Markov chains, mixing times, and coupling analysis

USTC, Hefei, China, Jan 15, 2025

Lecturer: Weiming Feng (ETH Zürich)

1 Sampling from high-dimensional distributions

Let $[q] = \{0, 1, \ldots, q-1\}$ be a finite domain of size q. Let V be a set of variables of size n. Let π be a *high-dimensional distribution* with support

$$\Omega = \{\sigma \in [q]^V \mid \pi(\sigma) > 0\}.$$

Example 1.1 (running example: graph coloring)**.** *Let $G = (V, E)$ be a graph. Let $[q]$ be a set of colors. Let $\Omega \subseteq [q]^V$ be the set of all proper colorings of G. We use π to denote the uniform distribution over Ω, e.g. the uniform distribution over all proper colorings in G.*

Example 1.2 (hardcore model)**.** *Let $G = (V, E)$ be a graph. For any $\sigma \in \{0, 1\}^V$, we say σ is an independent set if all vertices $v \in V$ such that $\sigma_v = 1$ form an independent set in G. Let Ω denote the set of all independent sets in G. Let $\lambda > 0$ be a weight parameter. We define π as a distribution over Ω by*

$$\forall \sigma \in \Omega, \quad \pi(\sigma) = \frac{\lambda^{|\sigma|}}{\sum_{\tau \in \Omega} \lambda^{|\tau|}},$$

where $|\sigma| = \sum_{v \in V} \sigma_v$ is the 1-norm of σ.

Example 1.3 (Ising model)**.** *Let $J \in \mathbb{R}^{V \times V}$ be a symmetric matrix such that $J_{uv} = J_{vu}$. Let $h \in \mathbb{R}^V$ be a vector. Let $\Omega = \{-1, +1\}^V$ be the set of all spin configurations. The Gibbs distribution over Ω is defined by*

$$\forall \sigma \in \Omega, \quad \pi(\sigma) = \frac{1}{Z} \exp\left(\frac{1}{2} \sum_{v,u \in V} J_{uv} \sigma_v \sigma_u + \sum_{v \in V} h_v \sigma_v\right),$$

where $Z = \sum_{\sigma \in \Omega} \exp(\frac{1}{2} \sum_{v,u \in V} J_{uv} \sigma_v \sigma_u + \sum_{v \in V} h_v \sigma_v)$.

We consider the following problem of sampling from high-dimensional distributions.

- **Input**: the *description* of π, where the description has size $\mathrm{poly}(n)$ but typically $|\Omega| = e^{\Omega(n)}$.
- **Output**: a (possibly approximate) random sample X from π.

For example, the uniform distribution of graph coloring can be described by the graph $G = (V, E)$ and an integer q. However, if $q \geq (1+\delta)\Delta$, where Δ is the maximum degree of G and $\delta > 0$ is a constant, then the number of proper colorings is at least $(q - \Delta)^n$.

The Markov chain Monte Carlo (MCMC) method is a popular method for sampling from high-dimensional distributions. For proper graph q-colorings, the following algorithm is the well-known *Metropolis-Hastings* chain [Jer95].

- Start from an arbitrary proper coloring $X \in [q]^V$.
- For each t from 1 to T:

1. Sample a vertex $v \in V$ uniformly at random and a color $c \in [q]$ uniformly at random.
2. Define the candidate coloring $X' \in [q]^V$ by $X'_v = c$ and $X'_u = X_u$ for all $u \neq v$.
3. If X' is a proper coloring, set $X = X'$.

- Return the coloring X.

The goal of this lecture is to show that if $q > (2 + \delta)\Delta$, then the Metropolis-Hastings chain returns a good approximate sample from π if $T = O(\frac{n}{\delta} \log n)$.

2 Basic definitions for Markov chains

Let Ω be a finite set which is the state space. A Markov chain $(X_t)_{t \geq 0}$ on Ω is specified by transition matrix $P \in \mathbb{R}_{\geq 0}^{\Omega \times \Omega}$ such that

$$\mathbf{Pr}\left[X_t = x_t \mid \forall t' < t, X_{t'} = x_{t'}\right] = \mathbf{Pr}\left[X_t = x_t \mid X_{t-1} = x_{t-1}\right] = P(x_{t-1}, x_t).$$

A distribution π (viewed as a row vector) on Ω is a *stationary* of P if

$$\pi P = \pi.$$

A Markov chain is *irreducible* if for any $x, y \in \Omega$, there is a $t \geq 0$ such that $P^t(x, y) > 0$.

Lemma 2.1. *An irreducible Markov chain has a unique stationary distribution.*

Proof Sketch. To show the existence of the stationary distribution, one can explicitly construct a π satisfying $\pi = \pi P$ using the stopping time [LPW17, Sec 1.5.3]. Specifically, we can fix an arbitrary $z \in \Omega$ and construct a vector $\tilde{\pi}$ such that for any $x \in \Omega$,

$$\tilde{\pi}(x) = \mathbf{E}\left[\text{strating from } z\text{, the number of visiting } x \text{ before returning to } z\ \right].$$

For irreducible Markov chains, one can show that

$$\tau_z^+ = \mathbf{E}\left[\text{strating from } z\text{, the number of steps before returning to } z\right] < \infty.$$

A stationary distribution is then given by $\pi(x) = \tilde{\pi}(x)/\tau_z^+$. (Exercise: verify it.)

To show the uniqueness, one can check the rank of the kernel space of the matrix $P - I$ [LPW17, Sec 1.5.4]. Consider any vector h such that $h = Ph$, which means h is an eigenvector of P with eigenvalue 1. Let $x \in \Omega$ be the state such that $h(x) = \max_z h(z)$. It holds that

$$h(x) = \sum_z P(x, z) h(z).$$

Consider all z's such that $P(x, z) > 0$. Since $h(x)$ is the average of there $h(z)$'s, it must hold that $h(z) = h(x)$. Otherwise, there exists a z such that $P(x, z) > 0$ but $h(z) > h(x)$. We can repeat this argument on all z's. Since the chain is irreducible, it must hold that h is a constant function. Note that $(P - I)h = 0$. Then, $P - I$ has rank $|\Omega| - 1$. Note that π is a solution to $\pi(P - I) = 0$. The solution space has dimension 1. Therefore, there is at most one vector π such that the sum of π is 1. The above proof explicitly construct a stationary distribution. This proves the uniqueness of the stationary distribution. $\square$

A Markov chain P is reversible with respect to π if the detailed balance equation holds

$$\forall x, y \in \Omega, \quad \pi(x)P(x,y) = \pi(y)P(y,x).$$

The detailed balance equation gives a quick way to verify the stationary distribution:

$$\forall x, \quad (\pi P)(x) = \sum_y \pi(y)P(y,x) = \sum_y \pi(x)P(x,y) = \pi(x).$$

Next, we say an irreducible Markov chain is *aperiodic* if for any $x \in \Omega$, $\gcd\{t > 0 \mid P^t(x,x) > 0\} = 1$. The Markov chain convergence theorem shows that if a Markov chain P is irreducible and aperiodic, then the distribution of X_t converges to the stationary π as $T \to \infty$. To make the formal statement, we need the following definition.

Definition 2.2 (total variation distance)**.** Let μ and π be two distributions over Ω. Their total variation distance (TV-distance) is defined by

$$d_{\mathrm{TV}}(\mu, \pi) = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \pi(x)| = \max_{A\subseteq\Omega} (\mu(A) - \pi(A)). \tag{1}$$

The TV-distance is also denoted by $\|\mu - \pi\|_{\mathrm{TV}}$.

Exercise 2.3. Prove the second equality in (1).

The following theorem is proved in [LPW17, Sec 4.3].

Theorem 2.4 (convergence theorem)**.** *If a Markov chain P is irreducible, aperiodic, and reversible with respect to π, then*

$$\lim_{t\to\infty} \max_{x\in\Omega} d_{\mathrm{TV}}\left(P^t(x,\cdot), \pi\right) = 0.$$

Proof Sketch. We give the sketch of the proof in [LPW17, Sec 4.3].

- Step-1: show that for irreducible and aperiodic Markov chains, there exists $r > 0$ such that for any $x, y \in \Omega$, $P^r(x,y) > 0$ [LPW17, Proposition 1.7].

- Step-2: let Π denote the matrix such that every row vector is π. From step-1, it holds that there exists a small $0 < \theta < 1$ such that $P^r = (1-\theta)\Pi + \theta Q$, where Q is a stochastic matrix (every entry is in $[0,1]$ and every row sum is 1). Verify that $\Pi P = \Pi$ and $Q\Pi = \Pi$ and use an induction argument to show that for any $k \geq 1$, $P^{rk} = (1-\theta^k)\Pi + \theta^k Q^k$.

- Step-3: show that for any $j > 0$, $P^{rk+j} - \Pi = \theta^k(Q^k P^j - \Pi P^j) = \theta^k(Q^k P^j - \Pi)$. The reason for adding j is that rk only capture the number that is a multiple of r, but $rk + j$ captures all large numbers. Bound the total variation distance by bounding the RHS.

You can either complete the proof by yourself or read the full proof in [LPW17, Sec 4.3]. □

One can verify that for graph q-coloring, the Glauber dynamics chain is aperiodic and reversible, furthermore, it is irreducible if $q \geq \Delta + 2$. Now, the main question is how many steps one needs to simulate a Markov chain in order to draw an approximate sample from π.

Definition 2.5 (mixing time)**.** The mixing time of Markov chain P is defined by

$$T_{\mathrm{mix}}(\varepsilon) = \min\{t \mid \max_{x\in\Omega} d_{\mathrm{TV}}\left(P^t(x,\cdot), \pi\right) \leq \varepsilon\}.$$

Simulating $T_{\mathrm{mix}}(\varepsilon)$ steps is enough to generate an ε-close sample in total variation distance because the TV-distance is non-increasing due to the data processing inequality

$$d_{\mathrm{TV}}\left(P^{t+1}(x,\cdot), \pi\right) = d_{\mathrm{TV}}\left(P^t(x,\cdot)P, \pi P\right) \leq d_{\mathrm{TV}}\left(P^t(x,\cdot), \pi\right).$$

3 Coupling of Markov chains

Definition 3.1 (coupling of distributions)**.** Let μ and π be two distributions over Ω. A coupling is a joint random variable $(X, Y) \in \Omega \times \Omega$ such that $X \sim \mu$ and $Y \sim \pi$.

Let $\Omega = \{0, 1\}$. Let $\mu(0) = \frac{1}{2}$ and $\mu(1) = \frac{1}{2}$. Let $\pi(0) = \frac{1}{3}$ and $\pi(1) = \frac{2}{3}$. There are many couplings between μ and π. For example, $X \sim \mu$ and $Y \sim \pi$ can be independent; or one can first sample a real number $r \in [0, 1]$ u.a.r. and then let $X = 0$ iff $r \le \frac{1}{2}$ and $Y = 0$ iff $r \le \frac{1}{3}$.

Lemma 3.2 (coupling lemma)**.** *Let μ and π be two distributions. For any coupling (X, Y),*

$$d_{\mathrm{TV}}(\mu, \pi) \le \mathbf{Pr}[X \ne Y].$$

The equality can be achieved by the optimal coupling.

Proof. For any coupling (X, Y), it must hold that $\mathbf{Pr}[X = Y = \sigma] \le \min\{\mu(\sigma), \pi(\sigma)\}$ for all $\sigma \in \Omega$. Otherwise, the coupling is invalid. We have

$$\mathbf{Pr}[X \ne Y] = 1 - \sum_{\sigma \in \Omega} \mathbf{Pr}[X = Y = \sigma] = 1 - \sum_{\sigma \in \Omega} \min\{\mu(\sigma), \pi(\sigma)\} = d_{\mathrm{TV}}(\mu, \pi).$$

To verify the last equation, we can write

$$\begin{aligned} d_{\mathrm{TV}}(\mu, \pi) = 1 - \min\{\mu(\sigma), \pi(\sigma)\} &= \sum_{\sigma \in \Omega} (\mu(\sigma) - \min\{\mu(\sigma), \pi(\sigma)\}) \\ &= \sum_{\mu(\sigma) > \nu(\sigma)} (\mu(\sigma) - \nu(\sigma)) \\ &= \max_{A \subseteq \Omega} (\mu(A) - \nu(A)). \end{aligned}$$

Exercise 3.3. Construct a coupling such that $\mathbf{Pr}[X = Y = \sigma] = \min\{\mu(\sigma), \pi(\sigma)\}$. □

Definition 3.4 (coupling of Markov chains)**.** Let μ_0, μ_1 be two distributions over Ω. Let $(X_t)_{t \ge 1}$ be a Markov chain with transition matrix P and $X_1 \sim \pi_0$. Let $(Y_t)_{t \ge 1}$ be a Markov chain with transition matrix P and $Y_1 \sim \pi_1$. A coupling of Markov chains is a joint process $(X_t, Y_t)_{t \ge 0}$ such that $(X_t)_{t \ge 0}$ and $(Y_t)_{t \ge 0}$ both follow their correct marginal distributions.

The above definition considers two Markov chains with the *same* transition matrix but start from two different initial distributions. In many applications, we often couple (X_t, Y_t) step by step. This kind of coupling is called the Markovian coupling. Due to the Markovian property, we often assume that in a coupling, once $X_t = Y_t$, then $X_{t'} = Y_{t'}$ for all $t' > t$. Suppose μ_0 is a Dirac distribution such that $\mu_0(x) = 1$ and $\mu_1 = \pi$ is the stationary distribution. We have

$$d_{\mathrm{TV}}\left(P^t(x, \cdot), \pi\right) \le \mathbf{Pr}[X_t \ne Y_t].$$

This is because $X \sim P^t(x, \cdot)$ and $Y_t \sim \pi$ as $Y_0 \sim \pi$ and $\pi P = \pi$. Hence, (X_t, Y_t) forms a coupling of the distributions $(P^t(x, \cdot), \pi)$. The above inequality follows from the coupling lemma.

Theorem 3.5 (geometric decay)**.** *Let $\tau = T_{mix}(\frac{1}{4e})$. For any $\varepsilon > 0$,*

$$T_{mix}(\varepsilon) \le O\left(\tau \log \frac{1}{\varepsilon}\right).$$

Proof. By the definition of τ, using triangle inequality of TV-distance, for any $x, y \in \Omega$, we have

$$d_{\mathrm{TV}}\left(P^t(x,\cdot), P^t(y,\cdot)\right) \le d_{\mathrm{TV}}\left(P^t(x,\cdot), \pi\right) + d_{\mathrm{TV}}\left(\pi, P^t(y,\cdot)\right) \le \frac{1}{2e}.$$

By coupling lemma, given $X_0 = x$ and $Y_0 = y$, we can couple X_τ and Y_τ such that $\mathbf{Pr}\left[X_\tau \neq Y_\tau\right] \le 1/(2e)$. If $X_\tau = Y_\tau$, we can couple two chains such that $X_t = Y_t$ for all $t > \tau$. Otherwise, we couple $X_{2\tau}$ conditional on X_τ and Y_τ. Repeating this process, we have

$$\max_{x\in\Omega} d_{\mathrm{TV}}\left(P^{k\tau}(x,\cdot), \pi\right) \le \max_{x,y\in\Omega} d_{\mathrm{TV}}\left(P^{k\tau}(x,\cdot), P^{k\tau}(y,\cdot)\right) \le \mathbf{Pr}\left[X_{k\tau} \neq Y_{k\tau}\right] \le \left(\frac{1}{2e}\right)^k,$$

which implies the bound on mixing time. □

4 Application to graph coloring

Theorem 4.1 ([Jer95]). *Let $\delta > 0$ be a constant, if $q \ge (2+\delta)\Delta$, then the mixing time of Metropolis-Hastings chain is $T_{mix}(\varepsilon) = O(\frac{n}{\delta} n \log \frac{n}{\varepsilon})$.*

Proof. Let x and y be two proper colorings. We construct a coupling of Metropolis-Hastings chains such that $X_0 = x$, $Y_0 = y$ and $\mathbf{Pr}\left[X_t \neq Y_t\right] \le \varepsilon$. Then, the mixing result follows from the coupling lemma:

$$\max_x d_{\mathrm{TV}}\left(P^t(x,\cdot), \mu\right) \le \max_{x,y} d_{\mathrm{TV}}\left(P^t(x,\cdot), P^t(y,\cdot)\right) \le \mathbf{Pr}\left[X_t \neq Y_t\right] \le \varepsilon.$$

Specifically, we show that there is a one step coupling $(X_{t-1}, Y_{t-1}) \to (X_t, Y_t)$ such that for any $x, y \in \Omega$, it holds that

$$\mathbf{E}\left[H(X_t, Y_t) \mid X_{t-1} = x \wedge Y_{t-1} = y\right] \le \left(1 - O\left(\frac{\delta}{n}\right)\right) H(x,y), \tag{2}$$

where $H(x,y) = |\{v \in V \mid x_v \neq y_v\}|$ is the Hamming distance between x and y. (2) implies

$$\mathbf{E}\left[H(X_T, Y_T)\right] \le \left(1 - O\left(\frac{\delta}{n}\right)\right)^T n \le \varepsilon.$$

By Markov inequality, $\mathbf{Pr}\left[X_T \neq Y_T\right] = \mathbf{Pr}\left[H(X_T, Y_T) \ge 1\right] < \varepsilon$.

To show (2), we first consider a special case where x, y disagree only at one vertex v_0. Say $X(v_0) = 0$ and $Y(v_0) = 1$. The coupling is defined as follows

- two chains sample the same vertex $v \in V$ u.a.r.
- if v is not a neighbor of v_0, then two chains sample the same color $c_X = c_Y \in [q]$ u.a.r.;
- if v is a neighbor of v_0, we first sample $c_X \in [q]$ u.a.r. and then set $c_Y = c_X$ if $c_X \notin \{0,1\}$ and $c_Y = 1 - c_X$ otherwise. In words, we swap the role of $\{0,1\}$ in two chains.

For any vertex $w \neq v_0$ and w is not a neighbor of v_0, it is easy to see $X_t(w) = Y_t(w)$ with probability 1.

For vertex v_0, the event $X_t(v_0) = Y_t(v_0)$ happens if v_0 is picked and the color $c_X = c_Y$ does not appear in the neighborhood of v_0. The probability of this event is at least

$$\mathbf{Pr}\left[X_t(v_0) = Y_t(v_0)\right] \ge \frac{1}{n} \cdot \frac{q - \Delta}{q}.$$

For vertex u is a neighbor of v_0. The event $X_t(u) \neq Y_t(u)$ happens only if u is picked, $c_X = 1$, and $c_Y = 0$. The probability of this event is at most

$$\mathbf{Pr}\left[X_t(u) \neq Y_t(u)\right] \leq \frac{1}{n} \cdot \frac{1}{q}.$$

Putting everything together, we have

$$\begin{aligned}
\mathbf{E}\left[H(X_t, Y_t) \mid X_{t-1} = x \wedge Y_{t-1} = y\right] &= 1 - \frac{1}{n} \cdot \frac{q - \Delta}{q} + \Delta \cdot \frac{1}{n} \cdot \frac{1}{q} \\
&\leq 1 - \frac{1}{n}\left(1 - \frac{2\Delta}{q}\right) \\
&\leq 1 - O\left(\frac{\delta}{n}\right).
\end{aligned}$$

What if x and y differ at many vertices. We can apply the *path coupling* technique [BD97]. Say $H(x, y) = k$. We can construct a sequence of colorings $\sigma_0, \sigma_1, \ldots, \sigma_k$ such that $\sigma_0 = x$ and $\sigma_k = y$ and σ_i and σ_{i+1} differ at exactly one vertex. Then, for each pair of σ_i, σ_{i+1}, we can construct a coupling $\mathcal{C}_i$ such that conditional on $X_{t-1} = \sigma_i$ and $Y_{t-1} = \sigma_{i+1}$, $\mathcal{C}_i$ generates a random pair $(\sigma'_i, \sigma'_{i+1})$ such that $\mathbf{E}\left[H(\sigma'_i, \sigma'_{i+1})\right]$ is at most $1 - O\left(\frac{\delta}{n}\right)$. We first use $\mathcal{C}_0$ to generate (σ'_0, σ'_1). Next, $\mathcal{C}_1$ defines a joint distribution of (σ'_1, σ'_2), we condition on the value of σ'_1 to generate σ'_2. This step is valid, because the marginal distribution of σ'_1 in $\mathcal{C}_0$ and $\mathcal{C}_1$ are identical. Repeating this process, we can generate $(\sigma'_1, \sigma'_2), \ldots, (\sigma'_{k-1}, \sigma'_k)$, which gives us a random sequence $\sigma'_0, \sigma'_1, \ldots, \sigma'_k$. The random pair σ'_0, σ'_k forms a one step coupling from x and y. We have

$$\begin{aligned}
\mathbf{E}\left[H(X_t, Y_t) \mid X_{t-1} = x \wedge Y_{t-1} = y\right] &= \mathbf{E}\left[H(\sigma'_0, \sigma'_k)\right] \\
\text{(triangle-inequality)} \qquad &\leq \mathbf{E}\left[\sum_{i=0}^{k-1} H(\sigma'_i, \sigma'_{i+1})\right] \\
\text{(linearity of expectation)} \qquad &\leq \sum_{i=0}^{k-1} \mathbf{E}\left[H(\sigma'_i, \sigma'_{i+1})\right] \\
&\leq \left(1 - O\left(\frac{\delta}{n}\right)\right) H(x, y).
\end{aligned}$$

There is still one missing part in the above proof. We split x, y into a path $\sigma_0, \sigma_1, \ldots, \sigma_k$, where σ_i in the middle can be an infeasible coloring. However, the above Metropolis chain is only defined over proper colorings, which make the random coloring σ'_i undefined. This issue can be fixed by consider the following more general Metropolis-Hastings chain.

- Start from an arbitrary ~~proper~~ coloring $X \in [q]^V$.
- For each t from 1 to T:
  1. Sample a vertex $v \in V$ uniformly at random and a color $c \in [q]$ uniformly at random.
  2. Define the candidate coloring $X' \in [q]^V$ by $X'_v = c$ and $X'_u = X_u$ for all $u \neq v$.
  3. ~~If X' is a proper coloring~~, If the coloring X' is locally feasible at vertex v, i.e., for any neighbor w of v, $X'_w \neq X'_v$, then set $X = X'$.
- Return the coloring X.

This Markov chain is defined over all colorings $[q]^V$ including infeasible ones. If we further restrict the chain to proper colorings, then the chain is equivalent to the Metropolis-Hastings chain defined in the beginning of this lecture. $\square$

There are many advanced couplings to analyze the mixing time of Markov chains for graph q-colorings. See the survey [FV07] by Frieze and Vigoda for more details. So far, the best known algorithm for sampling graph q-colorings is the flipping chain, which mixes in time $O(n \log n)$ when $q \geq 1.809\Delta$ [CV24].

References

[BD97] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *FOCS*, pages 223–231, 1997.

[CV24] Charlie Carlson and Eric Vigoda. Flip dynamics for sampling colorings: Improving (11/6-ε) using a simple metric. *CoRR*, abs/2407.04870, 2024.

[FV07] Alan M. Frieze and Eric Vigoda. A survey on the use of markov chains to randomly sample colourings. *Oxford Lecture Series in Mathematics and its Applications*, 34:53, 2007.

[Jer95] Mark Jerrum. A very simple algorithm for estimating the number of k-colorings of a low-degree graph. *Random Struct. Algorithms*, 7(2):157–165, 1995.

[LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2017.