A Stochastic process $\{X_t\}_{t \in T}$ is a collection of random variables indexed by some set T — we usually think of as time or space. These can be used to describe many phenomena, including evolution of a randomized algorithm.

A Markov chain is a particular kind of stochastic process. We will deal exclusively with discrete-time, homogeneous Markov chains over a set S of finite or countably infinite states. Such a chain is described by a transition matrix P whose i,jth entry describes the probability of moving from state i to state j in one time step.

Note that for all i, $\sum_j P_{ij} = 1$, because you will move to some state in the next time step. Call P a row-stochastic matrix

Markov chains have the property of memorylessness

$$\mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1}, \dots, X_0 = i_0)$$
$$= \mathbb{P}(X_t = i_t \mid X_{t-1} = i_{t-1})$$

It is not that X_t doesn't depend on the states before, it is that once we know X_{t-1}, we have all the relevant history from the past to predict X_t.

Note that $\mathbb{P}(X_t = j \mid X_{t-1} = i) = P_{ij}$

Markov Chains come up in many applications:

- might know of HMMs from signal processing, NLP, ML, etc.
- used in Bayesian statistic & ML to approximately sample from intractable distributions
- used to approximately count sets it is hard to sample from (e.g. perfect matchings in a graph) and approximate volume of complicated (convex) sets
- used to analyze behavior of randomized algorithms

In some of these applications, we use the <u>Markov Chain Monte Carlo</u> technique, which constructs a Markov Chain that has a specific limiting distribution for X_t, called a stationary distribution. Our goal is to reach that distribution as fast

as possible. Accordingly we care about

- when do stationary distributions exist?
- how do we measure convergence of distributions?
- what techniques can we use to show convergence?

These will be the focus of our intro to MCs.

Some examples:

- a random walk on $S = \mathbb{Z}$, where you move left/right w/ prob p or $1-p$

- a random walk on $\mathbb{Z}_m = \{0, \ldots, m-1\}$ where you increase/decrease the state $(\text{mod } m)$ w/ prob p or $1-p$

- random walks on a graph $G(V, E)$.
  Example: $P_{uv} = \begin{cases} \frac{1}{d(v)} & \text{if } (u,v) \in E \\ 0 & \text{otherwise} \end{cases}$

- random walk on the lattice $\mathbb{Z}^d$

$$P_{xy} = \begin{cases} \frac{1}{2d} & \text{if } y \text{ differs from } x \text{ by } 1 \text{ in exactly one coordinate} \\ 0 & \text{otherwise} \end{cases}$$

- lazy versions of Markov chains
  Given P, take $\tilde{P} = \alpha P + (1-\alpha) I$
  i.e. with some probability $(1-\alpha)$ stay in the same state.

- random transposition walk on the symmetric group S_n

$$P_{\pi_1 \pi_2} = \begin{cases} 1/\binom{n}{2} & \text{if } \pi_2 = \sigma \pi_1 \text{ for some transposition } \sigma \\ 0 & \text{otherwise} \end{cases}$$

This is often used to model card shuffling for example (famous Diaconis result on the mixing time for card shuffling)

- Ehrenfest urn model used to model particle diffusion. Have two urns 0 and 1, n balls. State is vector in $\{0,1\}^n$ describing location of balls. At each time pick a ball and change its location.

$$P_{xy} = \begin{cases} 1/n & \text{if } y \text{ differs from } x \text{ in one location} \\ 0 & \text{otherwise} \end{cases}$$

- A randomized alg for finding a satisfying assignment for the 2-SAT problem in n variables, which implicitly defines a Markov Chain. At each time pick an unsatisfied

clause and invert one of its literals. Note that the satisfying assignments are absorbing (probability of leaving them = 0). Can show using MC analysis that if a satisfying assignment exists, this alg. will find in $O(n^2)$ time. Same idea works for k-SAT, k-colorability, etc., but convergence time is different (also constructive version of Lovasz Local Lemma)

— TALK about DP example here

Basics

P — one-step transition matrix describes

$$\mathbb{P}(X_{t+1} = j \mid X_t = i)$$

General s-step transition matrix describes

$$\mathbb{P}(X_{t+s} = x_{t+s} \mid X_t = x_t) =$$

$$\sum_{x_{t+s-1}, \ldots, x_{t+1}} \mathbb{P}(X_{t+s} = x_{t+s} \mid X_{t+s-1} = x_{t+s-1}, \ldots, X_t = x_t)$$

Now recall we have seen before (e.g. in proof of LLL) that

$$\mathbb{P}(\mathcal{E}_n \mid \mathcal{E}_{n-1}, \ldots, \mathcal{E}_1) = \prod_{\ell=0}^{n-2} \mathbb{P}(\mathcal{E}_{n-\ell} \mid \mathcal{E}_{n-(\ell+1)})$$

ML-example: Differentially private SUD

TALK about earlier

Defn An algorithm A provides (ε, δ) differential privacy if

$$\mathbb{P}(A(D) \in S) \leq e^{\varepsilon}\mathbb{P}(A(D') \in S) + \delta$$

for all data sets D and D' differing in a single entry

So you can be sure that $A(D)$ does not inadvertently expose sensitive information on any one individual in D, if for example D computes some statistic on people.

Ex: take A to be max$_{\text{salary}}$, then if $A(D)$ and $A(D')$ changes when I remove one person, I know their salary

There is a tension b/w differential privacy and accuracy (e.g. if I add really high noise, trivially ensure DP) so one goal of DP researchers is: given a particular computation, how to achieve (ε, δ) privacy while ensuring as much accuracy as possible.

Consider case where want to DP compute

$$\underset{x \in S}{\operatorname{argmin}} f(x)$$

One way to achieve this is to sample from a distribution that is peaked around the minimum instead of exactly solving for the minimum, e.g. sample from dist

$$\mathbb{P}^{\gamma}(x \in A) \propto \int_{A} \exp(-\gamma f(x))\, dx \qquad (A \subseteq S)$$

Its clear that adjusting γ trades off b/w DP and accuracy

Ex: DP EVD (has optimal sample complexity)

given A, DP compute a basis for col span of A

so $A \approx U \Sigma U^T$ where $U \in \mathbb{R}^{n \times k}$ satisfies $U^T U = I$

exactly solve with

$$\underset{\substack{U \in \mathbb{R}^{n \times k} \\ U^T U = I}}{\operatorname{argmin}} -\operatorname{Tr}(U^T A U)$$

so DP solve by sampling from

$$\mathbb{P}(U \in A) \propto \int_A e^{-\frac{\gamma}{2} \mathrm{tr}(U^T A U)} \, dU$$

This is called the matrix Bingham distribution and has density

$$f(U) = \frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}n, A\right)} \exp\left(\mathrm{tr}(U^T A U)\right)$$

↑ confluent hypergeom fun

but we know no alg for directly sampling from

soln: use MCMC (in particular, Gibbs sampling)

so $\mathbb{P}(X_{t+s} \mid X_t) = \sum \mathbb{P}(X_{t+s} \mid X_{t+s-1}) \cdots \mathbb{P}(X_{t+1} \mid X_t)$

$$= \sum_{\substack{x_{t+s-1}, \ldots, \\ x_{t+1}}} P_{x_{t+s-1}, x_{t+s}} \cdots P_{x_t, x_{t+1}}$$

$$= (P^s)_{x_t, x_{t+s}}$$

So the s-step transition matrix is the s-th power of the 1-step transition matrix.

Denote our distribution over the states S by a row vector π^t at time t, then the distribution over the states at time $t+1$ are given by

$$(\pi^{t+1})_j = \mathbb{P}(X_{t+1} = j) = \sum_k \mathbb{P}(X_{t+1} = j \mid X_t = k)\mathbb{P}(X_t = k)$$

$$= \sum_k \pi_k P_{kj} = (\pi^t P)_j$$

so $\pi^{t+1} = \pi^t P$

In particular, if π^0 describes the initial state distribution, then

$$\pi^t = \pi^0 P^t$$

Stationary distributions

We call π a stationary distribution for the MC described by P if

$$\pi = \pi P$$

This means if we start the MC w/ state following π, the resulting chain always has its states distributed as π.

Note that a stationary distribution always exists: (argument for finite state spaces, can use diff technique for inf)

- P is a row-stochastic matrix, so

$$P\mathbf{1} = \mathbf{1}$$

means 1 is a right-eigenvalue of P,

i.e.

$$\det(P - \lambda I) \text{ has a zero at } \lambda = 1$$

$$\Rightarrow \det(P^T - \lambda I) \text{ has a zero at } \lambda = 1$$

$$\Rightarrow P^T \pi^T = \pi^T \text{ for some row-vector } \pi$$

Claim π has constant sign on the irreducible components of P

Pf

Let $P = P_1 \oplus \cdots \oplus P_m$ be the irreducible

$=$

components of F. Note that $P^n = P_1^n \oplus \cdots \oplus P_m^n$ and $\pi P^n = \pi_1 P_1^n \oplus \cdots \oplus \pi_m P_m^n$ where π_i indicates the portion of π restricted to component i

$$\pi P^n = \pi \Rightarrow \pi_i P_i^n = \pi_i$$

Since P_i is irreducible, there is an n_i such that $P_i^{n_i} > 0$ everywhere. Note that if π_i does not have constant sign,

$$|\pi_i P_i^{n_i}| < \pi_i$$

a contradiction ▨

Consequently $\pi P^n = \pi$ means π has constant sign on irreducible components. WLOG can take π to have positive sign on all components, showing π is a prob vector (after scaling) s.t.

$$\pi P = \pi$$

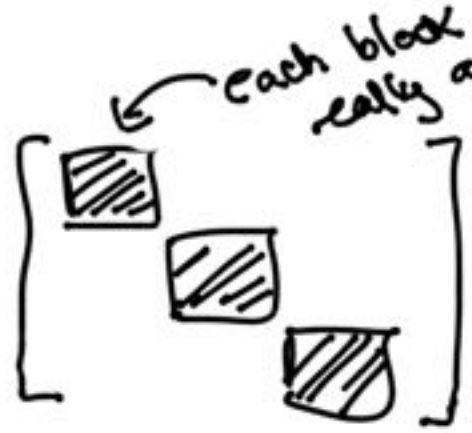
If the stationary distribution is unique, and we can handle the task, we can find it by solving the n+1 equations

$$\pi P = \pi$$
$$\pi 1 = 1$$

Ex of Markov Chains w/ multiple stationary distributions:

— $P = I$

— more generally $P =$



such chains are called <u>reducible</u>. Once your state reaches one of those blocks, you never leave it. So here we have e.g. three stationary distributions

Note that even if a Markov chain has a unique stationary distribution, π, it is not always the case that

$$\pi = \lim_{t \to \infty} \pi_0 P^t$$

for all initial distributions π_0

Example: random walk on a bipartite graph

stationary dist is unif dist,

but π_0 is supported only on L,

then π_1 is supported only on R,

and so on, so π^t has no limit as $t \to \infty$

In general call a MC periodic if the gcd of the set $\{n : (P^n)_{ii} > 0\}$ is not 1.

E.g. for the bipartite graph

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \Rightarrow P^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \Rightarrow P^3 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$P^4 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

so is periodic with period 1.

Using lazy versions of random walks ensures aperiodicity e.g.

$$P = \begin{bmatrix} .1 & .9 \\ .9 & .1 \end{bmatrix}$$

Important question: when are stationary distributions both unique & reachable from any starting distribution? Relevant for the design of MCMC methods.

Thm An irreducible MC on a finite state space has a unique stationary dist. If it is additionally aperiodic, then

$$\pi = \lim_{t\to\infty} \nu P^t$$

↑ π

for any starting distribution ν

NB: call aperiodic, irreducible MCs ergodic (time averages give statistical averages) ↪ enables MCMC

$\frac{1}{T}\sum_{t=1}^{T} X_t \to \mathbb{E}_\pi X$

One useful tool for proving stationarity and/or constructing MCs with given stationary distributions is the concept of reversibility.

We say a MC is reversible if there is a distribution π s.t.

$$\pi_i P_{ij} = \pi_j P_{ji}$$

This is called the detailed balance equation. This means if the chain is started w/ initial distribution π, then

$$\mathbb{P}(X_0 = i, X_1 = j) = \mathbb{P}(X_0 = j, X_1 = i)$$

so there is no "time's arrow"

Note if detailed balance holds then

$$(\pi P)_j = \sum_i \pi_i P_{ij} \overset{= \sum_i \pi_j P_{ji}}{=} \pi_j$$

so $\pi = \pi P$ and π is a stationary distribution

Ex of a reversible MC:

random walk on a graph w/ $\pi_j = \frac{d(v_j)}{\sum_k d(v_k) := D}$

$$\pi_i p_{ij} = \begin{cases} 0 & \text{if } (v_i, v_j) \notin E \\ \frac{d(v_i)}{D} \cdot \frac{1}{d(v_i)} & \end{cases} = \frac{1}{D} \text{ otherwise}$$

$$\pi_j p_{ji} = \begin{cases} 0 & \text{if } (v_i, v_j) \notin E \\ \frac{1}{D} & \text{otherwise} \end{cases}$$

so detailed balance holds

<u>Now we can consider the question of convergence to stationarity</u>

How to measure convergence? KL-divergence, Hellinger, S-J divergence, ... many choices, some equivalent

Well what do we care about? Imagine we know an algorithm A succeeds w/ prob at least $1-\delta$ if randomness comes from an expensive to sample dist P. Instead we can sample from a dist Q that is close to P in the sense that

$$\sup_{E \in \mathcal{F}} |P(E) - Q(E)| < \varepsilon$$

Why ergodicity is useful? It implies that if X_t is ergodic MC with limit dist π, then

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t) \longrightarrow \mathbb{E}_\pi f(X)$$

where $X \sim \pi$ (and f is sufficiently regular, e.g. continuous)

This allows us to do many things, e.g. estimate intractable integrals. Ex: assume $f \geq 0$, then can use Gibbs sampling or Metropolis-Hastings algorithm to approx

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n)\, dx$$

by sampling from an appropriate MC.

Important algorithmic application of reversibility:
Gibbs sampling

Imagine want to sample from function

$$f(x_1, \dots, x_n) \geq 0$$

e.g. to approximate an intractable integral

$$\int_{\mathbb{R}^n} f(x_1, \dots, x_n) \approx \frac{1}{m} \sum_{i=1}^{m} f(x_1^{(i)}, \dots, x_n^{(i)})$$

where $(x_1^{(i)}, \dots, x_n^{(i)})$ i.i.d. samples from prob dist w/ density f.

Don't know how to do this in general, so construct a MC w/ prob density given by f as its stationary dist (note this is not a finite-state MC)

$$\mathbb{P}\left((x_1^{(i+1)}, \dots, x_n^{(i+1)}) \,\middle|\, (x_1^{(i)}, \dots, x_n^{(i)}) \right) =$$

$$\begin{cases} 0 & \text{if the states differ in more than one position} \\ \frac{1}{n} \dfrac{f(x_1^{(i)}, \dots, x_{j-1}^{(i)}, z, x_{j+1}^{(i)}, \dots, x_n^{(i)})}{\sum_{\omega} f(x_1^{(i)}, \dots, x_{j-1}^{(i)}, \omega, x_{j+1}^{(i)}, \dots, x_n^{(i)})} & \text{if } x_j^{(i+1)} = z \end{cases}$$

so think of this as RW on an infinite weighted graph w/ stat. dist f.

Proof that f is stat. dist uses reversibility

$$f(\bar{x})\,P_{\bar{x}\bar{y}} = f(\bar{y})\,P_{\bar{y}\bar{x}} \quad \text{if } \bar{x},\bar{y} \text{ differ in more than one spot}$$

$$f(\bar{x})\,P_{\bar{x}\bar{y}} = f(\bar{x})\,\frac{1}{n}\sum_{\substack{\bar{z}=\bar{x} \\ \text{except poss.} \\ \text{at pos } j}} \frac{f(\bar{y})}{f(\bar{z})} \quad \text{if } \bar{x},\bar{y} \text{ differ in at most pos } j$$

$$= \frac{1}{n}\sum_{\substack{\bar{z}=\bar{y} \\ \text{except poss.} \\ \text{at pos } j}} \frac{f(\bar{x})f(\bar{y})}{f(\bar{z})} = f(\bar{y})\,\frac{1}{n}\cdots\sum\frac{f(\bar{x})}{f(\bar{z})}$$

$$= f(\bar{y})\,P_{\bar{y}\bar{x}}$$

shows stationarity, but what about irreducibility and aperiodicity? Must establish for specific case being looked at. e.g.

← Gibbs would give reducible MC

General condition: if $f>0$ everywhere, then

$P_{\bar{x}\bar{x}} > 0 \Rightarrow$ irreducible

$P_{\bar{x}\bar{y}}^{\,n} > 0$ for some $n \Rightarrow$ aperiodic (think of a path from $\bar{x}$ to $\bar{y}$ of length n)

so have ergodicity

In physics in particular, often have $f = \exp(H(x_1, \ldots, x_n))$ so $f > 0$ and Gibbs sampling works (also, the DP example from last lecture)

Drawback: Gibbs sampling assumes can sample from the marginals $f(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$. This is not always do-able

Another approach: Metropolis-Hastings

Assume we have a reversible MC P on our state space w/ stat. dist μ and we want samples from a stationary distribution π. All we have is that π is proportional to f, that is,

$$\pi_j = \frac{f_j}{Z} \quad \text{for all } j, \text{ for an unknown } Z$$

We construct a chain Q w/ transition probabilities

$$Q_{ij} = P_{ij} \min\left(1, \frac{\mu_i \pi_j}{\mu_j \pi_i}\right)$$

$$= P_{ij} \min\left(1, \frac{\mu_i f_j}{\mu_j f_i}\right) \quad \text{for } i \neq j$$

and $Q_{ii} = 1 - \sum_{j \neq i} Q_{ij}$

Then we can show that this chain is reversible:
WLOG, take $\mu_i \pi_j > \mu_j \pi_i$, then

$$Q_{ij} = P_{ij}, \text{ so}$$

$$\pi_i Q_{ij} = \pi_i P_{ij} = \pi_i \frac{(P_{ij}\mu_i)}{\mu_i}$$

$$= \pi_i \frac{(P_{ji}\mu_j)}{\mu_i} \quad \text{bc reversibility of } P$$

$$= P_{ji} \frac{(\mu_j \pi_i)}{\mu_i}$$

$$= \pi_j P_{ji} \left(\frac{\mu_j \pi_i}{\mu_i \pi_j}\right)$$

$$= \pi_j Q_{ji}$$

and Q is reversible and has π as its stationary distribution. We call P the <u>proposal distribution</u> and

$A = \min\left(1, \frac{\mu_i f_j}{\mu_j f_i}\right)$ is the acceptance function.

In practice, sample j from the proposal dist, then decide to accept by sampling a unif$(0,1]$ var U, then checking $U \leq A$

Then we can be sure that the alg using Q is cheaper and succeeds w/ prob at least $1-(\delta+\varepsilon)$.

We call this notion of distance total variation distance (TV distance):

Fix $(\Omega, \mathcal{F})$ a sigma algebra and let P, Q be prob measures over this space, then

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$$

By symmetry arguments it isn't hard to see that

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A)$$

Pf

Let S achieve $d_{TV}(P,Q) = \sup_{A \in \mathcal{F}} |P(A) - Q(A)|$

either $P(S) > Q(S)$ in which case nothing to show, or $P(S) \leq Q(S)$. In this case, note that

$$P(S^c) \geq Q(S^c) \text{ and}$$

$$P(S^c) - Q(S^c) = [1 - P(S)] - [1 - Q(S)] = Q(S) - P(S)$$

so $|P(S) - Q(S)| = Q(S) - P(S)$
$= P(S^c) - Q(S^c)$

thus S^c is a witness that

$$\sup_A P(A) - Q(A) \geq \sup_A |P(A) - Q(A)|$$

the other inequality is obvious, so

$$\sup_A P(A) - Q(A) = \sup_A |P(A) - Q(A)|$$

For finite-state space MCs, distributions can be written as vectors, and have

$$d_{TV}(P, Q) = \max_{S \subseteq [n]} |P(S) - Q(S)|$$

$$= \max_{S \subseteq [n]} \left| \sum_{i \in S} (\pi_P)_i - (\pi_Q)_i \right|$$

$$= \max_{S \subseteq [n]} \sum_{i \in S} (\pi_P)_i - (\pi_Q)_i$$

Pop Quiz

1) Argue if a MC w/ n states ($n>1$) has an absorbing state, then it is not ergodic

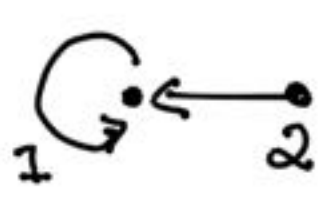
(a state i is absorbing if $P_{ii} = 1$)

but give an example of an MC w/ an absorbing state that has a unique stat dist. (and prove)

$$(P^n)_{ij} = \sum_k P_{ik}(P^{n-1})_{kj} = \begin{cases} 0 & \text{if } j \neq i \\ P_{ii}(P^{n-1})_{ki} \end{cases}$$

$$= \begin{cases} 0 & \text{if } j \neq i \\ P_{ii}^n \leftarrow \text{by induction} \end{cases} = \begin{cases} 0 & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases}$$

$\Rightarrow P$ is not irreducible since cannot reach any other state from i



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

clear $\pi = [1 \quad 0]$ is the onique stat dist

$$(x_1\ x_2)\,P = [x_1 + x_2 \quad 0] = [1, 0]$$

so just one step from any init dist reaches stat dist.

$\Rightarrow$ stat dist is unique & you converge to it regardless of starting pt.

Moral: irreducibility is sufficient for uniqueness & conv, not necessary.

Pop Quiz

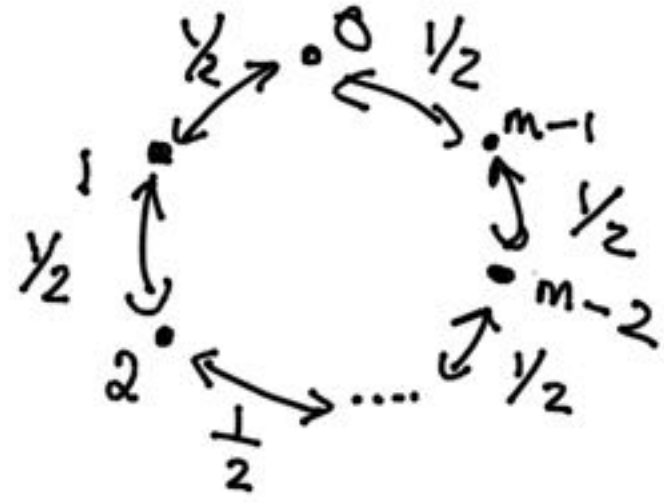
Let X be a Markov chain on $\mathbb{Z}_m = \{0, \dots, m-1\}$ w/ transition matrix P given by

$$P_{ij} = \begin{cases} 0 & \text{if } |(i-j) \bmod m| > 1 \\ \frac{1}{2} & \text{if } |(i-j) \bmod m| = 1 \end{cases}$$

Let

$$\tau_k = \min\{t \mid X_t = 0 \text{ given that } X_0 = k\}$$

be the hitting time for state 0 starting at state k



Note that τ_k is a random variable since it depends on the particular realization of the Markov Chain. E.g. if the chain looks like

$(1, 2, 3, 2, 1, 0, \dots)$ then $\tau_1 = 5$

$(1, 2, 3, 4, \dots, m-1, 0, \dots)$ then $\tau_1 = m-1$

Question: find a relationship b/w

$$\mathbb{E}[\tau_k], \mathbb{E}[\tau_{k-1}], \mathbb{E}[\tau_{k+1}]$$

Note that by total expectation,

$$\mathbb{E}[\tau_k] = \mathbb{E}[\tau_k \mid X_1 = k-1]\,\mathbb{P}(X_1 = k-1) + \mathbb{E}[\tau_k \mid X_1 = k+1]\,\mathbb{P}(X_1 = k+1)$$

(all addition is mod n)

and $\mathbb{E}[\tau_k \mid X_1 = k-1] = 1 + \mathbb{E}[\tau_{k-1}]$

$\mathbb{E}[\tau_k \mid X_1 = k+1] = 1 + \mathbb{E}[\tau_{k+1}]$

$$\mathbb{P}(X_1 = k-1) = \mathbb{P}(X_1 = k+1) = \frac{1}{2}$$

$$\Rightarrow \mathbb{E}[\tau_k] = \frac{1}{2}\left(1 + \mathbb{E}[\tau_{k-1}]\right) + \frac{1}{2}\left(1 + \mathbb{E}[\tau_{k+1}]\right)$$

$$= 1 + \frac{1}{2}\mathbb{E}[\tau_{k-1}] + \frac{1}{2}\mathbb{E}[\tau_{k+1}]$$

Claim

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

(recall $\|x\|_1 = \sum_{i=1}^{n} |x_i|$)

Pf

Let $A = \{i : x_i > y_i\}$

then $d_{TV}(x,y) = \max_{S \subseteq [n]} \sum_{i \in S} (x_i - y_i)$

$\leq \sum_{i \in A} (x_i - y_i) \Rightarrow d_{TV}(x,y) = \sum_{i \in A} (x_i - y_i)$

and likewise if $B = \{i : y_i > x_i\}$,

$$d_{TV}(x,y) = \sum_{i \in B} (x_i - y_i)$$

Thus $2 d_{TV}(x,y) = \sum_{i \in A} (x_i - y_i) + \sum_{i \in B} (x_i - y_i)$

$$= \sum_{i \in [n]} |x_i - y_i|$$

$$= \|x-y\|_1$$

Note that as a consequence, d_{TV} is a true metric on the probability measures on a given state space:

$$d_{TV}(P,Q) = \frac{1}{2}\|\pi_P - \pi_Q\|_1$$

$$\leq \frac{1}{2}\left(\|\pi_P - \pi_Z\|_1 + \|\pi_Z - \pi_Q\|_1\right)$$

$$= d_{TV}(P,Z) + d_{TV}(Z,Q)$$

$$d_{TV}(P,Q) \geq 0$$

$$d_{TV}(P,Q) = 0 \iff \frac{1}{2}\|\pi_P - \pi_Q\|_1 = 0 \iff \pi_P = \pi_Q$$
$$\iff P = Q$$

Of course can show all of these properties directly from defn of d_{TV} for arbitrary state spaces (not just finite)

If P, Q close in TV, can expect $\mathbb{E}_P f(x)$ is close to $\mathbb{E}_Q f(x)$ if f is well-behaved

Thm Let $|f(x)| \leq M$ for all $x \in \mathcal{X}$, then

$$|\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)| \leq 2M \, d_{TV}(P, Q)$$

Pf

$$|\mathbb{E}_P f(x) - \mathbb{E}_Q f(x)| = \left| \sum_x f(x)[P(x) - Q(x)] \right|$$

$$\leq M \sum_x |P(x) - Q(x)|$$

$$= M \, \|\pi_P - \pi_Q\|_1$$

$$= 2M \, d_{TV}(P, Q)$$

In particular, this implies a convergence rate for MCMC methods

Now we can start talking about how long it takes for a MC to converge to its stationary distribution — call this mixing time

Given a MC and a stationary distribution π, we say $t_{mix}(\varepsilon) = s$ if

for any initial distribution x,

$$d_{TV}(xP^t, \pi) \leq \varepsilon$$

when $t > s$.

Q: why does a mixing time exist. I.e., why is there <u>any</u> s s.t.

$$d_{TV}(xP^s, \pi) \leq \varepsilon$$

implies

$$d_{TV}(xP^t, \pi) \leq \varepsilon$$

for any $t \geq s$?

A: Because by our thm on convergence of an ergodic MC to its stationary dist, we know that $xP^t \to \pi$. In particular, $xP^t \to \pi$ in TV distance

(one proof is based on fact that if P is irreducible and aperiodic, it is eventually a contraction in $\|\cdot\|_1$ norm)

To prove results on mixing times, we use <u>couplings</u>. Given two distributions P & Q on state space Ω, we say a distribution Z on the state space $\Omega \times \Omega$ is a coupling of P and Q if the first marginal of Z is distributed according to P and the second marginal of Z is distributed according to Q. That is:

when $(x, y) \sim Z$,

first marginal $\rightarrow$ $Z(x \in A) = Z(x \in A, y \in \Omega) = P(A)$

second marginal $\rightarrow$ $Z(y \in A) = Z(x \in \Omega, y \in A) = Q(A)$

<u>Ex</u> Consider $P = Q$ is the uniform distribution over $\{0, 1, \ldots, n-1\}$, then one possible coupling is the <u>independent</u> coupling where

$$Z(x = i, y = j) = P(i)Q(j) = \frac{1}{n^2}$$

visually,

$Z =$

$\frac{1}{n^2}$	$\cdots$	$\frac{1}{n^2}$
$\vdots$		$\vdots$
$\frac{1}{n^2}$	$\cdots$	$\frac{1}{n^2}$

An example non-independent coupling is given by (visually)

$Z =$

where each row and column of Z has exactly two nonzero entries, both equal to $\frac{1}{2n}$.
It is then clear that

$$Z(x=i) = \sum_j Z(x=i, y=j) = 2 \cdot \frac{1}{2n} = \frac{1}{n} = P(i)$$

$$Z(y=j) = \sum_i Z(x=i, y=j) = 2 \cdot \frac{1}{2n} = \frac{1}{n} = Q(j)$$

so Z is a coupling of P & Q.

The <u>coupling lemma</u> is the key tool in applying coupling to determine the convergence of a MC

<u>Thm</u> Given a coupling Z of distributions P & Q, and a r.v. (X, Y) distributed according to Z,

$$d_{TV}(P, Q) \leq Z(X \neq Y)$$

Prf

We saw earlier that

$$d_{TV}(P,Q) = \sup_{A \in \mathcal{F}} P(A) - Q(A)$$

so it suffices to show that, given an arbitrary $A \in \mathcal{F}$,

$$P(A) - Q(A) \leq Z(X \neq Y).$$

To do so, note that

$$P(A) = Z(X \in A) = Z(X \in A \,\&\, Y \in \mathcal{X})$$
$$= Z(X \in A \,\&\, Y \in A) + Z(X \in A \,\&\, Y \in A^c)$$

and likewise

$$Q(A) = Z(X \in A \,\&\, Y \in A) + Z(X \in A^c \,\&\, Y \in A)$$

so

$$P(A) - Q(A) = Z(X \in A \,\&\, Y \in A^c) - Z(X \in A^c \,\&\, Y \in A)$$
$$\leq Z(X \in A \,\&\, Y \in A^c) + Z(X \in A^c \,\&\, Y \in A)$$
$$\leq Z(X \neq Y)$$

To use this result to prove convergence for a Markov chain starting in an arbitrary distribution π_0 to its stationary distribution π, given its transition matrix P, we will construct a series of couplings Z_t between the distributions $\pi_0 P^t$ and π such that if $(X_t, Y_t) \sim Z_t$ then:

1) $\{X_t\}$ is a Markov Chain with initial distribution π_0 and transition matrix P

2) $\{Y_t\}$ is a MC w/ init dist π and transition matrix P

3) If $X_{t_0} = Y_{t_0}$ at any time step t_0, then the two chains $\{X_t\}$ and $\{Y_t\}$ have coalesced:
for any $t \geq t_0$, $X_t = Y_t$

The first two properties & the coupling lemma imply

$$d_{TV}(\pi_0 P^t, \pi) \leq Z_t(X_t \neq Y_t)$$

Ex:

and the last property implies that

$$Z_t(X_t = \varphi_t \mid X_{t_0} = \varphi_{t_0}) = 1 \quad \text{if } t \geq t_0$$

so

$$\begin{aligned} Z_t(X_t \neq \varphi_t) &= 1 - Z_t(X_t = \varphi_t) \\ &\leq 1 - \left[Z_t(X_t = \varphi_t \mid X_{t_0} = \varphi_{t_0})\, Z_{t_0}(X_{t_0} = \varphi_{t_0}) \right] \\ &= 1 - Z_{t_0}(X_{t_0} = \varphi_{t_0}) \\ &= Z_{t_0}(X_{t_0} \neq \varphi_{t_0}) \quad \text{for } t \geq t_0 \end{aligned}$$

and by applying the coupling lemma, we have that

$$\begin{aligned} d_{TV}(\pi_0 P^t, \pi) &\leq Z_t(X_t \neq \varphi_t) \\ &\leq Z_{t_0}(X_{t_0} \neq \varphi_{t_0}) \quad \text{for } t \geq t_0 \end{aligned}$$

Hence if we find τ such that $Z_\tau(X_\tau \neq \varphi_\tau) \leq \varepsilon$, then

$$d_{TV}(\pi_0 P^t, \pi) \leq \varepsilon \quad \text{for } t \geq \tau$$

i.e. $t_{mix}(\varepsilon) \leq \tau$

So the coupling method is:

1) construct a coupling b/w the MC started in the arb state & the MC started at its stat dist, $(X_t, Y_t) \sim Z_t$, s.t. once $X_{t_0} = Y_{t_0}$, $X_t = Y_t$ for $t \geq t_0$

2) Find a τ s.t. $Z_\tau(X_\tau \neq Y_\tau) < \varepsilon$, then we have

$$d_{TV}(\pi_0 P^t, \pi) \leq \varepsilon \quad \text{for } t \geq \tau,$$

so

$$t_{mix}(\varepsilon) \leq \tau$$

Ex:

- Lazy Random Walk on the Ring $\mathbb{Z}_m$
- Random Sampling from the Hypercube

Example of coupling method: lazy: stay w/ $\frac{1}{2}$, up or down w/ $\frac{1}{4}$

Consider MC on the ring $\mathbb{Z}_{n+2}$ with stationary dist given by unif dist. What is an upper bound on $t_{mix}(\varepsilon)$? e.g. find a T s.t. $t > T \Rightarrow$

$$\max_A \left| \mathbb{P}(X_t \in A) - \sum_{i \in A} \pi_i \right| < \varepsilon$$

Idea: construct a coupling (X_t, Y_t) where X_t started at whatever initial dist and Y_t started at stationary distribution and evolve the coupling in a way s.t. its marginals are as described and

$$\mathbb{P}(X_t \neq Y_t)$$

if $X_t \neq Y_t$.

Idea: at time $t+1$ pick either X_t or Y_t w/ equal prob, and increase or decrease by 1 w/ prob $\frac{1}{2}$. If $X_t = Y_t$, then move processes together

Clear that marginals are as described, e.g.

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \begin{cases} 1/2 & \text{if } j = i \\ 1/4 & \text{if } j = i \pm 1 \end{cases}$$

Note that $\Delta_t = X_t - Y_t$ is by construction a random walk on $\mathbb{Z}_m$

and

$$\Delta_{t+1} \mid \Delta_t = \begin{cases} \Delta_t - 1 & \text{w.p. } \frac{1}{2} \\ \Delta_t + 1 & \text{w.p. } \frac{1}{2} \end{cases} \quad \text{if } \Delta_t \neq 0$$

with absorbing barrier at 0:

$$\mathbb{P}(\Delta_{t+1} = 0 \mid \Delta_t = 0) = 1$$

Key point is that this random walk on the ring converges to 0 in expected time less than $m^2/4$ from any starting state X_0

Thm Let $\Delta_{t+1} = \begin{cases} \Delta_t \mp 1 \text{ w.p. } 1/2, \Delta_t \neq 0 \\ 0 \text{ if } \Delta_t = 0 \end{cases}$ be

a random walk on $\mathbb{Z}_m$ with an absorbing barrier at 0, and

$$\tau = \min\{t : \Delta_t = 0\},$$

then for any starting distribution π,

$$\mathbb{E}[\tau] \leq \frac{m^2}{4}$$

Pf

By total expectation,

$$\mathbb{E}[\tau] = \sum_{k=0}^{m-1} \mathbb{E}[\tau | X_0 = k]\, \mathbb{P}(X_0 = k)$$

$$\leq \max_{k=0,\dots,m-1} \mathbb{E}[\tau | X_0 = k]$$

Write $f_k = \mathbb{E}[\tau | \Delta_0 = k]$ and note

$$\mathbb{E}[\tau | \Delta_0 = k] = \frac{1}{2}\left(1 + \mathbb{E}[\tau | \Delta_0 = k-1]\right)$$

$$+ \frac{1}{2}\left(1 + \mathbb{E}[\tau | \Delta_0 = k+1]\right) \qquad \text{when } k \neq 0$$

$$\text{so} \quad f_k = 1 + \frac{1}{2}\left(f_{k-1} + f_{k+1}\right)$$

$$\text{and} \quad f_0 = 0$$

$$\Rightarrow \quad f_k = 2\left(f_{k-1} - 1\right) - f_{k-2}$$

$$\Rightarrow \quad f_k - f_{k-1} = f_{k-1} - 2 - f_{k-2}$$

let $d_k = f_k - f_{k-1}$

$$\Rightarrow \quad d_k = d_{k-1} - 2 \qquad \Rightarrow \quad d_k = f_1 - 2(k-1)$$

$$d_1 = f_1 - f_0 = f_1$$

$$\Rightarrow f_k = d_k + f_{k-1} = d_k + d_{k-1} + f_{k-2}$$

$$= \cdots = \sum_{t=1}^{k} d_t = \sum_{t=1}^{k} (f_1 - 2(t-1))$$

$$= k f_1 - 2\sum_{t=1}^{k}(t-1) = k f_1 - 2\sum_{t=0}^{k-1} t$$

$$= k f_1 - k(k-1)$$

Note $f_1 = f_{m-1}$ by symmetry

$$\Rightarrow \quad f_1 = (m-1) f_1 - (m-1)(m-2)$$

$$\Rightarrow (m-1)(m-2) = (m-2) f_1$$

$$\Rightarrow \quad m-1 = f_1$$

$$\Rightarrow f_k = k(m-1) - k(k-1)$$

$$= km - k^2$$

so $\max_{k \in [m]} \mathbb{E}[\tau \mid \Delta_0 = k] = \max_{k \in [m]} km - k^2$

occurs at $k = \frac{m}{2}$ (if m is even, otherwise this is an upper bound)

$$\Rightarrow \mathbb{E}[\tau] \leq m\left(\frac{m}{2}\right) - \left(\frac{m}{2}\right)^2 = \frac{m^2}{4}$$

Now since $Z_t = X_t - Y_t$ is a random walk on $\mathbb{Z}_m$ w/ absorbing state at 0,

$$Z_t \neq 0 \Rightarrow \tau > t$$

Let $T = \frac{m^2}{4}$, then if $t > 2T$,

$$\mathbb{P}(Z_t \neq 0) \leq \mathbb{P}(\tau > 2T) \leq \frac{\mathbb{E}\tau}{\left(\frac{m^2}{2}\right)} \leq \frac{2}{m^2}\frac{m^2}{4} = \frac{1}{2}$$

so the chain is likely to have coupled after $\tau = \frac{m^2}{2}$ steps w/ prob greater than $\frac{1}{2}$

Now consider $t > 2cT$

$$\mathbb{P}(Z_t \neq 0) \leq \mathbb{P}(Z_{2cT} \neq 0)$$

$$= \prod_{i=0}^{c-1} \mathbb{P}\left(Z_{2(c-i)T} \neq 0 \mid Z_{2(c-i-1)T} \neq 0\right) \mathbb{P}(Z_0 \neq 0)$$

$$\leq \left(\frac{1}{2}\right)^c$$

so the chain has coupled after $\tau = 2cT$ steps w/ prob at least $1 - \frac{1}{2^c}$

consequently

$$d_{TV}(X_t, Y_t) \leq \mathbb{P}(X_t \neq Y_t) \leq \varepsilon$$

if

$$t \geq \log_2\left(\frac{1}{\varepsilon}\right)\frac{m^2}{2}$$

or

$$\| \pi_0 P^t - \pi \|_1 \leq 2\varepsilon \quad \text{when} \quad t \geq \log_2\left(\frac{1}{\varepsilon}\right)\frac{m^2}{2}$$

π_0: arbitrary starting dist; π: stat dist

Note: if $X_0 \sim$ unif on the ring, can get $t \geq \log_2\left(\frac{1}{\varepsilon}\right)\frac{m^2}{3}$, b/c can show $\mathbb{E}\tau \leq \frac{m^2}{6}$ (use total probability)

Another example of coupling: random walk on hypercube

Start w/ a bit vector of length n and at each time step, flip a random bit w.p. $1/2$

$X_t \in \{0,1\}^n$

$Y_t \in \{0,1\}^n$

$$P_{(i_1,\ldots,i_n),(j_1,\ldots,j_n)} = \begin{cases} 0 & \text{if different in } >1 \text{ position} \\ \frac{1}{2n} & \text{otherwise} \end{cases}$$

Coupling, take $Z_t = (X_t, Y_t)$ where X_t has arb init pos and Y_t has stat dist, and we evolve Z_t by picking exactly the same random coordinate in X_t, Y_t at each time step and flipping to exactly the same value.

Clearly that marginal probs are both MCs given by P, and

$$X_t = Y_t$$

as soon as we have touched all n bits

Recall from coupon collector that if

τ = time to draw all n unique coupons

then

$$\mathbb{E}\tau = nH_n$$

so $\mathbb{P}(X_t \neq Y_t) \leq \mathbb{P}(X_{2nH_n} \neq Y_{2nH_n}) \leq \frac{1}{2}$

for $t \geq 2nH_n$

$\Rightarrow$ as before, $d_{TV}(X_t, \pi) \leq \varepsilon$ when $t \geq O(\log_2(\frac{1}{\varepsilon}) nH_n)$

note that there are 2^n states, so have <u>fast mixing</u>: $t_{mix}(\varepsilon) \approx O(n \log n)$ is logarithmic in # states!

Final Markov Material

Ex 2-SAT on n literals (analyze as MC)

Consider alg: 0. randomly assign vals to literals
1. randomly pick an unsatisfied clause
2. pick one of its literals, randomly set to T/F
3. repeat until all clauses satisfied

Q: how long does this run?

Think of as MC w/ absorbing states corresponding to satisfying assignments on $\{0,1\}^n$

Fix a satisfying assignment A and ask how long we expect alg to run before we reach this satisfying state.

Let X = # correct literals at time

$$X_{t+1} = \begin{cases} X_t - 1 & \text{w.p. } \frac{1}{2} \\ X_t + 1 & \text{w.p. } \frac{1}{2} \\ n & \text{if } X_t = n \end{cases} \quad \text{if } X_t \neq n$$

and letting $\tau = \min\{t : X_t = n\}$,

$$\mathbb{E}[\tau \mid X_0 = k] = \frac{1}{2}\left(1 + \mathbb{E}[\tau \mid X_0 = k-1]\right) + \frac{1}{2}\left(1 + \mathbb{E}[\tau \mid X_0 = k+1]\right)$$

$$\Rightarrow \quad f_k := 1 + \frac{1}{2}(f_{k-1} + f_{k+1})$$

$\Rightarrow \tau \leq \frac{m^2}{4}$ as argued before for r.w. on the ring

$\Rightarrow$ this alg returns a satisfying assignment in $O(m^2)$ steps