Markov Chain Convergence

Final Project for 18.408: Algorithmic Aspects of Machine Learning

Lucas E. Morales `{lucasem}`

1 Overview of Markov Chains

A Markov chain is defined by a state space $\mathcal{X}$ and a transition matrix P which satisfy the **Markov property**: for all $x_0, x_1, \ldots, x_t \in \mathcal{X}$ and $t \geq 1$,

$$\begin{aligned}&\mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= \mathbb{P}[X_t = x_t | X_{t-1} = x_{t-1}] = P(x_{t-1}, x_t)\end{aligned}$$

where $P(x_{t-1}, x_t)$ is an entry in the transition matrix. The Markov property means that the conditional probability on a state over previous states only depends on the immediately preceding state.

It follows that P is **stochastic**, that is each x-th row $P(x, \cdot)$ is a probability distribution. Suppose we start with a distribution μ_0 on our state space. After one stochastic iteration, or step of the Markov chain, our new distribution is $\mu_1 = \mu_0 P$. By induction it can be shown that the distribution at iteration $t \geq 1$ is $\mu_t = \mu_0 P^t$.

A Markov chain is **irreducible** if it is possible to get to any state from any other state. That is, for all $x, x' \in \mathcal{X}$, there exists a $t \geq 1$ such that $P^t(x, x') > 0$.

A state $x \in \mathcal{X}$ has period k if every return from x to x occurs in a multiple of k iterations. The **period** of a state x is defined as the largest such k:

$$k_x = \gcd\{t \geq 1 : P^t(x, x) > 0\}$$

If every state has period 1, then the Markov chain is **aperiodic**, else it is **periodic**.

Lemma 1. *If a Markov chain is irreducible and aperiodic, there exists some m such that for all $m' \geq m$, $P^{m'}(x, x') > 0$ for all $x, x' \in \mathcal{X}$.*

Proof. Because the Markov chain is irreducible, for every $x, x' \in \mathcal{X}$ there exists an $r = r(x, x')$ where $P^r(x, x') > 0$. By aperiodicity, for every $x \in \mathcal{X}$ there exists a $t = t(x)$ where $P^t(x, x) > 0$. Because the product of two entry-wise positive matrices is also entry-wise positive, we take $m = \max_x (t(x) + \max_{x'} r(x, x'))$. □

A **stationary distribution** π of a Markov chain is a probability distribution over the state space that satisfies $\pi = P\pi$.

2 Markov Chains in Finite State Spaces

Theorem 1 (Perron-Frobenius Theorem)**.** *Let $A \in \mathbb{R}^{n\times n}$ be entry-wise positive with eigenvalues*

$$|\lambda_1| \geq |\lambda_2| \geq \cdots |\lambda_n| \geq 0$$

and corresponding eigenvectors s_i. Then $|\lambda_1| > |\lambda_2|$ and s_1 is component-wise positive.

Proposition 1. *Every irreducible and aperiodic Markov chain on finite state space has a unique stationary distribution.*

Proof. Because the state space is finite, let n be the cardinality of the state space so that $P \in \mathbb{R}^{n\times n}$. By Lemma 1, there exists some m such that P^m is entry-wise positive. By Theorem 1 on $A = (P^m)^\top$ we have a unique maximum eigenvalue λ_1 and a corresponding component-wise positive eigenvector $\pi^\top$. We claim that $\lambda_1 = 1$ and that π is the unique stationary distribution.

Let μ_0 be a probability distribution on $\mathcal{X}$ and consider the limit as $t = mk \to \infty$. If $\lambda_1 < 1$ then $|\mu_0 P^t| = |A^k \mu_0^\top| \to 0$, and similarly if $\lambda_1 > 1$ then $|\mu_0 P^t| = |A^k \mu_0^\top| \to \infty$. However, because P is a stochastic matrix, we know exactly that $|P^t \mu_0| = 1$, therefore it must be the case that $\lambda_1 = 1$. By the definition of eigenvector, $A\pi^\top = \pi^\top$ so $\pi = \pi P^m$. Because every other eigenvalue is smaller than 1, π is the *unique* vector for which $\pi = \pi P^m$.

Again by Lemma 1, P^{m+d} is entry-wise positive for non-negative integer d — meaning that once the chain transitions becomes positive after m iterations, they stay positive. Consider the case $d = 1$ and we take our approach again, yielding the *unique* $\pi' = \pi' P^{m+1}$. This means $\pi' = \pi' \left(P^{m+1}\right)^m = \pi' P^{m(m+1)}$, but we similarly have $\pi = \pi \left(P^m\right)^{m+1} = \pi P^{m(m+1)}$. Because both π and π' are the *unique* such vectors, they must be the same vector. Therefore $\pi = \pi P^{m+1} = \left(\pi P^m\right) P = \pi P$ and we complete the proof. □

We will assume finite state spaces for the rest of this paper.

3 Convergence and Mixing Time

We define a metric that will be useful for measuring the distance between distributions, prove that certain Markov chains converge to the stationary distribution, and examine how quickly such a chain converges.

The **total variation distance** between two probability distributions μ and ν on $\mathcal{X}$ is the largest difference between the probabilities they each assign the same event:

$$\|\mu - \nu\|_{\mathrm{TV}} = \max_{A\subseteq\mathcal{X}} |\mu(A) - \nu(A)| = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|$$

Theorem 2 (Power iteration)**.** *Let A be a matrix with a unique largest eigenvalue $|\lambda_1| > |\lambda_2|$ that has eigenvector v_1, and let b_0 be a vector that is not orthogonal to v_1. The power iteration described by*

$$b_k = \frac{A^k b_0}{\|A^k b_0\|}$$

converges geometrically to a multiple of v_1:

$$|b_k - v_1| \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k$$

for some $C > 0$.

Theorem 3 (Convergence Theorem)**.** *Given an irreducible and aperiodic Markov chain with transition matrix P and stationary distribution π, there exists an $\alpha \in (0, 1)$ and $C > 0$ such that*

$$\max_{x \in \mathcal{X}} \|P^t(x, \cdot) - \pi\|_{\text{TV}} \leq C\alpha^t$$

Proof. We will prove this statement for finite state spaces, though it has been proven for infinite spaces as well. By Proposition 1, we showed that therefore P has unique largest eigenvalue of 1 with corresponding left eigenvector π. Because P is stochastic, $P(x, \cdot)$ is a probability distribution whose elements sum to 1 and is therefore *not* orthogonal to π for all $x \in \mathcal{X}$. Let $\alpha = |\frac{1}{\lambda_2}| \in (0, 1)$ then by Theorem 2 we have that for all $x \in \mathcal{X}$:

$$|P^t(x, \cdot) - \pi| \leq C\alpha^t$$

from which the theorem follows. □

From this statement of the convergence theorem comes the study of *rapid mixing*, where a P^* is found which minimizes the second largest eigenvalue without changing the stationary distribution — and hence getting close to the stationary distribution in fewer iterations. We will revisit this in Section 4.

We define the maximal distance from the stationary distribution after t iterations:

$$d(t) = \max_{x \in \mathcal{X}} \|P^t(x, \cdot) - \pi\|_{\text{TV}}$$

which will be useful for establishing a convergence bound over a number of iterations on the chain. We more specifically define the maximal distance between any two distributions on our state space after some number of iterations:

$$\bar{d}(t) = \max_{x, x' \in \mathcal{X}} \|P^t(x, \cdot) - P^t(y, \cdot)\|_{\text{TV}}$$

Proposition 2. *For positive integer t,*

$$d(t) \leq \bar{d}(t) \leq 2d(t)$$

and for positive integer c,

$$\bar{d}(ct) \leq \bar{d}(t)^c$$

Proof sketch. The first statement can be proven by considering the distance between $P^t(x, \cdot)$ and π evaluated at a particular event and relating it to the distance between any $P^t(x, \cdot)$ and $P^t(x', \cdot)$ using the triangle inequality. The second statement holds because $\bar{d}$ is submultiplicative — i.e. $\bar{d}(a+b) \leq \bar{d}(a)\bar{d}(b)$ — which can be proven using an *optimal coupling*, a topic we will not discuss here. □

The **mixing time** is how we measure rate of convergence:

$$t_{\text{mix}}(\epsilon) = \min\{t : d(t) < \epsilon\}$$

By Proposition 2, for positive integer ℓ, we have

$$d(\ell t_{\text{mix}}(\epsilon)) \leq \bar{d}(t_{\text{mix}}(\epsilon))^\ell \leq (2d(t_{\text{mix}}(\epsilon)))^\ell \leq (2\epsilon)^\ell$$

Let $t^* = t_{\text{mix}}(1/4)$ be the mixing time to get the total variation distance to the stationary distribution within 1/4. We then have that

$$d(\ell t^*) \leq 2^{-\ell}$$

and that

$$t_{\text{mix}}(\epsilon) \leq \left\lceil \log_2 \frac{1}{\epsilon} \right\rceil t^*$$

4 Rapid Mixing on Undirected Graphs using Semidefinite Programming

We observed in the proof of Theorem 3 that convergence is exponential in $|\frac{1}{\lambda}_2|$. It follows that for small λ_2, mixing is fast. Recall that, as in the convention of Theorem 1, λ_2 is the second-largest eigenvalue by magnitude (or second-largest eigenvalue modulus, SLEM). The **rapid mixing** problem is to find some stochastic P^* that minimizes the mixing time without changing the stationary distribution π:

$$\begin{aligned} &\text{minimize } |\lambda_2(P)| \\ &\text{subject to } P \geq 0,\ P\mathbf{1} = \mathbf{1} \\ &\qquad\qquad s(P) = \pi,\ f(P) \end{aligned}$$

where $s(P)$ is the stationary distribution under P, and $f(P)$ are additional constraints which depend on the particular problem.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected undirected graph with vertices $\mathcal{V} = \{1, \ldots, n\}$ and edges $\mathcal{E}$ where $(i,j) \in \mathcal{E} \Leftrightarrow (j,i) \in \mathcal{E}$. We define a Markov chain on state space $\mathcal{V}$ with symmetric transition matrix $P \in \mathbb{R}^{n \times n}$ where $P_{ij} = 0$ if $i \neq j$ and $(i,j) \notin \mathcal{E}$.

Proposition 3. *If a Markov chain has symmetric transition matrix P, then its stationary distribution $\pi = \frac{1}{n}\mathbf{1}$ is uniform.*

Proof. Because P is symmetric and stochastic,

$$(\mathbf{1}P)_j = \sum_{i=1}^{n} P_{ij} = 1$$

Therefore $\mathbf{1}P = \mathbf{1}$ from which it follows, by normalization, that $\pi = \frac{1}{n}\mathbf{1}$. □

With Proposition 3 we can simplify the rapid mixing problem for $\mathcal{G}$:

$$\begin{aligned} \text{minimize } & |\lambda_2(P)| \\ \text{subject to } & P \geq 0,\ P\mathbf{1} = \mathbf{1},\ P = P^\top \\ & P_{ij} = 0,\ (i,j) \notin \mathcal{E},\ i \neq j \end{aligned}$$

Theorem 4. *The rapid mixing problem for $\mathcal{G}$ can be solved by a semidefinite program.*

Proof. We first reformulate λ_2 by taking the spectral norm of the orthogonal projection of P against $\mathbf{1}$:

$$\lambda_2(P) = \|P - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2$$

We then bound this norm by a scaled identity matrix:

$$\begin{aligned} \text{minimize } & s \\ \text{subject to } & -sI \preceq P - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \preceq sI \\ & P \geq 0,\ P\mathbf{1} = \mathbf{1},\ P = P^\top \\ & P_{ij} = 0,\ (i,j) \notin \mathcal{E},\ i \neq j \end{aligned}$$

□

With a semidefinite program (SDP) in hand, the dual can be used to determine optimality conditions. These conditions are developed in [Boyd et al., 2004, Section 4].

5 The Metropolis Algorithm

The transition probabilities for a **random walk** on a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where d_i is the degree of vertex i:

$$P^{\text{rw}}_{ij} = \begin{cases} 1/d_i & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases}$$

For this chain, the stationary distribution at each vertex is proportional to vertex's degree.

We can construct a Markov chain with an arbitrary stationary distribution $\pi = (\pi_1, \ldots, \pi_n)$ by modifying the random walk: let $R_{ij} = (\pi_j P^{\text{rw}}_{ji})/(\pi_i P^{\text{rw}}_{ij})$ and define the **Metropolis algorithm** as a Markov chain with the following transition probabilities:

$$M_{ij} = \begin{cases} P^{\text{rw}}_{ij} \min\{1, R_{ij}\} & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \\ 1 - \sum_{\{i,k\} \in \mathcal{E}} P^{\text{rw}}_{ik} \min\{1, R_{ik}\} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

where P^{m}_{ii} is effectively the probability of a "rejected proposal" coming from state i. More generally, the Metropolis algorithm can modify a transition matrix P to obtain a new transition matrix M which converges to a particular stationary distribution π. Let $R_{ij} = (\pi_j P_{ji})/(\pi_i P_{ij})$:

$$M_{ij} = \begin{cases} P_{ij} \min\{1, R_{ij}\} & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \\ P_{ii} + \sum_{\{i,k\} \in \mathcal{E}} P_{ik} \left(1 - \min\{1, R_{ik}\}\right) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

The Metropolis chain based on a random walk with uniform stationary distribution π is

$$M = \begin{cases} \min\{1/d_i, 1/d_j\} & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j \\ \sum_{(i,j) \in \mathcal{E}} \max\{0, 1/d_i - 1/d_k\} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

5.1 Metropolis with Independent Proposals

The Metropolis algorithm can have the base chain set to repeated independent samples. For state space $\mathcal{X} = \{1, \ldots, n\}$ (numbered without loss of generality) and a probability distribution $p(x)$, the base chain we use is described by transition matrix $P(x, x') = p(x')$. Let M be the Metropolis chain using this base chain with $R_{ij} = \frac{\pi(j)p(i)}{\pi(i)p(j)}$.

Theorem 5 ([Liu, 1996])**.** *The Metropolis chain M with base chain of repeated independent samples from distribution p and SLEM λ_2 has bounded variation distance from start state x:*

$$\|M^t(x,\cdot),\pi\|_{\mathrm{TV}} \leq \frac{\lambda_2^k}{\sqrt{\pi(x)}}$$

It follows from Theorem 5 that the mixing time for M is

$$t_{\mathrm{mix}}(\epsilon) = \left\lceil \log_{\lambda_2}\left(\epsilon\sqrt{r}\right)\right\rceil$$

where $r = \min_{x\in\mathcal{X}} \pi(x)$.

It is difficult to rigorously determine similar bounds for base chains which do not follow this repeated-independent model.

In [Boyd et al., 2004], empirical comparisons are made between mixing times of the Metropolis chain versus solutions to the SDP from Theorem 4. These experiments found the SDP to never yield a slower-mixing chain than the Metropolis algorithm. However, the constrained domain under which the SDP can be applied — in which the stationary distribution is uniform and the transitions are symmetric — is far more restrictive than the broad use-cases of the Metropolis algorithm (or, even more so, the Metropolis-Hastings algorithm, which is an extension of the Metropolis algorithm). If SDP can be used, and especially if the dual proves that its solution is optimal, then it should be applied in favor of the Metropolis-Hastings algorithm.

6 Conclusion

In this survey, we reviewed at how to measure convergence for Markov chains and analyzed at two different applications: rapid mixing and the Metropolis algorithm. Markov chains are foundational to many techniques for Bayesian inference, and establishing rigorous convergence bounds for such techniques is a desirable direction for future research in probabilistic approaches of machine learning.

References

Stephen Boyd, Persi Diaconis, and Lin Xiao. Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

Stephen P. Boyd, Persi Diaconis, Pablo A. Parrilo, and Lin Xiao. Fastest mixing markov chain on graphs with symmetries. *SIAM Journal on Optimization*, 20: 792–819, 2009.

Persi Diaconis and Laurent Saloff-Coste. What do we know about the metropolis algorithm? In *STOC*, 1995.

David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Society, 2008.

Jun S Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.