Markov Chain Mixing Time

Algorithm design

- The Fundamental Theorem of Markov Chains states that for an ergodic chain we have a unique stationary distribution π such that for all starting μ we have $\mu P^t \to \pi$ as $t \to \infty$.
- We design Markov chain algorithms by taking an “interesting” distribution π over some Ω and crafting some ergodic chain P with $\pi P = \pi$.
- But we care about running time, how long do we have to wait until μP^t is “close” to π?
- What does **close** mean?

Total variation distance

- We need a distance measure for distributions.
- **Total Variation (TV) Distance:**

$$\| \mu - \pi \|_{TV} = \frac{1}{2} \sum_{i} |\mu_i - \pi_i| = \max_{A \subset \Omega} |\mu(A) - \pi(A)|$$

- **Mixing Time:** The time it takes for the TV distance from the worst-case starting state to drop below ϵ:

$$\tau(\epsilon) = \max_{x \in \Omega} \min \left\{ t \geq 0 : \| P^t(x, \cdot) - \pi \|_{TV} \leq \epsilon \right\}$$

Classic application: card shuffling

- One of the first applications of this math was to analyze card shuffling.
- The target distribution is uniform over the permutations of the deck.
- $|\Omega| = 52!$, which is **HUGE**.
- Define a “step” to be a standard riffle shuffle: divide the deck in half (roughly) and drop cards at random from each half, somewhat evenly.
- Aldous and Diaconis proved a **cutoff phenomenon**: the TV distance is close to 1 for ~6 steps but after the 7th it falls close to 0.

CS application: lazy random walks

- Standard random walks on bipartite graphs (like even cycles) are periodic!
- The lazy walk fixes this by staying put with probability 1/2 at every state.
- Recall that for a connected d-regular graph this means the walk is ergodic, reversible, and has uniform stationary distribution.
- We will now understand the **rate** of convergence.

Spectral gap controls mixing

- For a reversible, ergodic Markov chain, its eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are real and there is a unique stationary distribution π.
- Let $\lambda_P = \max\{|\lambda_2|, |\lambda_n|\}$.
- The **spectral gap** is defined $\gamma_P = 1 - \lambda_P$.
- **Theorem:** For any starting distribution μ:

$$\| \mu P^t - \pi \|_{TV} \leq \frac{\lambda_P^t}{\min_i \pi_i} \leq \frac{e^{-\gamma_P t}}{\min_i \pi_i}$$

- For this to be at most ϵ we need $t = \frac{1}{\gamma_P} \log \frac{1}{\epsilon \min_i \pi_i}$.

Intuition

- If you know about diagonalizing matrices then you can see why the spectral gap controls mixing. In a basis such that P is diagonal, we have

$$P^t = \begin{pmatrix} \lambda_1^t & 0 & \dots & 0 \\ 0 & \lambda_2^t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n^t \end{pmatrix}$$

- For $i > 1$ we have $|\lambda_i^t| \le \lambda_P^t \le e^{-t\gamma_P}$, so larger γ_P means faster decay of the influence of these eigenvalues.
- π is the unique probability vector with eigenvalue 1, so $\mu P^t \to \pi$ with exponential rate determined by γ_P.

Analyzing st-Connectivity

- If t is not in the same component as s then we will never find it, and always correctly return False.
- If t is in the same component as s then we run the random walk enough times that we find t with probability at least $1/4$.
- If you want better success probability, run the whole thing more times.

Mixing time analysis

- Standard spectral graph theory shows that for a connected d-regular graph with n vertices, the lazy random walk has spectral gap $1/(2dn^2)$.
- Also, $\pi_i = 1/n$ for each i.
- Mixing time to reach ϵ TV distance:

$$t \geq \frac{1}{\gamma_P} \log \frac{1}{\epsilon \cdot 1/n} = 2dn^2 \log(n/\epsilon)$$

- If we take $\epsilon = 1/(4n)$ and $T = 2dn^2 \log(4n^2)$ then after T steps the probability we are at a vertex t is at least $3/(4n)$
- Repeat this $4n \log(4)/3$ times to get a failure probability at most $1/4$.

Spectral graph theory

Theorem: (See e.g. [Boll98, Section 9.4].) For a connected d-regular graph on n vertices, let the adjacency matrix A have eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. Then

- $\mu_1 = d$,
- $\mu_2 \leq d - 1/n^2$,
- $\mu_n \geq -d$.

Since $P = \frac{1}{2}I + \frac{1}{2d}A$ the eigenvalues of P are $\lambda_i = \frac{1}{2} + \frac{\mu_i}{2d}$ and hence

- $\lambda_2 \leq \frac{1}{2} + \frac{1}{2} - \frac{1}{2dn^2} = 1 - \frac{1}{2dn^2}$,
- $\lambda_n \geq \frac{1}{2} - \frac{1}{2} = 0$,
- $\lambda_P \leq 1 - \frac{1}{2dn^2}$ which means $\gamma_P \geq \frac{1}{2dn^2}$.

Application: PageRank

PageRank

- Goal: Rank the importance of pages on the World Wide Web.
- Important pages are pointed to by other important pages.
- **The Random Surfer Model:**
  - With probability θ: click a uniform random link on the current page.
  - With probability $1 - \theta$: teleport to a uniform random page on the entire web.
- Teleportation ensures the random surfer is irreducible and aperiodic.
- There are entire books on the subject [LaMe11].

Computing PageRank (The Google Matrix)

- If H is the adjacency matrix and D is the out-degree matrix, the transition matrix is called the **Google Matrix**:

$$G = (1 - \theta)\frac{\mathbf{1}^T\mathbf{1}}{n} + \theta D^{-1}H$$

- The **PageRank** is exactly the unique stationary distribution π of G.
- We can compute it by finding the left eigenvector, or simulating the chain for enough steps.
- Convergence rate is governed by the spectral gap of G, largely controlled by the damping parameter θ.

The main idea

- The stationary probabilities of the random surfer constitute a ranking system for importance
- Clever tricks mean that you can avoid computing $\pi G = \pi$ and estimate the top ranks without ever storing G in memory.
- This idea has utility in smaller networks such as recommending movies based on who watches what.
- I believe that this technology is all being replaced with deep neural networks now…

References

[Boll98] Bollobás, B. (1998). *Modern graph theory*. Graduate texts in mathematics, Springer. Catalog link.

[LaMe11] Langville, A. N., Meyer, C. D. (2011). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Google's PageRank and Beyond, Princeton University Press. DOI: https://doi.org/10.1515/9781400830329. Catalog link. Library link.