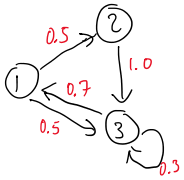
3 Markov chains

Sunday, January 19, 2020 4:11 PM

Looked ahead, and decided this is the best time to do Markov chains, since you need them for some other optimization & clustering problems. Also, Mystery Hunt.



Consider a directed graph $G=(V, E)$, $E \subseteq V^2$.
A graph is **strongly connected** if for any $x, y \in V$, there exists a path $P=\{v_0, \dots, v_k\}$, $v_i \in V$, $(v_i, v_{i+1}) \in E$, $v_0 = x$ (starting at x), $v_k = y$ (ending at y).

A random walk on a graph is a sequence of vertices generated from a start vertex, where at each step you probabilistically select the next vertex by traveling along an edge.

Ex. Start at node 1, then the next state is 2 w.p. $\frac{1}{2}$, 3 w.p. $\frac{1}{2}$.
After that prob(3) = 0.65, prob(1) = 0.35. And so on.

Let the matrix P have p_{ij} = Prob. (transition from i to j). Then $\vec{p}(t+1) = \vec{p}(t)\,P$.

Ex.
$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0.7 & 0 & 0.3 \end{bmatrix}$$
$$[1\ 0\ 0]\,P = [0\ \ 0.5\ \ 0.5]$$
$$[1\ 0\ 0]\,P^2 = [0.35\ \ 0\ \ 0.65]$$

We are going to study the limiting behavior of random walks, as well as mixing time, hitting time, etc.

Random walk on graph $\Leftrightarrow$ Markov chain
vertices $\Leftrightarrow$ states
strongly connected graph $\Leftrightarrow$ connected Markov chain

---

Let $\vec{p}(t)$ be the prob. distribution after t steps of a random walk.

Def. The long-term avg prob. dist $\vec{a}(t)$ is
$$\vec{a}(t) = \frac{1}{t}\left(\vec{p}(0) + \vec{p}(1) + \cdots + \vec{p}(t-1)\right).$$

Goal: $\lim_{t\to\infty} \vec{a}(t) = \vec{x}$ s.t. $\vec{x}P = \vec{x}$ for a connected Markov chain.

Technical lemma

Lemma 4.1: Let P be the transition matrix for a connected Markov chain.
The $n \times (n+1)$ matrix $A = [P-I, \vec{1}_n^T]$ has rank n. $\left(\vec{1}_n^T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\} n\right)$

proof. Suppose $\text{rank}(A) \neq n$. Then $\text{rank}(A) < n \Rightarrow \dim(\text{Null}(A)) \geq 2$.
Each row in P sums to 1, so each row in $P-I$ sums to 0.
Then $A \begin{bmatrix} \vec{1}_n^T \\ 0 \end{bmatrix} = (P-I)\vec{1}_n^T = 0$.

Assume $\exists [\vec{x}, \alpha] \perp [\vec{1}_n, 0]$ s.t. $A\vec{x}^T = 0$. (second solution)

Then $(P-I)\vec{x}^T + \alpha 1_n = 0$ so for each i, $x_i = \sum_{j=1}^{n} p_{ij} x_j + \alpha$. (Each x_i a convex combination of x_j's $+ \alpha$)

Assume $\exists [\vec{x}, \alpha] \perp [1_n, 0]$ s.t. $A\vec{x} = 0$.

Then $(P - I)\vec{x}^T + \alpha 1_n = 0$ so for each i, $x_i = \sum_{j=1}^{n} p_{ij} x_j + \alpha$. (Each x_i a convex combination of x_j's $+ \alpha$)

Since $\vec{x} \perp \vec{1}_n$, if $\vec{x} \neq 0$, then some $x_i > 0$ and some $x_j < 0$.

Let $x_i \geq x_j$ for all j. Then $\sum_{j=1}^{n} p_{ij} x_j < \sum_{j=1}^{n} p_{ij} x_i = 1$.

But $x_i = \sum_{j=1}^{n} p_{ij} x_j + \alpha$, so $\alpha > 0$.

Alternately, let $x_i \leq x_j$ for all j. Then $\alpha < 0$. ✗

Contradiction, so $\text{rank}(A) = n$.

Fundamental thm of Markov Chains: For a connected Markov chain, there is a unique probability vector $\vec{\pi}$ satisfying $\vec{\pi} P = \vec{\pi}$. Moreover, for any starting dist $\vec{p}(0)$, $\lim_{t \to \infty} \vec{a}(t) = \vec{\pi}$

proof. $\vec{a}(t) = \frac{1}{t}\left(\vec{p}(0) + \cdots + \vec{p}(t)\right)$, so $\vec{a}(t)$ is also a prob. dist.

Let $\vec{b}(t) = \vec{a}(t)P - \vec{a}(t) = \frac{1}{t}\left(p(t) - p(0)\right)$.

Then $\left|\vec{a}(t)P - \vec{a}(t)\right| = \frac{1}{t}\left|p(t) - p(0)\right| \leq \frac{1}{t}\left(|p(t)| + |p(0)|\right) = \frac{2}{t} \to 0$ as $t \to \infty$.

Thus, the limit exists.

By Lemma, let $A = [P - I, 1_n^T]$. $\text{rank}(A) = n$.

Let B be an $n \times n$ submatrix of A with all but the first column.

$\text{rank}(B) = n$ because the first column is negative the sum of the other columns (besides the last)

Thus, B is invertible.

Let $\vec{c}(t)$ be $\vec{b}(t)$ with the first column removed.

Note $\vec{b}(t) = \vec{a}(t)[P - I]$.

So $\vec{a}(t) A = [\vec{b}(t), \vec{a}(t) 1_n^T] = [\vec{b}(t), 1]$

$\vec{a}(t) B = [\vec{c}(t), 1]$

$\vec{a}(t) = [\vec{c}(t), 1] B^{-1} \longrightarrow [0, 1] B^{-1}$ as $t \to \infty$

Thus $\vec{\pi} = [0, 1] B^{-1}$ satisfies the thm. ▨

Lemma: For a random walk on a strongly connected graph with probabilities on the edges, if the vector $\vec{\pi}$ satisfies $\pi_x p_{xy} = \pi_y p_{yx}$ and $\sum_x \pi_x = 1$, then $\vec{\pi}$ is the stationary distribution.

proof. $\pi_x = \sum_y \pi_x p_{xy} = \sum_y \pi_y p_{yx}$, so $\vec{\pi} = \vec{\pi} P$. ▨

Suppose you want to sample a complicated prob. dist. in high dimensions

Markov Chain Monte Carlo (MCMC)

Given a prob. dist. $p(\vec{x})$, want to estimate $\mathbb{E} f = \sum_{\vec{x}} p(\vec{x}) f(\vec{x})$.

Markov Chain Monte Carlo (MCMC)

Given a prob. dist. $p(\vec{x})$, want to estimate $\mathbb{E} f = \sum_{\vec{x}} p(\vec{x}) f(\vec{x})$.

If each x_i has at least 2 possibilities, then exponential number of possible $\vec{x}$ ($\sim 2^n$)

If we can instead sample points $\vec{x}$ according to p, then we don't need to evaluate at all possible $\vec{x}$, reducing computation time.

e.g. equivalent of finding mean by drawing random samples

MCMC allows drawing a sample $\vec{x}$ w.p. $p(\vec{x})$ by designing a Markov Chain whose stationary dist. is $p(\vec{x})$.

- ↳ Metropolis-Hastings
- ↳ Gibbs Sampling

} But first prove that MCMC works in general

$\mathbb{E}_{\vec{x} \sim p} f(\vec{x}) = \sum_{\vec{x}} p(\vec{x}) f(\vec{x})$. Let's use notation $\mathbb{E} f = \sum_i p_i f_i$, where i is our state.

Consider a random walk on our Markov chain. Let Y be the average of the values of f along nodes of our Markov chain in a t-step walk.

Then Y is an estimator for $\mathbb{E} f$ as $t \to \infty$.

$$\mathbb{E}_{\text{over } t\text{-step walks}} Y = \sum_i f_i \cdot \left(\frac{1}{t} \sum_{j=1}^{t} \text{Prob(walk is in state } i \text{ at time } j) \right) = \sum_i f_i \cdot a_i(t).$$

Let $f_{max} = \max_i |f_i|$.

Then $\left| \sum_i f_i p_i - \mathbb{E} Y \right| \le f_{max} \sum_i \left| p_i - a_i(t) \right| = f_{max} \underbrace{\left| \vec{p} - \vec{a}(t) \right|_1}$

ℓ_1-norm
total variation distance

So we can bound the performance of an MCMC estimate by the f_{max} and the total variation distance between $\vec{p}$ and $\vec{a}(t)$ prob. dist.

The **rate of convergence** depends on how quickly $\vec{a}(t) \to \vec{p}$, so we want to design Markov chains that **rapidly mix.**

Prop 4.4. For two prob dist. p and q,

$$|p - q|_1 = 2 \sum_i (p_i - q_i)^+ = 2 \sum_i (q_i - p_i)^+,$$

where $x^+ = x$ if $x \ge 0$ and $x^+ = 0$ if $x < 0$. (ReLU)

Metropolis-Hastings

Let $\vec{\pi} = (\pi_1, \dots, \pi_n)$ be our desired stationary dist. on states $1, \dots, n$.

Start with any connected undirected graph on the states with max. degree r.

Then at node i, for each adjacent node, we have $\frac{1}{r}$ chance of choosing it (and $\frac{r - \deg(i)}{r}$ chance of choosing none of them).

Then at node i, for each adjacent node, we have $\frac{1}{r}$ chance of choosing it (and $\frac{r - \deg(i)}{r}$ chance of choosing none of them).

If we've chosen a node j, if $\pi_j \geq \pi_i$, go to j.
if $\pi_i > \pi_j$, go to j w.p. $\frac{\pi_j}{\pi_i}$.

Otherwise, stay in place.

i.e.
$$p_{ij} = 0 \quad \text{if edge } (i,j) \text{ does not exist and } i \neq j.$$
$$p_{ij} = \frac{1}{r} \min\left(1, \frac{\pi_j}{\pi_i}\right), \quad \text{if edge } (i,j) \text{ exists and } i \neq j.$$
$$p_{ii} = 1 - \sum_j p_{ij}$$

Thus, if edge (i,j) does not exist $\pi_i p_{ij} = 0 = \pi_j p_{ji}$.
if edge (i,j) does exist, $\pi_i p_{ij} = \frac{\pi_i}{r} \min\left(1, \frac{\pi_j}{\pi_i}\right) = \frac{1}{r} \min(\pi_i, \pi_j) = \frac{\pi_j}{r} \min\left(1, \frac{\pi_i}{\pi_j}\right) = \pi_j p_{ji}$.

$\Rightarrow$ π_i is the stationary dist. of the Markov chain.

Note that this is true for whatever connected undirected graph we choose, but the graph structure will affect mixing time, (time to convergence)

Gibbs sampling

Let $p(\vec{x})$ be the target distribution, where $\vec{x} = (x_1, \ldots, x_d)$.

Let $G = (V, E)$ be an undirected graph where V is the set of all states $\vec{x}$ and $(\vec{x}, \vec{y}) \in E$ iff $|\vec{x} - \vec{y}|_0 = 1$. (i.e. if the two states differ in only one coordinate)

Each step, we choose only one coordinate to update, based on marginal prob. with other coordinates fixed. (e.g. randomly choose coordinate, or sequentially scan)

Suppose $|\vec{x} - \vec{y}|_0 = 1$. WLOG, suppose the coordinate that differs is the first.

Then set $p_{\vec{x}\vec{y}} = \frac{1}{d} p(y_1 \mid x_2, x_3, \ldots, x_d)$ (randomly choose coordinate)

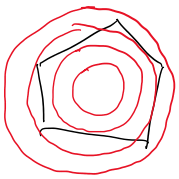
$$\Rightarrow p(\vec{x}) p_{\vec{x}\vec{y}} = \frac{p(\vec{x})}{d} p(y_1 \mid x_2, \ldots, x_d) = \frac{p(x_1, \ldots, x_d) \cdot p(y_1, \ldots, y_d)}{d \cdot p(x_2, \ldots, x_d)} = \frac{p(\vec{y}) \cdot p(x_1 \mid y_2, \ldots, y_d)}{d} = p(\vec{y}) \cdot p_{\vec{y}\vec{x}}.$$

Estimating high-dim. volume using MCMC

Hard for irregular figures, though we computed them for spheres & cubes.

Rejection sampling alone fails in high dimensions.

Rejection sampling alone fails in high dimensions.



Instead, for a convex set R, choose concentric circles $S_1, \ldots, S_k$ s.t. $S_1 \subseteq R \subseteq S_k$

Then $\mathrm{Vol}(R) \approx \mathrm{Vol}(S_k \cap R)$

$$= \frac{\mathrm{Vol}(S_k \cap R)}{\mathrm{Vol}(S_{k-1} \cap R)} \cdot \frac{\mathrm{Vol}(S_{k-1} \cap R)}{\mathrm{Vol}(S_{k-2} \cap R)} \cdots \frac{\mathrm{Vol}(S_2 \cap R)}{\mathrm{Vol}(S_1 \cap R)} \cdot V(S_1)$$

If the radius of S_i is $1 + \frac{1}{d}$ times the radius of S_{i-1},

then $\mathrm{Vol}(S_i) = \left(1 + \frac{1}{d}\right)^d \mathrm{Vol}(S_{i-1})$.

$$\Rightarrow 1 \le \frac{\mathrm{Vol}(S_i \cap R)}{\mathrm{Vol}(S_{i-1} \cap R)} \le \left(1 + \frac{1}{d}\right)^d \approx 2$$

(by convexity of R and the circles being centered in R)

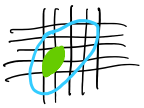
So we can estimate this ratio by rejection sampling, provided we can randomly sample from $S_i \cap R$.

The number of spheres needed is $O(\log_{1+\frac{1}{d}} r) = O(rd)$, where r is the ratio of radii b/t S_k and S_1.

So if each ratio is estimated to error $1 \pm \frac{\varepsilon}{4rd}$, then the overall error will be $1 \pm \varepsilon$.

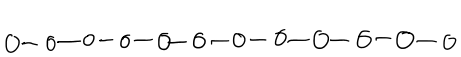
How to estimate ratios?

Use Metropolis-Hastings on a grid imposed on a region, and just walk along the grid.
The stationary prob we want is uniform on all grid points inside $S_i \cap R$,
which implies that it's just an ordinary random walk on the undirected graph with nodes in $S_i \cap R$.
This allows us to sample $\frac{S_{i-1} \cap R}{S_i \cap R}$.

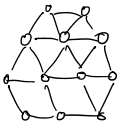


Fast convergence due to the grid structure. (not proven right now).

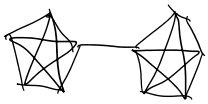
Convergence of random walks on undirected graphs



slow convergence



fast convergence.



slow convergence because of constriction

Given an edge-weighted graph with w_{xy} the weight between vertices x & y, ($w_{xy}=0$ if no edge).

Let $w_x = \sum_y w_{xy}$, and let $p_{xy} = \frac{w_{xy}}{w_x}$ be the Markovian transition prob.

(If adjacency matrix $A = \{w_{xy}\}$, let $D = \text{diag}(A\vec{1})$, so $P = D^{-1}A$.)

Then the stationary dist. $\vec{\pi}$ has $\pi_x = \frac{w_x}{w_{total}}$, $w_{total} = \sum w_x$, because

$$\pi_x p_{xy} = \frac{w_x}{w_{total}} \cdot \frac{w_{xy}}{w_x} = \frac{w_{xy}}{w_{total}} = \frac{w_y}{w_{total}} \cdot \frac{w_{yx}}{w_y} = \pi_y \cdot p_{yx} = \pi_y \cdot p_{yx}.$$

Let's determine how quickly the Markov chain converges to the stationary dist. $\vec{\pi}$.

Def. Fix $\varepsilon > 0$. The ε-mixing time of a Markov chain is

$$\min_{\substack{\vec{p}(0) \\ \text{prob. dist.}}} \left(t \mid |\vec{a}(t) - \vec{\pi}|_1 < \varepsilon \right), \quad \text{where } \vec{a}(t) = \frac{1}{t}\left(\vec{p}(0) + \cdots + \vec{p}(t)\right).$$

Def. For a subset S of vertices, let $\pi(S) = \sum_{x \in S} \pi_x$. The normalized conductance

$$\Phi(S) = \frac{\sum_{(x,y)\in(S,\bar{S})} \pi_{xy} p_{xy}}{\min\left(\pi(S), \pi(\bar{S})\right)}$$

Interpretation, WLOG, say $\pi(S) \le \pi(\bar{S})$. Then

$$\Phi(S) = \sum_{x \in S} \underbrace{\frac{\pi_x}{\pi(S)}}_{\substack{\text{prob in} \\ \text{stationary} \\ \text{dist. cond.} \\ \text{on in } S}} \underbrace{\sum_{y \in \bar{S}} p_{xy}}_{\substack{\text{prob of} \\ \text{leaving} \\ S \text{ in} \\ 1 \text{ step}}}$$

Note $\mathbb{E}[\#\text{ steps to go from } S \text{ to } \bar{S}] = \Phi(S) + 2(1-\Phi(S))\Phi(S) + 3(1-\Phi(S))^2\Phi(S) + \cdots$

$$= \Phi(S)\left[1 + 2(1-\Phi(S)) + 3(1-\Phi(S))^2 + \cdots\right]$$

$$= \Phi(S) \cdot \frac{1}{\left(1-(1-\Phi(S))\right)^2} = \frac{1}{\Phi(S)}.$$

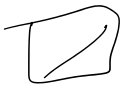
Recall $\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i$

$\frac{1}{(1-x)^2} = \sum_{i=1}^{\infty} i x^{i-1}$

Obviously need to be able to reach all regions in order to mix, so must reach any S. So $\frac{1}{\Phi(S)}$ is a lower bound on mixing time.

Def. The normalized conductance of a Markov chain, denoted Φ, is

$$\Phi = \min_{S \subseteq V,\, S \neq \emptyset} \bar{\Phi}(S).$$



states

state

$\frac{1}{t}(\vec{p}(t) - \vec{p}(0))$

to A