Lower Bounds on Mixing Times

Yanheng Wang

Intuition

We want to bound t_{mix} from below. What are the possible reasons accounting for a large t_{mix}?

- The state space is too large and cannot be covered in a short period.
- The state space has "bottlenecks", so we have difficulty in reaching certain states.
- ...

We shall describe three methods that exploit the structure of state space. We will also introduce another method that directly investigates the evolution of Markov chain.

Method 1: Cardinality

Suppose the stationary distribution π is uniform over $\mathcal{X}$. Let G be the transition graph of the chain. Denote by Δ its maximum out-degree. What are the states that can be reached at time t? Clearly no more than Δ^t.

Remark. If the chain is reversible, then G is undirected and we may attain a better upperbound $1 + \Delta \cdot \sum_{i=1}^{t-1} (\Delta - 1)^i$.

Theorem. $t_{\mathrm{mix}}(\epsilon) \geq \log_\Delta((1-\epsilon)|\mathcal{X}|)$.

Proof. Let t equals RHS, and take S to be the set of all reachable states at time t. Then $|S| \leq \Delta^t \leq (1-\epsilon)|\mathcal{X}|$, i.e. only a small portion of $\mathcal{X}$ is reachable. Then we have

$$\|P^t(x,\cdot) - \pi\| \geq |P^t(x,S) - \pi(S)| = 1 - \frac{|S|}{|\mathcal{X}|} \geq \epsilon \qquad \blacksquare$$

Method 2: Diameter

Theorem. For all $\epsilon < 1/2$ we have $t_{\mathrm{mix}}(\epsilon) \geq \mathrm{diam}(G)/2$.

Proof. Take $x, y \in \mathcal{X}$ to be the most distant vertices on G. Let S_t^x and S_t^x stand for the set of reachable states at time t from x and y, respectively. Write $r := \mathrm{diam}(G)/2$. Obviously, $S_r^x \cap S_r^y = \emptyset$, hence

$$\|P^r(x,\cdot) - P^r(y,\cdot)\| \geq |P^r(x, S_r^x) - P^r(x, S_r^y)| = 1 - 0 = 1.$$

By relation between $\bar{d}$ and d, we know

$$\|P^r(x,\cdot) - \pi\| \geq 1/2$$

which implies $t_{\mathrm{mix}}(\epsilon) \geq r$ for all $\epsilon < 1/2$. ■

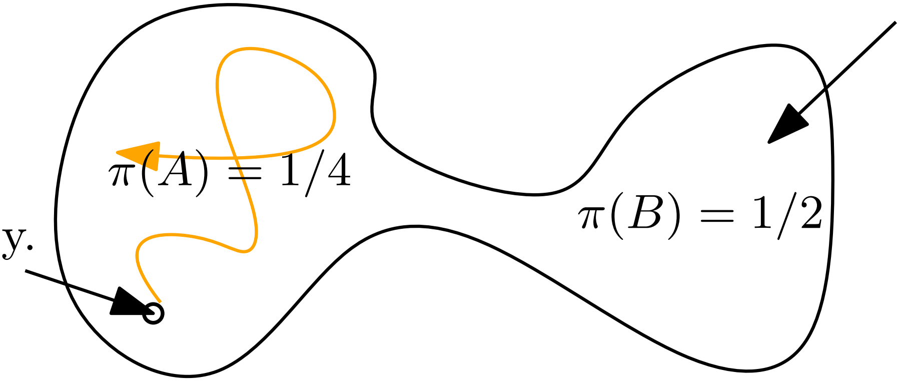
The previous two methods are almost trivial; they neglects the rich structure in the transition matrix as well as in the underlying graph.

The next method incorporates these structural information, and will be of practical interest.

Method 3: Bottleneck

Roughly speaking, a bottleneck in a graph is a narrow passage that bridges two or more components. But a mere appearance of bottleneck says nothing valuable, unless the *stationary probability* in some component is high. In that case, the chain will find itself “trapped” in a low-probability region and fail to get out.

Probability ought to be high in stationarity.



Probability concentrates here in reality.

Definition. We define the bottleneck ratio (or conductance) of $S \subseteq \mathcal{X}$ to be

$$\Phi(S) := \frac{\sum_{x \in S, y \in \overline{S}} \pi(x) P(x, y)}{\pi(S)} = \frac{\mathbb{P}_\pi \left(X_t \in S, X_{t+1} \in \overline{S} \right)}{\pi(S)}.$$

Clearly, the numerator indicates "escape probability from S".

Theorem. Let $\Phi_\star := \min_{\pi(S) < 1/2} \Phi(S)$. Then we have $t_{\text{mix}} \geq 1/(4\Phi_\star)$.

Remark. Why do we optimise over $S : \pi(S) < 1/2$? Intuitively, this corresponds to our discussion that S is a trapping region – one that ought to have low stationary distribution yet hard to escape in reality. We shall see in the proof that such requirement is useful (but can be relaxed).

Theorem. Let $\Phi_\star := \min_{\pi(S)<1/2} \Phi(S)$. Then we have $t_{\text{mix}} \geq 1/(4\Phi_\star)$.

Proof. Consider a stationary chain (X_t). Under what consequences can $X_t \in \overline{S}$? A cursory necessary condition is: the chain must move from S to $\overline{S}$ at some step before t. So

$$\mathbb{P}_\pi(X_t \in \overline{S}) \leq \sum_{i=0}^{t-1} \mathbb{P}_\pi\left(X_i \in S, X_{i+1} \in \overline{S}\right) = t \cdot \pi(S)\Phi(S)$$

Therefore,

$$\mathbb{P}_\pi\left(X_t \in \overline{S} \mid X_0 \in S\right) \leq t \cdot \Phi(S)$$

So there is some $x \in S$ such that $P^t(x, \overline{S}) \leq t\Phi(S)$. (Why?) In other words, $P^t(x, S) \geq 1 - t\Phi(S)$. Hence,

$$\|P^t(x, \cdot) - \pi\| \geq |P^t(x, S) - \pi(S)| \geq 1 - t\Phi(S) - \pi(S).$$

Note that the discussion above holds for all S. Under the condition that $\pi(S) < 1/2$, the total variation distance is always greater than $1/2 - t\Phi(S)$, giving $t_{\text{mix}} \geq 1/(4\Phi(S))$. ■

Method 4: Statistics

This method marks a diversion from the previous ones: It is more “mathematical” since it looks into the distribution at time t directly.

The intuition is simple: In attempt to prove $\|\mu - \nu\| \geq \epsilon$, we turn to design some appropriate statistical quantity to distinguish $X \sim \mu$ and $Y \sim \nu$. That is, we design a function $f : \mathcal{X} \to \mathbb{R}$ and try to argue that the random variables $f(X)$ and $f(Y)$ are “distant”. Since $f(X)$ and $f(Y)$ are real-valued, they are often much more convenient for discussion.

Theorem. Fix $f : \mathcal{X} \to \mathbb{R}$. Let $X \sim \mu$ and $Y \sim \nu$. If $|\mathbb{E}(fX) - \mathbb{E}(fY)| \geq r\sigma$, where $\sigma^2 = \max\{\mathrm{Var}(fX), \mathrm{Var}(fY)\}$, then $\|\mu - \nu\| \geq 1 - 8/r^2$.

Proof. This is just an application of Chebyshev's inequality. To avoid clutter, we denote $X' := fX$ and $Y' := fY$. Without loss of generality, assume $\mathbb{E}(X') \geq \mathbb{E}(Y')$. Take interval $I := (\mathbb{E}(Y') + r\sigma/2,\ \infty)$. By Chebyshev,

$$\mathbb{P}(X' \in I) \geq 1 - (2/r)^2$$

and

$$\mathbb{P}(Y' \in I) \leq (2/r)^2$$

We take $S := f^{-1}(I)$, then $\mu(S)$ and $\nu(S)$ correspond to the probabilities above, respectively. Therefore,

$$\|\mu - \nu\| \geq |\mu(S) - \nu(S)| \geq 1 - 8/r^2. \ \blacksquare$$

Remark. The converse is not true. If the f is poorly designed, it could diffuse the distinguishing features into $\mathbb{R}$.

The lower bound could be improved slightly:

Theorem. Fix $f : \mathcal{X} \to \mathbb{R}$. Let $X \sim \mu$ and $Y \sim \nu$. If $|\mathbb{E}(fX) - \mathbb{E}(fY)| \geq r\sigma$, where $\sigma^2 = [\mathrm{Var}(fX) + \mathrm{Var}(fY)]/2$, then $\|\mu - \nu\| \geq r^2/(4 + r^2)$.

Proof sketch. Note that the total variation distance is translation invariant. So we can always assume $\mathbb{E}(X') = E$ and $\mathbb{E}(Y') = -E$. Then $2E \geq r\sigma$ by condition.
Assume $X' \sim \alpha$ and $Y' \sim \beta$. Then define $r := \frac{2\alpha}{\alpha+\beta}$ and $s := \frac{2\beta}{\alpha+\beta}$. (They are not distributions!) Then we have

$$(2E)^2 = \big(\mathbb{E}(X') - \mathbb{E}(Y')\big)^2 = \left(\sum_{z\in\mathbb{R}} z(\alpha(z) - \beta(z))\right)^2$$

By Cauchy-Schwarz, this is less than

$$\left(\sum_z \frac{\alpha(z) + \beta(z)}{2} z^2\right)\left(\sum_z \frac{\alpha(z) + \beta(z)}{2} (r(z) - s(z))^2\right)$$

$$\begin{tikzcd}
\left(\sum_z \frac{\alpha(z)+\beta(z)}{2} z^2\right) \arrow[d] & \left(\sum_z \frac{\alpha(z)+\beta(z)}{2}(r(z)-s(z))^2\right) \arrow[d] \\
\frac{\mathbb{E}(X'^2)+\mathbb{E}(Y'^2)}{2} \arrow[d] & 2\sum_z \frac{(\alpha(z)-\beta(z))^2}{\alpha(z)+\beta(z)} \arrow[d, "|\wedge"] \\
\frac{\mathrm{Var}(X')+\mathrm{Var}(Y')+\mathbb{E}^2(X')+\mathbb{E}^2(Y')}{2} \arrow[d] & 2\sum_z |\alpha(z)-\beta(z)| \arrow[d] \\
\sigma^2+M^2 & 2\|\alpha-\beta\| \qquad \blacksquare
\end{tikzcd}$$

How can we design a proper statistical quantity? The rule of thumb is to “encode” the most important feature in a given problem.

For instance, in the random walk on hypercube $\{0, 1\}^d$, we can design f to be the Hamming distance. For detailed analysis, see the textbook by Levin, Peres and Wilmer.

Philosophical View

Just as in the relation between algorithm and complexity theory, the investigation on lower bound of mixing times leads to better understanding of the chains: it points out the restriction in chain design and suggests improvements. The negative results also motivate the classification of hardness.

It seems intriguing to explore the relations between certain families of chains – just as we did in relating different models of computing. If we could somehow find a universal model of Markov chains in the sense of running time, then it would be possible to classify the problems by their intrinsic hardness!