Markov chains lecture Dan D 01/08/2018, 15:19

INTRODUCTION TO FINITE MARKOV CHAINS

– DANIAL DERVOVIC

MAIN REF: Levin, Peres, Wilmer

Theory

- Def^n of M.C.
- Irreducibility + Aperiodicity
- Simple random walks on graphs
- Stationary distribution (simple walk, symmetric chain)
- Existence + Uniqueness of stationary distribution for irreducible Markov chains
- Total variation distance
- Convergence Theorem
- Mixing, bounds on mixing time

Applications

- Card shuffling
- Metropolis–Hastings
- Combinatorial Optimisation

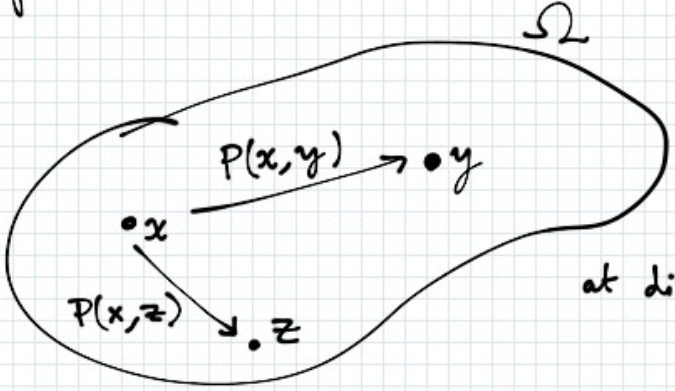
Quantum

- Discussion of mixing in quantum walks, bounds

Markov chains theory is fundamentally the study of the long-term behaviours of stochastic processes that are memoryless. That is, the systems we will study have dynamics obeying the Markov property; informally, the state tomorrow only depends on the state today.

How do we describe these dynamics formally?

State space Ω



at discrete timestep t

Markov property: transition probability $x \mapsto y$ given by a constant $P(x,y)$.

This gives us a transition matrix $P \in \mathbb{R}^{|\Omega| \times |\Omega|}$

Remark Conservation of probability demands that

$$\sum_{y \in \Omega} P(x,y) = 1, \quad P(x,y) \geq 0, \forall x \in \Omega$$

$\Rightarrow P$ is a stochastic matrix.

every row sums to 1

P is a map between probability distributions on Ω

$P: \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$. In fact, linear map from $\mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ iff stochastic matrix

The state at time t is a row vector $\mu_t \in \mathcal{P}(\Omega)$

$$\mu_t = \mu_{t-1} P \Rightarrow \mu_t = \mu_0 P^t$$

Def^n

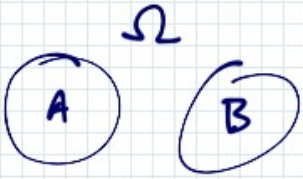
We can specify a Markov chain with the triple $\mathcal{M} = (\Omega, P, \mu_0)$, where $P: \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ and $\mu_0 \in \mathcal{P}(\Omega)$.

Practically: Sample an initial state from Ω according to μ_0. Then random walk according to P.

Irreducibility and Aperiodicity

Def^n A chain (Ω, P, μ_0) is irreducible if for any $x, y \in \Omega$, $\exists t \in \mathbb{N}$ such that $P^t(x,y) > 0$.

Example



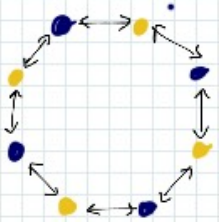
$\Omega = A \cup B$ NOT irreducible: consider $x \in A$, $y \in B$

For any $x \in \Omega$, let $T(x)$ be the set of lengths possible paths starting and ending at x.

$T(x) = \{4, 7\}$

Def^n If $\gcd(T(x_1), T(x_2), \dots) = 1$, then M is aperiodic, where gcd is "greatest common divisor".

Example consider even-length cycles C_{2k}, with equal transition probability for clockwise or anticlockwise move.



PERIODIC with period two

Note Any periodic chain can be aperiodic using a simple trick. Sub $Q = \frac{I+P}{2}$ as the transition matrix. This is called a **lazy walk**.

Prop. If P is aperiodic & irreducible, then $\exists r \in \mathbb{N}$ such that $P^r(x,y) > 0$ **for all** $x, y \in \Omega$.

Proof idea Can take max over t in irreducibility def^n + constant.

Note the "for all" is what the aperiodicity gives us. See the walk on C_{2k} above for example where irreducible + periodic means $\nexists r$.

Def^n An **ergodic** M.C. is one which is aperiodic + irreducible

Stationary Distributions

A distribution $\pi \in \mathcal{P}(\Omega)$ satisfying $\pi = \pi P$ is called a **stationary distribution** of M.

Example Let G be a graph. Define the **simple random walk** on G to be the M.C. with state space $V(G)$ and transition matrix

$$P(x,y) = \begin{cases} 1/\deg(x), & \text{if } y \sim x; \\ 0, & \text{otherwise.} \end{cases}$$

Claim $\pi(y) = \dfrac{\deg(y)}{2|E|} \quad \forall y \in \Omega$.

Proof Need to show $\pi(y) = \sum_{x \in \Omega} \pi(x) P(x,y)$

$$\sum_{x \in \Omega} \pi(x) P(x,y) = \sum_{x \in \Omega} \frac{\deg(x)}{2|E|} \delta_{x \sim y} \frac{1}{\deg(x)} = \frac{1}{2|E|} \sum_{x \sim y} 1 = \frac{\deg(y)}{2|E|}$$

Example A **symmetric** M.C. is one is an M.C. with transition matrices satisfying $P = P^T$.

Claim $\pi = \frac{1}{|\Omega|} \vec{1}^T$, i.e. is uniform for symmetric M.C.

Proof $\pi P = \pi P^T = \frac{1}{|\Omega|} \vec{1}^T P^T = \frac{1}{|\Omega|} (P \vec{1})^T = \frac{1}{|\Omega|} \vec{1}^T = \pi$

↑ def^n symmetric

↑ P stochastic

Existence & uniqueness of stationary distributions for irreducible M.C.'s

Linear algebra gives us this for free, via the Perron-Frobenius theorem.

Thm (Perron - Frobenius)
Let A be a square nonnegative irreducible matrix with spectral radius $\rho(A) = r$. Then,

1) r is an eigenvalue of A, $r > 0$.
2) Corresponding eigenvector, $\vec{x}$, satisfies $x_i > 0 \ \forall i$.
3) r has multiplicity 1.
4) $\min_i \sum_j a_{i,j} \leq r \leq \max_i \sum_{i,j} a_{i,j}$.
5) $\vec{x}$ is only eigenvector with *all* $x_i > 0$

For irreducible $\mathcal{M} = (\Omega, P, \mu_0)$, P is square, irreducible & nonnegative.

Applying P.F., $\sum_j P_{ij} = 1$ since P is stochastic.

(4) $\Rightarrow r = 1$. + (2) $\Rightarrow$ Existence of π. $\left(\pi = \frac{\vec{x}}{\|\vec{x}\|_1}\right)$

(3) gives uniqueness

Total variation distance

"The" distance measure between probability distributions.

Def" Let μ & ν be probability measures over Ω. Then, the total variation distance between μ & ν is

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)|$$

"Maximum difference assigned to an event A by μ & ν"

Lemma If Ω is finite:

$$\|\mu - \nu\|_{TV} = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| = \frac{1}{2} \|\mu - \nu\|_1$$

The Convergence Theorem

We now have all of the ingredients to prove the convergence theorem!

Thm (Convergence) Let $\mathcal{M} = (\Omega, P, \mu_0)$ be an ergodic M.C. with stat. dist. π. Then, $\exists\, \alpha \in (0,1)$, $C > 0$ such that

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

Proof Since P is ergodic, $\exists r$ such that $P^r > 0$.

Define Π as the matrix with $|\Omega|$ rows, each of which is π, i.e. $\Pi := \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}$. We can set $\delta > 0$ sufficiently small that

$$P^r(x,y) \geq \delta \pi(y) \quad \forall x,y \in \Omega .$$

Now, define $\theta := 1-\delta$. Now, implicitly define a stochastic matrix Q by

$$P^r = (1-\theta)\Pi + \theta Q$$

Why is Q stochastic? $\theta Q = P^r - (1-\theta)\Pi \Rightarrow Q = \frac{1}{\theta}P^r - \left(\frac{1-\theta}{\theta}\right)\Pi$

P^r, Π are stochastic, so each row of Q sums to $\frac{1}{\theta} + -\left(\frac{1-\theta}{\theta}\right) = 1$

Elements of Q are nonnegative. Why? suppose $Q(x,y) < 0$. Then,

$$\frac{1}{\theta}P^r(x,y) - \left(\frac{1-\theta}{\theta}\right)\pi(y) < 0 \Rightarrow P^r(x,y) - (1-\theta)\pi(y) < 0$$
$$\Rightarrow P^r(x,y) - \delta\pi(y) < 0 \quad \text{⚔ contradiction}$$

Fact 1 $M\Pi = \Pi$ for any stochastic M.

Proof $[M\Pi](x,y) = \sum_{z\in\Omega} M(x,z)\Pi(z,y) = \sum_{z\in\Omega} M(x,z)\pi(y) = \pi(y)$ (M stochastic)

Fact 2 $\Pi M = \Pi$ for any M such that $\pi M = \pi$

Proof $[\Pi M](x,y) = \sum_{z\in\Omega} \Pi(x,z)M(z,y) = \sum_{z\in\Omega} \pi(z)M(z,y) = \pi(y)$

Claim $P^{rk} = (1-\theta^k)\Pi + \theta^k Q^k$ for $k \geq 1$

Proof Induction. For $k=1$, def^n of Q gives it for free.
Now assume hypothesis holds for $k=n$,

$$P^{r(n+1)} = P^{rn} \cdot P^r = \left[(1-\theta^n)\Pi + \theta^n Q^n\right] P^r$$
$$= (1-\theta^n)\Pi P^r + \theta^n Q^n \left((1-\theta)\Pi + \theta Q\right)$$
$$= (1-\theta^n)\Pi P^r + \theta^n(1-\theta) Q^n \Pi + \theta^{n+1} Q^{n+1}$$

Fact 2 $\Rightarrow \Pi P^r = \Pi$, Fact 1 $\Rightarrow Q^n \Pi = \Pi$, so

$$P^{r(n+1)} = \left[(1-\theta^n) + \theta^n(1-\theta)\right]\Pi + \theta^{n+1} Q^{n+1}$$
$$= (1-\theta^{n+1})\Pi + \theta^{n+1} Q^{n+1}$$

Multiply P^{rk} by P^j:

$$P^{rk+j} = (1-\theta^k)\underbrace{\Pi P^j}_{\Pi} + \theta^k Q^k P^j$$

$$\Leftrightarrow P^{rk+j} - \Pi = \theta^k (Q^k P^j - \Pi)$$

Take the 1-norm of i^{th} row of each side

LHS: $\| P^{rk+j}(i,\cdot) - \pi \|_1 = 2 \| P^{rk+j}(i,\cdot) - \pi \|_{TV}$

RHS: $\theta^k \| Q^k P^j(i,\cdot) - \pi \|_1 = 2\theta^k \| Q^k P^j(i,\cdot) - \pi \|_{TV}$

$\leq 2\theta^k \cdot 1$ (max T.V is 1)

$\Rightarrow \| P^{rk+j}(i,\cdot) - \pi \|_{TV} \leq \theta^k$

Can adjust constants C, α such that statement is true non-asymptotically ■

Mixing

Have shown that ergodic M.C.s converge in the infinite-time limit. What about finite-time convergence?

Def^n (Mixing time)

$$t_{mix}(\varepsilon) = \min_{t \in \mathbb{N}} \left\{ t \;\middle|\; \max_{x \in \Omega} \| P^t(x,\cdot) - \pi \|_{TV} \leq \varepsilon \right\}$$

For a certain large class of M.C.s, can bound mixing time w.r.t spectrum of transition matrix.

Def^n An ergodic M.C. is <u>reversible</u> if

$$\pi(x) P(x,y) = \pi(y) P(y,x)$$

Let γ be the spectral gap of P, i.e. the difference between its largest two eigenvalues (note the largest is 1 & is unique by P-F thm)

Thm Let M be an ergodic M.C. with spectral gap γ. Then,

$$\left(\frac{1}{\gamma} - 1\right) \log\left(\frac{1}{2\varepsilon}\right) \leq t_{mix}(\varepsilon) \leq \log\left(\frac{1}{\varepsilon \pi_{min}}\right) \frac{1}{\gamma}$$

Note NON asymptotic bounds

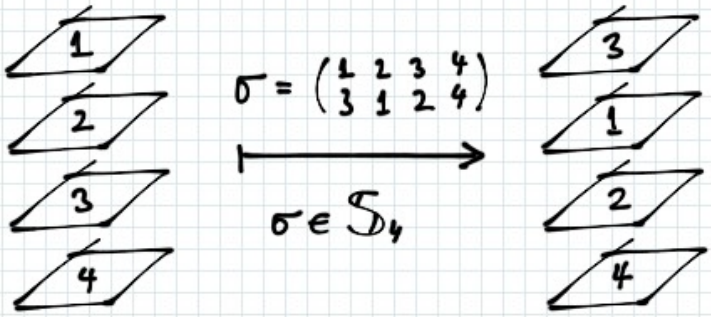
Note In Hamiltonian complexity, other way around. Can compute t_{mix} to bound γ.

APPLICATIONS

Card Shuffling

A stack of n cards can be viewed as an element of the symmetric group, S_n, consisting of all permutations of the set $\{1, 2, \ldots, n\}$.

Interpret permutations as positions of cards



Question How many riffle shuffles does it take to mix a pack of cards?

Define a M.C. on S_{52} that models a riffle shuffle, then compute $t_{mix}(\varepsilon)$ for some suitably small ε.

What is the M.C.?

- Draw $M \sim \text{Binom}(n, \frac{1}{2})$ (cut the deck)
- There are $\binom{n}{M}$ ways to riffle the two piles together, preserving the ordering of each pile. Choose one ordering uniformly at random.

These transitions faithfully model a riffle shuffle

Why $\binom{n}{M}$ riffles?

There are n positions to be filled. Call these "buckets" and the cards "balls". There are $\binom{n}{M}$ ways to put M balls in n buckets. Since the ordering of the top M cards is fixed, fill these cards in a given placement. The remaining $n-M$ cards are also ordered, so we just fill in the blank positions with them.

Thm (Bayer-Diaconis '92)

$$t_{mix}(\varepsilon) \leq 2\log_2\left(4n/3\right)$$

For $n = 52$, $\varepsilon = 0.01$, $t_{mix} \approx 12$

MARKOV CHAIN MONTE-CARLO

Metropolis-Hastings Algorithm

Given an irreducible transition matrix P, $\exists$ unique stat. dist. π s.t. $\pi P = \pi$.

How about the other way around? Given a probability distribution π can we find a transition matrix P for which π is the stationary distribution?

Suppose we have the <u>symmetric</u> chain (Ω, Ψ, μ_0). We will modify this chain so that π is its stationary distribution. (Note: we can relax the symmetric requirement, it simplies the analysis).

Modify M.C. as follows: proceed with transitions Ψ, but now "accept" a move with probability $a(x,y)$.

New chain:

$$P(x,y) = \begin{cases} \Psi(x,y)\,a(x,y), & y \neq x \\ 1 - \sum_{z : z \neq x} \Psi(x,z)\,a(x,z), & y = x \end{cases}$$

Lemma (Detailed balance)

Note M reversible iff satisfies detailed balance

Suppose $\pi \in \mathcal{P}(\Omega)$ satisfies

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

Then π is stationary for π

Proof Sum both sides over y

$$\pi(x) = \sum_{y\in\Omega} \pi(x)P(x,y) = \sum_{y\in\Omega} \pi(y)\,P(x,y)$$

↑ P stochastic

Thus we want

$$\pi(x)\,\Psi(x,y)\,a(x,y) = \pi(y)\,\Psi(y,x)\,a(y,x) \quad \forall x \neq y$$

Since Ψ is symmetric,

$$a(x,y) \leq \frac{\pi(y)}{\pi(x)}, \quad a(y,x) \leq \frac{\pi(y)}{\pi(x)}$$

Also, $a(x,y) \leq 1$ since it's a probability

To improve mixing time, we want $a(x,y)$ as large as possible.

Try $a(x,y) = \min\{1, \frac{\pi(y)}{\pi(x)}\}$ This satisfies detailed balance!

So we have derived M.C., with stat. dist. π, requiring transition probs.

$$P(x,y) = \begin{cases} \Psi(x,y)\min\{1, \frac{\pi(y)}{\pi(x)}\}, & y \neq x \ ; \\ 1 - \sum_{z : z \neq x} \Psi(x,z)\min\{1, \frac{\pi(z)}{\pi(x)}\}, & y = x \ . \end{cases}$$

MCMC & combinatorial optimisation

Let $f : \Omega \to \mathbb{R}$. We want to solve

$$\max_{x \in \Omega} f(x)$$

Exhaustive search is slow, can we M.C. Which one?

Metropolis with designed stationary distribution.

Fix $\lambda \geq 1$ & define

$$\pi_\lambda(x) = \frac{\lambda^{f(x)}}{Z(\lambda)}, \quad \text{where } Z(\lambda) = \sum_{x \in \Omega} \lambda^{f(x)}$$

Note: we don't need $Z(\lambda)$ explicitly

Define

$$\Omega^* = \left\{ x \in \Omega \;\middle|\; f(x) = f^* := \max_{x \in \Omega} f(x) \right\}$$

Then,

$$\lim_{\lambda \to \infty} \pi_\lambda(x) = \lim_{\lambda \to \infty} \frac{\lambda^{f(x)} / \lambda^{f^*}}{|\Omega^*| + \sum_{x \in \Omega \setminus \Omega^*} \lambda^{f(x)} / \lambda^{f^*}} = \frac{1}{|\Omega^*|} \delta_{\{x \in \Omega^*\}}$$

multiply by $1/\lambda^{f^*}$ / $1/\lambda^{f^*}$

This is uniform dist. over optimal set!

QUANTUM WALKS

c.f. talk on Szegedy walks for general defn.

Certain graphs, e.g. cycle have quadratic speedup in terms of mixing time, as compared with optimal classical walk.

Generically, $O\left(\frac{1}{\gamma}\right)$ bound becomes $O\left(\sum_{\lambda_i \neq \lambda_j} \frac{1}{|\lambda_i - \lambda_j|}\right)$, where $\{\lambda_i\}$ are the eigenvalues of the (unitary) walk matrix.