_Y.

We need to further explain the third step. Since both $L_X(v)$ and $L_Y(v)$ are subsets of $[q]$, the two uniform distributions μ_X and μ_Y satisfy $\mu_X(c) = \frac{1}{L_X(v)}$ and $\mu_Y(c) = \frac{1}{L_Y(v)}$ for a color c in $L_X(v)$ and $L_Y(v)$ respectively. It is not hard to verify that, in the optimal coupling μ of μ_X and μ_Y, we have

$$\mathbf{Pr}_{(c_X, c_Y) \sim \mu}[c_X = c_Y] = \frac{\left|L_X(v) \cap L_Y(v)\right|}{\max\left\{\left|L_X(v)\right|, \left|L_Y(v)\right|\right\}}. \tag{2}$$

For every (X_t, Y_t) in our coupling, we use $d(X_t, Y_t)$ to denote the number of vertices on which X_t and Y_t differ, namely $d(X_t, Y_t) = \sum_{v \in V} \mathbf{1}(X_t(v) \neq Y_t(v))$. We shall bound the random variable $\mathbf{E}\left[d(X_{t+1}, Y_{t+1}) \,\middle|\, (X_t, Y_t)\right]$, which is the expected distance after one step. We first divides V into two disjoint sets A_t and D_t where $A_t = \left\{v \in V : X_t(v) = Y_t(v)\right\}$ and $D_t = \left\{v \in V : X_t(v) \neq Y_t(v)\right\}$. We use m to denote the number of edges between A_t and D_t. For every $v \in A_t$, we let $d(v) \triangleq \{u \in D_t : \{u, v\} \in E\}$ and for every $v \in D_t$, we let $d(v) \triangleq \{u \in A_t : \{u, v\} \in E\}$. Then it is clear that

$$\sum_{v \in A_t} d(v) = \sum_{v \in D_t} d(v) = m.$$

In the first step of our coupling, we get a vertex v that is either in A_t or in D_t.

- The quantity $d(X_{t+1}, Y_{t+1})$ is equal to $d(X_t, Y_t) + 1$ only if v is in A_t and in the third step, the coupling fails, namely we have $c_X \neq c_Y$. We assume w.l.o.g. that $\left|L_X(v)\right| \geq \left|L_Y(v)\right|$. It follows from eq. (2) that $c_X \neq c_Y$ happens with probability

$$1 - \frac{\left|L_X(v) \cap L_Y(v)\right|}{\left|L_X(v)\right|} \leq 1 - \frac{L_X(v) - d(v)}{L_X(v)} = \frac{d(v)}{L_X(v)} \leq \frac{d(v)}{q - \Delta}.$$

  Therefore, the probability that $d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1$ is at most

$$\frac{1}{n} \sum_{v \in D_t} \frac{d(v)}{q - \Delta} = \frac{m}{n(q - \Delta)}.$$

- The quantity $d(X_{t+1}, Y_{t+1})$ is equal to $d(X_t, Y_t) - 1$ only if v is in D_t and in the third step, the coupling succeeds, namely we have $c_X = c_Y$. This event happens with probability

$$\frac{\left|L_X(v) \cap L_Y(v)\right|}{\left|L_X(v)\right|} \geq 1 - \frac{\Delta - d(v)}{\left|L_X(v)\right|} \geq \frac{q - 2\Delta + d(v)}{q - \Delta}.$$

  Therefore, the probability that $d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1$ is at least

$$\frac{1}{n} \sum_{v \in A_t} \frac{q - 2\Delta + d(v)}{q - \Delta} = \frac{(q - 2\Delta)\,|D_t| + m}{n(q - \Delta)}.$$

Therefore, if we let $a = \frac{m}{n(q-\Delta)}$, $b = \frac{(q-2\Delta)|D_t|+m}{n(q-\Delta)}$, then

$$\begin{aligned}
\mathbf{E}\left[d(X_{t+1}, Y_{t+1}) \,\middle|\, (X_t, Y_t)\right] &\leq a \cdot (d(X_t, Y_t) + 1) + b \cdot (d(X_t, Y_t) - 1) + (1 - a - b)\, d(X_t, Y_t) \\
&= a - b + d(X_t, Y_t) \\
&= \left(1 - \frac{q - 2\Delta}{n(q - \Delta)}\right) \cdot d(X_t, Y_t).
\end{aligned}$$

Therefore, if we have $q \geq 2\Delta + 1$, then for every $t \geq 0$,

$$\mathbf{E}\left[d(X_{t+1}, Y_{t+1}) \,\middle|\, (X_t, Y_t)\right] \leq \left(1 - \frac{1}{2n}\right) d(X_t, Y_t).$$

Taking expectation on both sides implies

$$\mathbf{E}\left[d(X_{t+1}, Y_{t+1})\right] \leq \left(1 - \frac{1}{2n}\right) \mathbf{E}\left[d(X_t, Y_t)\right]$$

holds for every $t \geq 0$. Since $d(X_0, Y_0) \leq n$, we have

$$\mathbf{E}\left[d(X_t, Y_t)\right] \leq n \left(1 - \frac{1}{2n}\right)^t \leq n e^{-\frac{t}{2n}}.$$

If we let $X_0 \sim \mu$ and $Y_0 \sim \pi$, then for $t = 2n \log \frac{n}{\varepsilon}$, we have

$$d_{\mathrm{TV}}(\pi^T P^t, \pi) \leq \mathbf{Pr}\left[X_t \neq Y_t\right] = \mathbf{Pr}\left[\left|d(X_t, Y_t)\right| \geq 1\right] \leq \mathbf{E}\left[d(X_t, Y_t)\right] \leq \varepsilon.$$

This means $\tau(\varepsilon) \leq 2n \log \frac{n}{\varepsilon}$.