Lecture 17: Markov Chain Monte Carlo (MCMC)

Sampling: given a potentially incomplete description of a probability distribution, or indirect access to it, generate new, fresh samples from this distribution.

e.g. generative models:

Stable diffusion:
"generate a picture of a cat" → 

input: a dataset of images of cats

output: another picture of a cat

not from dataset! But still from distribution of images of cats!

next lecture: diffusion models!

Monte Carlo methods:

Sometimes, sampling is a useful way of efficiently computing/approximating some hard process or system.

Such approaches are referred to as Monte Carlo Simulations

e.g. approximating π: suppose I have the following dartboard:

What is the probability that a random dart falls in the circle?

$$= \frac{\text{area of circle of radius } r}{\text{area of square w/ sidelength } 2r} = \frac{\pi r^2}{(2r)^2} = \frac{\pi}{4}.$$

Idea: throw n darts, and guess $\pi \approx 4 \cdot$ (fraction of darts in circle).

Let $X_i = \mathbb{1}[\, i^{th} \text{ dart is in circle} \,]$.

Then fraction of darts in circle $= \frac{1}{n}\sum_{i=1}^{n} X_i$ (✓ independent.)

$\mathbb{E}[X_i] = \pi/4$.

Chernoff: $\Pr\left[\left|\frac{1}{n}\sum X_i - \frac{\pi}{4}\right| \geq \varepsilon\right] \leq \exp(-cn\varepsilon^2)$ (some universal constant)

So if we want an ε-approximation of $\pi/4$ w.p. $1-\delta$, can take $n = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$.

See also: Buffon's needle

An important special case of Monte-Carlo methods: Markov Chain Monte Carlo (MCMC)

Def: A **Markov process** is specified by:

1. A set of states $S = \{s_1, \dots, \}$
2. A transition function
$$P: S \times S \to [0,1].$$

Here, for any s_i, s_j, $P(s_i, s_j) = \Pr[\text{1 transition to } s_j \mid \text{in state } s_i]$.

In particular, $\forall s_i$, $\sum_{s_j} P(s_i, s_j) = 1$.

Given a Markov process, and an initial state X_0, this defines a **random walk**

X_0 X_1 X_2 ...
(all random!)

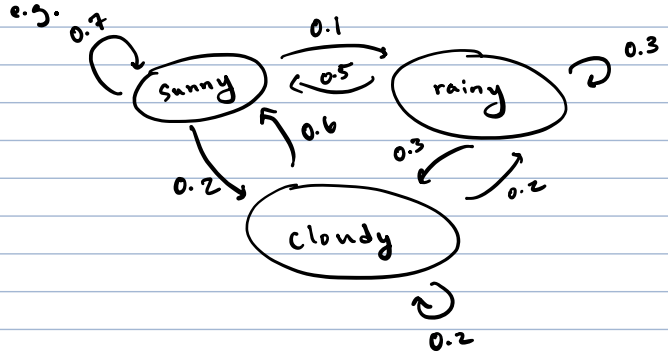
where for all $i \geq 1$, we sample X_i by choosing a random transition from X_{i-1}

$$\Pr[X_i = s \mid X_{i-1} = t] = P(s,t).$$

When there are n states, we usually represent transitions as a $n \times n$ transition matrix

$$P_{ij} = P(s_i, s_j).$$

e.g.



	sunny	cloudy	rainy
sunny	0.7	0.2	0.1
cloudy	0.6	0.2	0.2
rainy	0.5	0.2	0.3

Key property of Markov Chains: the transition probability at time t depends only on the state X_t, and not how we got to X_t.

"Markov property". "memory-less"

Examples:

- weather systems
- (some) board games: Monopoly, snakes + ladders
- genetics and inheritance

Some examples of not Markov chains:

- blackjack (if you track state of deck)
- stock market (if people learn from the past...?)

Computing with Markov Chains:

Suppose we have initial state $X_0 = s$. Let v_t be the probability distribution of X_t, i.e.:

$$v_t(x) = \Pr[X_t = x \mid X_0 = s].$$

1). How can we compute v_t?

2). What is the behavior of v_t as $t \to \infty$?

Def: A probability distribution π on $\{s_1, \cdots, s_n\}$ is a stationary distribution if for all initial states X_0,

$$v_t(x) \to \pi(x) \quad \forall \; x \in \{s_1, \cdots, s_n\}.$$

e.g. Page Rank (Google's claim to fame).

states = {webpages on the internet}.

transition: w.p. 0.8, choose a random outgoing link from current website.

w.p. 0.2, transition to uniformly random website.

Claim: π (the stationary distribution) roughly corresponds to importance of pages

π is known as the Page Rank distribution.

Approach 1: Think of v_t as a n-dimensional row vector.

Initially, $v_0 = (0, 0, \cdots, \underset{\uparrow}{1}, 0, \cdots 0)$ (1 in position s)

Claim: $v_t = v_{t-1} \cdot P$

pf: Fix j. Then

$$\Pr[X_{t+1} = j \mid X_0 = s] = \sum_i \Pr[X_{t+1} = j \mid X_t = i] \cdot \Pr[X_t = i \mid X_0 = s]$$

$\Pr[X_{t+1} = j \mid X_t = i] = P_{ij}$, $\Pr[X_t = i \mid X_0 = s] = v_t(i)$

$$\begin{bmatrix} & \end{bmatrix} \begin{bmatrix} \\ \\ \end{bmatrix}$$

So by induction,

$$v_t = v_0 \cdot P^t. \quad \leftarrow \text{power iteration!}$$

under appropriate conditions, $v_t \to \pi$ quite quickly as we'll discuss later.

When n is small, can approximate quickly via iterated squaring:

$$P \to P^2 \to (P^2)^2 = P^4 \to (P^4)^2 = P^8 \cdots$$

What to do when n is very large?

MCMC: A way of approximating stationary distributions for "nice" Markov chains, even when the # of states can be huge.

Idea: just directly simulate the Markov Chain!

Initialize X_0 somehow
for $t = 1, \cdots, T$
- pick X_t at random from transition probability given X_{t-1}.

Common Algorithmic Framework:

1. Define a random walk / Markov chain so that transition probability induces a stationary distribution π w/ desirable properties
   - it encodes useful info
   - it puts more weight on good sol'ns.
2. Run MCMC
3. Hope it finds a good solution.

Often, even when # of states is very large, MCMC "mixes" rapidly to a good solution.

Analyzing mixing times

Q. Is the stationary distribution unique?

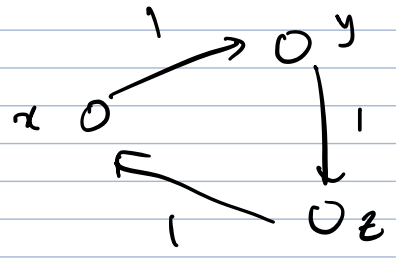
Q. How long do we have to run MCMC / power iteration to sample from something close to the stationary distribution?

Two barriers to unique stationary distributions:

1). Not strongly connected

Def: A MC is reducible if its digraph is not strongly connected. Otherwise, it is irreducible.

2. Periodicity



$X_0 = x$

$X_t = x \iff t \equiv 0 \mod 3$.

No stationarity for this starting distribution!

Def: A chain is aperiodic if

$$\gcd\left(\{t : \underbrace{P_t(x,y)}_{\text{prob 1 transition from } x \text{ to } y \text{ in } t \text{ steps}} > 0\}\right) = 1 \quad \forall x, y \in S.$$

In above: gcd = 3 so chain is periodic.

Fundamental theorem of Markov Chains: If a Markov Chain is irreducible and aperiodic, then it has a unique stationary distribution.

How quickly do we mix? ← what does it even mean to be "close" to a stationary distribution?

Def: Given two probability distributions μ, ν on S, the total variation distance is defined as

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_x |\mu(x) - \nu(x)| = \frac{1}{2}\|\mu - \nu\|_1$$

$$= \max_E \left| \Pr_\mu[E] - \Pr_\nu[E] \right|.$$

Def: The mixing time of a chain is defined to be:

$$\tau_{mix} = \min_t \left\{ d_{TV}(\pi, P_t) \leq \underbrace{1/4}_{\text{specific const not very important}} \text{ for all initial state } X_0 = s \right\}.$$

Fact: After $c \cdot \tau_{mix}$ steps, $d_{TV}(\pi | P_t) \leq e^{-c}$.

How to understand mixing time? One approach: spectral

Consider the special case where P is symmetric ("reversible"):

$$P(x,y) = P(y,x).$$

Recall: For any initial distribution v_0,

$$v_t = v_0 \cdot P^t$$

For a stationary distribution π, note that

$$\lim_{t\to\infty} v_t = \pi$$

$$\lim_{t\to\infty} v_t P = \pi P$$

$$\Big\downarrow$$

$$= \lim_{t\to\infty} v_{t+1} = \lim_{t\to\infty} v_t = \pi$$

so $\pi = \pi P \Rightarrow \pi$ is a eigenvector of P w/ eigenvalue 1.

Let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of P.

Fact: $|\lambda_i| \le 1 \ \forall i$

Fact: If P is irreducible, then $\lambda_{n-1} < 1$.

Fact: If P is aperiodic, then $\lambda_1 > -1$.

$\Rightarrow$ FT of Markov Chains!

Let $v_1, \dots, v_n \propto \pi$ be the eigenvectors of P. Since P is symmetric, it admits a spectral decomposition:

$$P = \sum \lambda_i v_i v_i^T = \frac{\pi\pi^T}{\|\pi\|_2^2} + \sum_{i\neq n} \lambda_i v_i v_i^T$$

$$P^t = \sum \lambda_i^t v_i v_i^T = \frac{\pi\pi^T}{\|\pi\|_2^2} + \sum_{i\neq n} \lambda_i^t v_i v_i^T$$

so $$vP^t = \frac{\langle v, \pi\rangle}{\|\pi\|_2^2}\pi^T + \sum_{i\neq n} \lambda_i^t \langle v_i, v_i\rangle v_i^T$$

$\lambda_i^t \to 0$ as $t \to \infty$.

(actually, this shows that for symmetric MC, π is uniform!).