4. Markov chains and MCMC

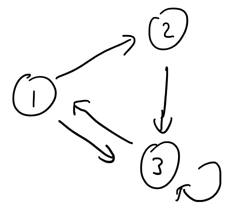
Wednesday, September 15, 2021 11:22 PM

Previously: Sampling from high-dimensional unit balls is hard naively, but we can use Gaussians to do so.

Today:
- How can we sample from complicated distributions in high dimensions?
- Markov Chain Monte Carlo (MCMC)

Markov Chains and random walks on graphs

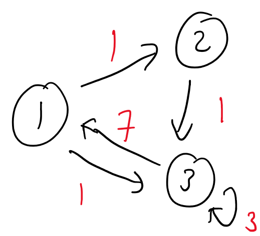
- Consider a directed graph $G=(V,E)$, $E \subseteq V^2$.



- A path P in G starting at x and ending at y is a sequence of vertices $P=(v_0,\dots,v_k)$ where $v_i \in V$, $(v_i, v_{i+1}) \in E$, $v_0 = x$, and $v_k = y$.
  - └ $\text{length}(P) = k$

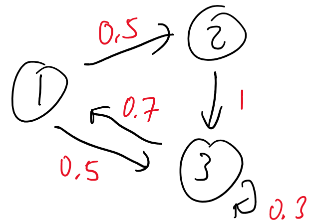
Ex. $(1,2,3,1,2)$

- A graph G is strongly connected if $\forall x,y \in V$, $\exists$ a path from x to y.

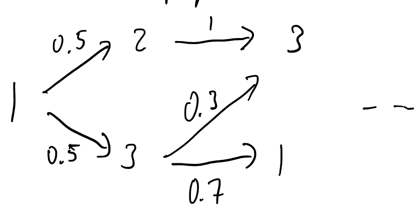


- A weighted graph additionally assigns a positive value to each edge, called a weight
  - └ We can view an unweighted graph as a graph where every weight is 1!

- A random walk on a graph is a path generated by starting from a vertex, and then iteratively choosing the next vertex by travelling along edges
  - └ The standard random walk fixes the transition probabilities as proportional to edge weight.



Ex.



Let the matrix P have $p_{ij} = \text{Prob.}(\text{transition from } i \text{ to } j)$. Then $\vec{p}(t+1) = \underbrace{\vec{p}(t)}_{\text{prob. dist. at time } t} P$

Ex.

$$P = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0.7 & 0 & 0.3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} P = \begin{bmatrix} 0 & 0.5 & 0.5 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} P^2 = \begin{bmatrix} 0.35 & 0 & 0.65 \end{bmatrix}$$

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

Random walk on graph $\Longleftrightarrow$ Markov chain

vertices $\Longleftrightarrow$ states

strongly connected graph $\Longleftrightarrow$ connected Markov states

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

Def. The long-term average 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 $A = \left[P - I,\ \vec{1}_n^{\,T}\right]$ has rank n. $\left(\vec{1}_n^{\,T} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}\Big\} n\right)$

proof. Suppose $\operatorname{rank}(A) \neq n$. Then $\operatorname{rank}(A) < n \Rightarrow \dim(\operatorname{Null}(A)) \geq 2$.

$P\vec{1}_n^{\,T} = \vec{1}_n^{\,T}$ because each row in P sums to 1 as a prob. distribution.

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. $A[\vec{x}, \alpha]^T = 0$ (second solution)

$\Rightarrow \left[P - I,\ \vec{1}_n^{\,T}\right]\begin{bmatrix} \vec{x}^{\,T} \\ \alpha \end{bmatrix} = (P - I)\vec{x}^{\,T} + \alpha\vec{1}_n^{\,T} = 0$

Thus, $\forall i$, $\sum_{j=1}^{n} p_{ij}x_j - x_i + \alpha = 0 \Rightarrow x_i = \sum_{j=1}^{n} p_{ij}x_j + \alpha$ (each x_i is a convex comb of x_j's and α)

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 $\sum_{j=1}^{n} p_{ij}x_j < \sum_{j=1}^{n} p_{ij}x_i = x_i$ (we use connectedness here)

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

However, repeat logic letting $x_i \leq x_j$ for all j. Then $\alpha < 0$.

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

Fundamental Thm of Markov Chains For a connected Markov chain, there is a unique prob. vector $\vec{\pi}$ satisfying $\vec{\pi}P = \vec{\pi}$. Moreover, for

Fundamental Thm of Markov Chains For a connected Markov chain, there is a unique prob. 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 $\vec{b}(t) = \vec{a}(t) P - \vec{a}(t) = \frac{1}{t}\left(\vec{p}(t) - \vec{p}(0)\right)$.

Aside: We are going to use the 1-norm today $|\cdot| = \|\cdot\|_1$.

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

Thus, the limit exists.

By Lemma, let $A = [P - I, \ \mathbb{1}_n^T]$. $\text{rank}(A) = n$. $A = [C_1 \ C_2 \ \cdots \ C_{n+1}]$, $C_{n+1} = \mathbb{1}_n^T$.

Let $B = [C_2 \ C_3 \ \cdots \ C_{n+1}]$, an $n \times n$ submatrix.

$\text{rank}(B) = n$ because $C_1 + C_2 + \cdots + C_n = 0$, so this set is linearly dep.

$\Rightarrow B$ is invertible.

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

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

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

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

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

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

- Aside: why did we not just say $\vec{\pi} = \lim_{t\to\infty} \vec{p}(t)$?

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 $\vec{\pi} = \vec{\pi} P$. $\boxed{/}$

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 exponentially many possible $\vec{x}$.
Rather want to sample points $\vec{x}$ according to p, so we don't need to evaluate everywhere
e.g. equiv of finding mean by drawing random samples

MCMC draws a sample $\vec{x}$ from $p(\vec{x})$ by designing a Markov chain whose stationary distr is $p(\vec{x})$.

↳ Common variations include Metropolis-Hastings & Gibbs Sampling whose stationary distr is $p(\vec{x})$.

↳ common variations include Metropolis-Hastings & Gibbs Sampling

first, let's prove it works in general.

$\mathbb{E}_{\vec{x}\sim p} f(\vec{x}) = \sum_{\vec{x}} p(\vec{x}) f(\vec{x})$.   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 avg of f along nodes in a t-step walk $(w_0, \ldots, w_t)$

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

$$\underset{\substack{\uparrow \\ \text{over } t\text{-step walks}}}{\mathbb{E}} Y = \sum_i f_i \cdot \left(\frac{1}{t} \sum_{j=1}^{t} \text{Prob}(w_j = i) \right) = \sum_i f_i \cdot a_i(t).$$

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

Then $\left| \sum_i f_i p_i - \mathbb{E} Y \right| \le f_{max} \sum_i |p_i - a_i(t)| = f_{max} \overbrace{\left| \vec{p} - \vec{a}(t) \right|_1}^{\substack{\text{total variation distance} \\ 1\text{-norm}}}$

So we can bound the performance of 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 define Markov chains that rapidly mix.