Markov Chain Convergence

- Total variation distance & coupling
- MC convergence theorem
- Mixing times via
  - coupling
  - strong stationary times
- Markov chain Monte Carlo
- Perfect sampling
  - strong Doeblin cond
  - Coupling from the past

Thm (Convergence Thm for finite MC) Given finite MC P that is irreducible, aperiodic & positive recurrent, with stationary distribution Π. Then $\exists\, \alpha \in (0,1]$ and $C > 0$ s.t.

$$\max_{x \in \mathcal{X}} \| P^t(x,\cdot) - \Pi \|_{TV} \leq C\alpha^t$$

We first show this for finite state chains

Pf - Since P is irreducible, aperiodic $\Rightarrow \exists\, r \geq 1$ s.t. $P^r > 0$ (i.e. $P^r(x,y) > 0\ \forall x,y \in \mathcal{X}$). Thus for some $\delta > 0$, we have

$$P^r(x,y) \geq \delta \pi(y) \ \forall x,y \in \mathcal{X}$$

- Let $\alpha = 1-\delta$. We can write $P^r = (1-\alpha)\mathbb{1}^T\pi + \alpha Q$, where Q is a stochastic matrix, and we denote $\Pi = \mathbb{1}^T\pi = \begin{pmatrix} \pi \\ \pi \\ \pi \\ \vdots \end{pmatrix}$

- Now we claim: $\forall k \geq 1$, $P^{rk} = (1-\alpha^k)\Pi + \alpha^k Q^k$

Can show this by induction - suppose true for k, then

$$\begin{aligned} P^{r(k+1)} &= P^{rk}P^r = ((1-\alpha^k)\Pi + \alpha^k Q^k)((1-\alpha)\Pi + \alpha Q) \\ &= (1-\alpha^k)\Pi + \alpha^k(1-\alpha)\Pi + \alpha^{k+1}Q^{k+1} \\ &= (1-\alpha^{k+1})\Pi + \alpha^{k+1}Q^{k+1} \end{aligned}$$

Since $M\Pi = \Pi$ for any stochastic matrix M

- Thus $P^{rk+j} - \Pi = \alpha^k(Q^kP^j - \Pi)$

Now for any $x \in \mathcal{X}$, we have that

$$\| P^{rk+j}(x_0,\cdot) - \pi \|_{TV} \leq \alpha^k$$

This follows from the fact that $\|\mu - \nu\|_{TV} = \|\mu - \nu\|_1/2$, and since $\|Q^kP^j(x,\cdot) - \pi\|_1 \leq 2$ for stochastic matrices Q, P ▥

- The total variation distance between any two probability distributions μ and ν on Ω is defined as $\|\mu-\nu\|_{TV} = \sup_{A\subset\Omega} |\mu(A)-\nu(A)|$

- Properties of $\|\mu-\nu\|_{TV}$ (or $d_{TV}(\mu,\nu)$)

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

ii) $\|\mu-\nu\|_{TV} = \sum_{x\in\Omega \mid \mu(x)\geq\nu(x)} (\mu(x)-\nu(x))$

test function

iii) $\|\mu-\nu\|_{TV} = \frac{1}{2}\sup\left\{\sum_{x\in\Omega} f(x)\mu(x) - \sum_{x\in\Omega} f(x)\nu(x) \;\middle|\; \|f\|_\infty \leq 1\right\}$

- A coupling of 2 probability distributions F and G is a pair of r.v (X,Y) defined on the same probability space $(\Omega,\mathcal{F},\mathbb{P})$ s.t.
$\forall x,y,\ \mathbb{P}[X\leq x] = F(x),\ \mathbb{P}(Y\leq y) = G(y)$

– In other words, we 'design' X and Y so that they have the correct marginal distr (but can have any joint distr)

Eg - $F \sim \mathrm{Ber}(p)$, $G \sim \mathrm{Ber}(p/2)$

Coupling 1 - $X \sim \mathrm{Ber}(p)$, $Y \sim \mathrm{Ber}(p/2)$ independently

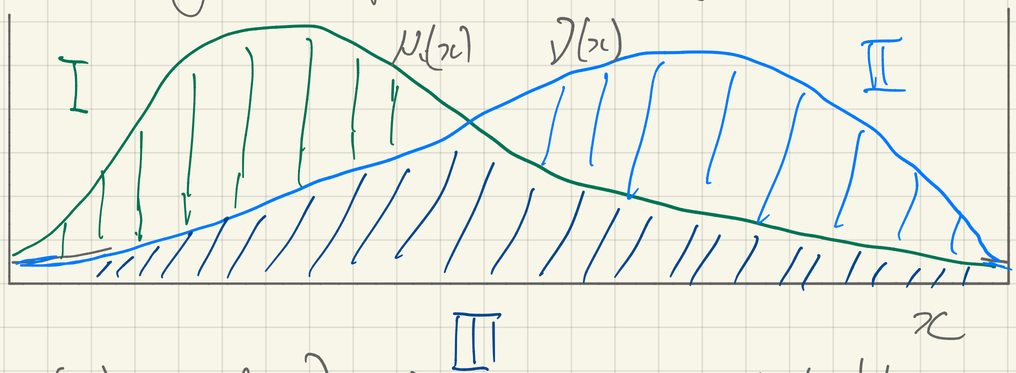
Coupling 2 - $X \sim \mathrm{Ber}(p)$; If $X = 0$ then $Y = 0$, and if $X = 1$, then $Y \sim \mathrm{Ber}(1/2)$

$$\left(\text{ie, } (X,Y) = \begin{cases} (1,1) & \text{wp } p/2 \\ (1,0) & \text{wp } p/2 \\ (0,0) & \text{wp } 1-p \end{cases} \right)$$

Lemma (the 'optimal' coupling) Let μ and ν be two probability distributions on Ω. Then

$$\|\mu - \nu\|_{TV} = \inf \{ \mathbb{P}[X \neq Y] \mid (X,Y) \text{ is a coupling of } \mu \text{ and } \nu \}$$

Pf - Intuitively the proof is as follows



Given $\mu(x)$ and $\nu(x)$, we can plot them as above

Now by defn, the area of regions I and II are $\|\mu-\nu\|_{TV}$ and area of III is $1 - \|\mu-\nu\|_{TV}$. Now we sample from III w.p. $1 - \|\mu-\nu\|_{TV}$ and set $X = Y$, else sample X from I and Y from II

- Formally, for arbitrary $A \subset \Omega$, and $X \sim \mu$, $Y \sim \nu$

$$\mathbb{P}[X \neq Y] \geq \mathbb{P}[X \in A, Y \in \bar{A}] = \mathbb{P}[X \in A] - \mathbb{P}[X \in A, Y \in A]$$
$$\geq \mathbb{P}[X \in A] - \mathbb{P}[Y \in A]$$
$$= \mu(A) - \nu(A)$$

Thus $\|\mu - \nu\|_{TV} \leq \inf \{\mathbb{P}[X \neq Y] \mid (X, Y) \text{ coupling of } \mu, \nu\}$

- To show equality, we construct coupling (X, Y) as follows – let $U \in \{0, 1\}$ and $Z, V, W \in \Omega$ be independent r.v. defined as

$$U = \text{Ber}(1 - \|\mu - \nu\|_{TV})$$
$$Z \equiv \mathbb{P}[Z = i] = (\mu(i) \wedge \nu(i)) / (1 - \|\mu - \nu\|_{TV})$$
$$V \equiv \mathbb{P}[V = i] = (\mu(i) - \nu(i))^+ / \|\mu - \nu\|_{TV}$$
$$W \equiv \mathbb{P}[W = i] = (\nu(i) - \mu(i))^+ / \|\mu - \nu\|_{TV}$$

- Now let $(X, Y) = \begin{cases} (Z, Z) & \text{if } U = 1 \\ (V, W) & \text{if } U = 0 \end{cases}$

Note that $V \neq W \Rightarrow \mathbb{P}[X \neq Y] = \|\mu - \nu\|_{TV}$, and

$$\mathbb{P}[X = i] = \frac{(\mu(i) \wedge \nu(i))}{(1 - \|\mu - \nu\|_{TV})} \cdot (1 - \|\mu - \nu\|_{TV}) + \|\mu - \nu\|_{TV} \left(\frac{(\mu(i) - \nu(i))^+}{\|\mu - \nu\|_{TV}} \right) = \mu(i)$$

Similarly $\mathbb{P}[Y = i] = \nu(i)$ ∎

$\boxed{\text{Thm}}$ (Convergence Thm for countable MC) Given P on countable $\mathcal{X}$ that is irreducible, aperiodic & positive recurrent, with stationary distribution Π. Then $\forall x \in \mathcal{X}$

$$\lim_{t \to \infty} \| P^t(x,\cdot) - \Pi \|_{TV} = 0$$

Pf - Consider the 2-dimensional MC (X_t, Y_t) defined as $\tilde{P}\left[(X_{t+1}, Y_{t+1}) = (i', j') \mid (X_t, Y_t) = (i,j)\right] = P(i,i')P(j,j') \; \forall i, i', j, j'$

- Can view this as 2 particles starting at $X_0 = x_0$, $X_1 = x_1$ and evolving independently
- Claim - P irreducible, aperiodic $\Rightarrow$ the joint chain (X_t, Y_t) with transition $\tilde{P}$ is also irreducible and aperiodic (check!)
- Also $\tilde{\Pi}(i,j) = \Pi(i)\Pi(j)$ is a stationary distribution for $\tilde{P}$!

$\Rightarrow$ $\boxed{\tilde{P} \text{ is an ergodic MC with stat dist } \tilde{\Pi}}$

- Now let $\tau_c = \inf\{t \geq 0 \mid X_t = Y_t\}$ (Coupling time)

Claim - for any μ, ν, we have $\mathbb{P}_{x,y}[\tau_c < \infty] = 1$

This follows from the fact $\tilde{P}$ is ergodic, and since τ_c is the hitting time of set $A = \{(x,x) \mid x \in \mathcal{X}\}$

- Next, for any pair of dist (μ, ν), we can construct coupling (X_t, Y_t), where $X_0 \sim \mu$, $Y_0 \sim \nu$, and the chains evolve as $(\forall (i,j) \in \mathcal{X}^2, (i',j') \in \mathcal{X}^2)$

$$\mathbb{P}\left[(X_{t+1}, Y_{t+1}) = (i', j') \mid (X_t, Y_t) = (i, j)\right] = \begin{cases} P(i,i')P(j,j') ; & i \neq j \\ P(i,i')\mathbb{1}\{i'=j'\} ; & i = j \end{cases}$$

- From above, we know coupling time τ_c is a.s. finite

- Now using this claim, we can prove the theorem as follows. Recall P ergodic $\Rightarrow$ unique stationary distribution π. Now let $\nu = \pi$, and $\mu = \delta_x$ (i.e., $X_0 = x$). Then we have

$$\| P^t(x, \cdot) - \pi \|_{TV} \leq \mathbb{P}_{\delta_x, \pi}[X_t \neq Y_t]$$
$$= \mathbb{P}_{\delta_x, \pi}[\tau_c > t]$$

and $\lim_{t \to \infty} \mathbb{P}_{\delta_x \times \pi}[\tau_c > t] = \sum_{y \in \mathcal{X}} \pi(y) \lim_{t \to \infty} \mathbb{P}_{x,y}[\tau_c > t] = 0$

- Thus $\lim_{t \to \infty} \| P^t(x, \cdot) - \pi \|_{TV} = 0 \quad \forall x \in \mathcal{X}$

$\blacksquare$

As a result, π is often referred to as the equilibrium dist

Mixing Times

We know use the above idea to understand how 'fast' $\pi_0^T P^t$ converges to π^T.

Def - For ergodic MC P with stationary dist^n π, we define $\forall\ t \geq 0$

$$d(t) = \max_{x \in \mathcal{X}} \| P^t(x,\cdot) - \pi \|_{TV}$$

$$\bar{d}(t) = \max_{x,y \in \mathcal{X}} \| P^t(x,\cdot) - P^t(y,\cdot) \|_{TV}$$

(distance from stationarity)

- **Lemma** - $d(t) \leq \bar{d}(t) \leq 2d(t)$

Pf - For the upper bound, since $\| \ \|_{TV}$ is a norm, we have by Δ inequality -

$$\max_{x,y} \| P^t(x,\cdot) - P^t(y,\cdot) \|_{TV} \leq \max_{x,y} \left(\| P^t(x,\cdot) - \pi \|_{TV} + \| P^t(y,\cdot) - \pi \|_{TV} \right)$$

$$\leq 2 \max_{x} \| P^t(x,\cdot) - \pi \|_{TV} = 2d(t)$$

- For the lower bound, note that $\forall A \subseteq \mathcal{X}$, $\pi(A) = \sum_{x \in \mathcal{X}} \pi(x) P^t(x,A)$

$$\Rightarrow \| P^t(x,\cdot) - \pi \|_{TV} = \sup_{A \subseteq \Omega} | P^t(x,A) - \pi(A) |$$

$$= \sup_{A \subseteq \Omega} \left| \sum_y \pi(y) \left(P^t(x,A) - P^t(y,A) \right) \right|$$

$$\leq \sup_{A \subseteq \Omega} \sum_y \pi(y) | P^t(x,A) - P^t(y,A) | = \bar{d}(t)$$

- <u>Lemma</u> – $\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t) \ \forall s,t \geq 0$

Pf – Fix $x,y \in \mathcal{X}$, and for any $s \geq 0$, let (X_s, Y_s) be the optimal coupling of $P^s(x,\cdot)$ and $P^t(y,\cdot)$, ie, $X_s \sim P^s(x,\cdot)$, $Y_s \sim P^s(y,\cdot)$, $\|P^s(x,\cdot) - P^s(y,\cdot)\|_{TV} = \mathbb{P}[X_s \neq Y_s]$

- Now $P^{s+t}(x,w) = \sum_z \mathbb{P}[X_s = z] P^t(z,w) = \mathbb{E}[P^t(X_s,w)]$

Similarly $P^{s+t}(y,w) = \mathbb{E}[P^t(Y_s,w)]$

$\Rightarrow P^{s+t}(x,w) - P^{s+t}(y,w) = \mathbb{E}[P^t(X_s,w) - P^t(Y_s,w)]$

(<u>Note</u> - we can do this as X_s, Y_s are on the same $(\Omega, \mathcal{F}, \mathbb{P})$)

$$\Rightarrow \|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\|_{TV} = \frac{1}{2}\sum_w |P^{s+t}(x,w) - P^{s+t}(y,w)|$$
$$= \mathbb{E}\left[\frac{1}{2}\sum_w |P^t(X_s,w) - P^t(Y_s,w)|\right]$$
$$= \mathbb{E}\left[\|P^t(X_s,\cdot) - P^t(Y_s,\cdot)\|_{TV}\right]$$

- However $\because P^t(X_s,\cdot) = P^t(Y_s,\cdot)$ when $X_s = Y_s$, we have

$$\|P^{s+t}(x,\cdot) - P^{s+t}(y,\cdot)\|_{TV} \leq \bar{d}(t)\,\mathbb{E}[\mathbb{1}\{X_s \neq Y_s\}]$$
$$= \bar{d}(t)\bar{d}(s) \quad \blacksquare$$

Thus, $\bar{d}(t)$ is Submultiplicative $\Rightarrow d(ct) \leq \bar{d}(ct) \leq \bar{d}(t)^c$

- **Def (Mixing Time)** - For any P, and $\varepsilon > 0$
  - $t_{mix}(\varepsilon) = \inf\{t \geq 0 \mid d(t) \leq \varepsilon\}$
  - $t_{mix} = t_{mix}(1/4)$

- Why $1/4$? Using previous lemma we have

$$d(\ell t_{mix}) \leq \bar{d}(\ell t_{mix}) \leq (\bar{d}(t_{mix}))^{\ell} \leq (2d(t_{mix}))^{\ell} = 2^{-\ell}$$

$$\Rightarrow \quad t_{mix}(\varepsilon) \leq \lceil \log_2(1/\varepsilon) \rceil \, t_{mix}$$

- We now want to understand how t_{mix} behaves

Thm (Mixing from coupling) Let (X_t, Y_t) be a coupling s.t. $X_0 = x$, $Y_0 = y$, and if $X_s = Y_s$ then $X_t = Y_t \; \forall \, t \geq s$.

Define $\tau_{couple} = \inf\{t \geq 0 \mid X_t = Y_t\}$.

Then we have $\bar{d}(t) \leq \mathbb{P}_{x,y}[\tau_{couple} > t]$

Pf - $P^t(x,z) = \mathbb{P}_{x,y}[X_t = z]$, $P^t(y,z) = \mathbb{P}_{x,y}[Y_t = z]$

$\Rightarrow$ for coupling (X_t, Y_t), $\|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \leq \mathbb{P}_{x,y}[X_t \neq Y_t]$

Also $\mathbb{P}[X_t \neq Y_t] = \mathbb{P}[\tau_{couple} > t]$, and we are done! $\blacksquare$

Eg - Lazy walk on circle - $\mathcal{X} = \{0, 1, \dots, n-1\}$

$$Z_t = \begin{cases} -1 & \text{wp } 1/4 \\ 0 & \text{wp } 1/2 \\ 1 & \text{wp } 1/4 \end{cases}, \quad X_{t+1} = (X_t + Z_{t+1}) \bmod n$$

- Check that $\pi = (1/n \ 1/n \dots 1/n)^T$
- Coupling $(X_t, Y_t) \equiv X_0 = x, Y_0 = y$
  - At time t, let $W_t = \text{Ber}(1/2)$.
    - If $W_t = 1 \Rightarrow X_t = X_{t-1} \mp 1$ wp $1/2$
    - $W_t = 0 \Rightarrow Y_t = Y_{t-1} \mp 1$ wp $1/2$
  - If $X_s = Y_s$ then $X_t = Y_t \ \forall s \geq t$
- Let D_t = clockwise distance betn particles at t,
  - $\tau = \min\{t \geq 0 \mid D_t \in \{0, n\}\} = \tau_{couple}$
  - $\mathbb{E}_{x,y}[\tau] = \underbrace{k(n-k)}$, where k = clockwise dist betn x, y

from gambler's ruin - Solves $\mathbb{E}[\tau|k] = 1 + \frac{1}{2}(\mathbb{E}[\tau|k+1] + \mathbb{E}[\tau|k-1])$

$$\Rightarrow \bar{d}(t) \leq \max_{x,y} \mathbb{P}_{x,y}[\tau_{couple} > t] \leq \frac{\max_{x,y} \mathbb{E}_{x,y}[\tau]}{t}$$

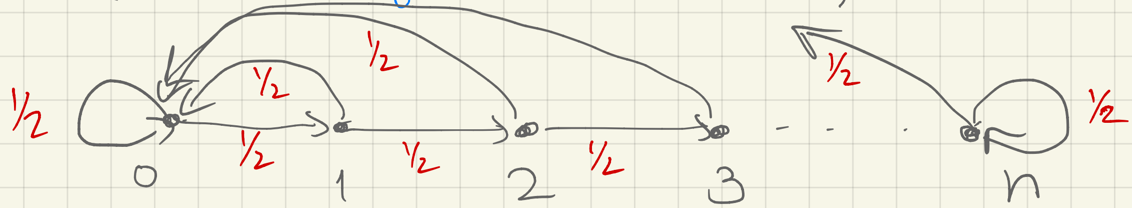
$$\leq \max_{k \in \{0, \dots, n\}} \frac{k(n-k)}{t} = \frac{n^2}{4t}$$

Thus if $t > n^2 \Rightarrow d(t) \leq \bar{d}(t) \leq 1/4$

$\Rightarrow t_{mix} \leq n^2$

KK

Eg- (the 'winning streak' chain)



$$X_{t+1} = \begin{cases} \min(X_t+1, n) & \text{wp } 1/2 \\ 0 & \text{wp } 1/2 \end{cases}$$

- Let $X_0 = x$, $Y_0 = y$, $Z_t \sim Ber(1/2)$

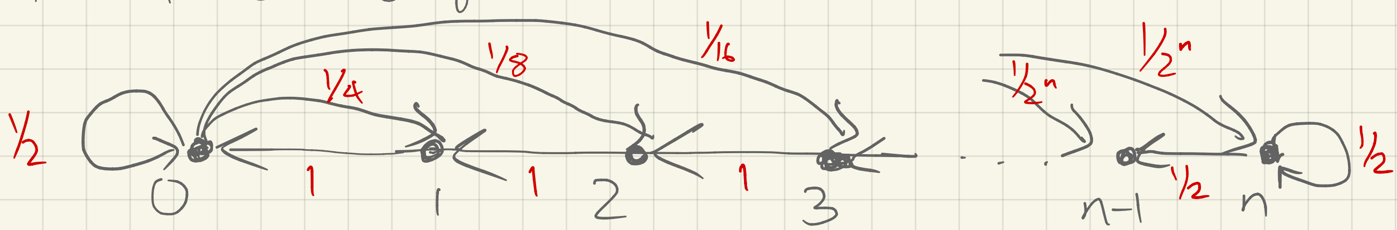
If $Z_t = 1$, then $X_t = (X_{t-1}+1)\wedge n$, $Y_t = (Y_{t-1}+1)\wedge n$

If $Z_t = 0$, then $X_t = Y_t = 0$!

$\Rightarrow \mathbb{P}[\tau_{couple} > t] \le 2^{-t} \Rightarrow \boxed{t_{mix} \le 2 \ \forall n}$!

- Note that $\pi(i) = 1/2^{i+1}$ if $i \le n-1$, $1/2^n$ if $i = n$

- Next consider the reverse chain $\hat{P}$



- $\pi(i) = 2^{-((i+1)\wedge n)} \Rightarrow \hat{P}(i,i-1) = 1$, $\hat{P}(n,n) = \hat{P}(n,n-1) = 1/2$ and $\hat{P}(0,i) = \pi(i) \ \forall 0 \le i \le n$

- If $X_t = 0 \Rightarrow X_{t+1} \sim \pi$! Now suppose $X_0 = n \Rightarrow$ MC spends $K \sim Geom(1/2)$ turns at n, then reaches 0 in $n-1$ turns, then mixes.

Also $\hat{P}^n(n,\cdot) = \pi \Rightarrow \pi_t = \pi \ \forall t \ge n$!

Finally $\hat{P}^{n-1}(n,1) = 1/2 \Rightarrow d(n-1) \ge \|\hat{P}^{n-1}(n,\cdot) - \pi\|_{TV} \ge 1/2 \Rightarrow \boxed{t_{mix} = n}$

Eg - Random walk on the hypercube - $\mathcal{X} = \{0,1\}^n$ Hamming distance of 1

- $P(x,y) = 1/n$ if x and y differ in 1 bit, else 0
- Lazy RW on hypercube - $P(x,x) = 1/2$, $P(x,y) = 1/2n$ if $d_H(x,y) = 1$
- $\pi(x) = 1/2^n$ ($\because |\mathcal{X}| = 2^n$, by symmetry / doubly stoch)
- Alternate description for LRW - Pick index $i \in [n]$ uar, flip bit $X(i)$ wp $1/2$
- Coupling $(X_t, Y_t) \equiv$ Given (X_t, Y_t)
  - i) Pick index $I_{t+1} \in [n]$ uar, and $Z_{t+1} = \{0,1\}$ wp $1/2$
  - ii) Set $X_{t+1}(I_t) = Y_{t+1}(I_t) = Z_{t+1}$
- Note - $d_H(X_0, Y_0) \leq n \Rightarrow \tau_{couple} \leq$ 'coupon collector' on n bins

$\Rightarrow \mathbb{E}[\tau_{couple}] = n H_n \Rightarrow \bar{d}(t) \leq \mathbb{E}[\tau_{couple}]/t \leq n H_n / t$

$\Rightarrow t_{mix} \leq 4 n H_n$ (can be improved to $\frac{1}{2} n \ln n$)

(Grand Coupling) - For any MC $X_{t+1} = f(X_t, U_{t+1})$, we can always construct a coupling $(X_{t+1}, Y_{t+1}) = (f(X_t, U_{t+1}), f(Y_t, U_{t+1}))$

- This is a generic way to construct a coupling for any chain! It gives a standard way to measure τ_c

$$\tau_c = \min_{t \geq 0} \{ f(\cdots f(f(X_0, U_1), U_2) \cdots, U_t) = f(\cdots f(f(Y_0, U_1), U_2, \cdots, U_t) \}$$

- Note that this is not always easy to compute. However this gives us another approach to computing mixing times

Strong Stationary Times

- Suppose MC has representation $X_{t+1} = f(X_t, Z_{t+1})$ for some iid sequence Z_t

Def - (Randomized Stopping Time) - A random time τ for MC X_t is a randomized stopping time if its a stopping time for Z_t

Def - (Stationary Time) For MC X_t with stationary dist π, a stationary time τ is a randomized stopping time (possibly dependent on x) s.t $\mathbb{P}_x[X_\tau = y] = \pi(y)$

- Thus, a stationary time is in a sense a signal that a chain has mixed. However, to bound mixing times, we need a slightly stronger definition.

Def (Strong stationary time) - A strong stationary time for MC X_t is a randomized stopping time, possibly dependent on starting state x s.t $\forall y,t: \mathbb{P}_x[\tau = t, X_\tau = y] = \mathbb{P}_x[\tau = t]\,\pi(y)$

Why the stronger defn? Consider a rw on n-cycle, and the following τ: wp $1/n$, set $\tau = 0$, else set $\tau = \inf\{t \geq 0 \mid$ every state $x \in \mathcal{X}$ visited once$\}$. In the latter case, the terminal state is uniform over $\mathcal{X} \setminus x_0 \Rightarrow \mathbb{P}[X_\tau = x] = 1/n \;\forall x \in \mathcal{X}$! However, $\tau \not\perp X_\tau$, as $\tau = 0 \Rightarrow X_\tau = x_0$!

- **Lemma** - For X_t ergodic MC with stationary dist π, if τ is a strong stationary time, then $\forall\, t \geq 0$

$$\mathbb{P}_x[\tau \leq t, X_t = y] = \mathbb{P}_x[\tau \leq t]\,\pi(y)$$

Pf - Let $Z_1, Z_2, \ldots$ be the sequence in the random mapping.

For any $s \leq t$

$$\mathbb{P}_x[\tau = s, X_t = y] = \sum_{z \in \mathcal{X}} \mathbb{P}_x[X_t = y \mid \tau = s, X_s = z]\,\mathbb{P}_x[\tau = s, X_s = z]$$

$\because$ τ is a stopping time for Z_t, the event $\{\tau = s\}$ is adapted to $\sigma(Z_1, Z_2, \ldots, Z_s)$, and $X_{s+r} = f_r(X_s, Z_{s+1}, \ldots, Z_{s+r})$ for some fn f_r. Also since $(Z_1, \ldots, Z_s) \perp\!\!\!\perp (Z_{s+1}, \ldots, Z_{s+r}) \Rightarrow$

$$\mathbb{P}_x[X_t = y \mid \tau = s, X_s = z] = \mathbb{P}_x[f_{t-s}(z, Z_{s+1}, \ldots, Z_t) = y \mid X_s = z, \sigma(Z_1, \ldots, Z_s)]$$
$$= P^{t-s}(z, y)$$

$\Rightarrow$ by definition of strong stopping times

$$\mathbb{P}_x[X_t = y, \tau = s] = \sum_{z \in \mathcal{X}} P^{t-s}(z, y)\,\mathbb{P}_x[\tau = s]\,\pi(z)$$

Also since $\pi^T P = \pi^T$, we have that the RHS is $\pi(y)\,\mathbb{P}_x[\tau = s]$. Now summing over all $s \leq t$, we get the result $\blacksquare$

- Now to use strong stationary times to bound t_{mix}, we need an additional definition.

Defn (Separation Distance) - For any x, t, given MC P, define

$$s_x(t) = \max_{y \in \mathcal{X}} \left[1 - \frac{P^t(x,y)}{\pi(y)} \right], \quad s(t) = \max_{x \in \mathcal{X}} s_x(t)$$

Lemma - $\| P^t(x,\cdot) - \pi \|_{TV} \leq s_x(t)$

Pf - $$\| P^t(x,\cdot) - \pi \|_{TV} = \sum_{\substack{y \in \mathcal{X} \mid \\ P^t(x,y) < \pi(y)}} \left(\pi(y) - P^t(x,y) \right)$$

$$= \sum_{\substack{y \mid \pi(y) \\ > P^t(x,y)}} \left(1 - \frac{P^t(x,y)}{\pi(y)} \right) \pi(y)$$

$\sum x_i y_i \leq \|x\|_1 \|y\|_\infty$ Hölder's Ineq

$$\leq \sum_{y \in \mathcal{X}} \pi(y) \max_y \left(1 - \frac{P^t(x,y)}{\pi(y)} \right) \leq s_x(t)$$

Lemma - If τ is a strong stationary time, then

$$s_x(t) \leq \mathbb{P}_x[\tau > t]$$

Pf - For any $x \in \mathcal{X}$, $(\forall y)$

$$1 - \frac{P^t(x,y)}{\pi(y)} = 1 - \frac{\mathbb{P}_x[X_t = y]}{\pi(y)} \leq 1 - \frac{\mathbb{P}_x[X_t = y, \tau \leq t]}{\pi(y)}$$

by the earlier lemma, $\mathbb{P}_x[X_t = y, \tau \leq t] = \mathbb{P}[\tau \leq t]\, \pi(y)$ ∎

Combining we get $\forall x \in \mathcal{X}$ and strong stationary time τ

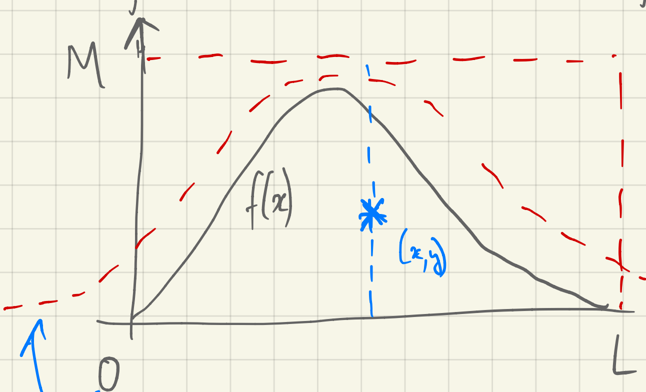
$$\| P^t(x,\cdot) - \pi \|_{TV} \leq \mathbb{P}_x[\tau > t]$$

Markov chains & Sampling

One of the most important uses of MCs is to sample from complex distributions.

- If we want $X \sim F$ for some known F on $\mathbb{R}$, then given $U \sim \text{Unif}[0,1]$, can set $X = F^{-1}(U)$ (inversion method)

- If we know pdf f for X, and can bound it in a box (Eg, in $[0,L]\times[0,M]$; see figure), then we can generate (x,y) uniformly in the box, and set $X = x$ if $y \leq f(x)$, else sample another (x,y)... (rejection sampling)



(importance sampling)

- Suppose now we want to sample something more complex:
  - Given $G(V,E)$, sample a spanning tree uniformly at random
  - Sample $x \in \mathcal{X}$ for some complex $\mathcal{X}$, proportional to some $g(x)$ (i.e., wp $Z g(x)$, where $Z = \int_{\mathcal{X}} g(x)\,dx$ may be unknown)
  - Sample a random independent set X of $G(V,E)$, proportional to $\lambda^{|X|}$

- Markov-chain Monte Carlo (MCMC) is a set of generic techniques that let us do this!
  - Idea - Set up MC s.t. π = target dist^n. Run till mixing...

- ## MCMC recipe
  - π = target distribution
    P ≡ sampling algorithm (typically reversible)
    Q ≡ Candidate-generator matrix
    (Some given random walk on $\mathcal{X}$)

  - Given state $x \in \mathcal{X}$, generate candidate y wp $Q(x,y)$
  - Accept y wp $\alpha(x,y) = P(x,y)/Q(x,y)$, else stay at x
  - Run till 'MC has mixed sufficiently'.
- We want P s.t. $P(x,y)\pi(x) = P(y,x)\pi(y) \; \forall x,y$
  Q s.t. it is ergodic, easy to sample from.

Examples of P -

1) Metropolis Algorithm - $\alpha(x,y) = \min\left(1, \left(\frac{\pi(y)Q(y,x)}{\pi(x)\,Q(x,y)}\right)\right)$

2) Barker's Algorithm - $\alpha(x,y) = \dfrac{\pi(y)\,Q(y,x)}{\pi(y)Q(y,x)+\pi(x)Q(x,y)}$

3) Gibb's Algorithm - If we want $\pi(x(1), x(2), \ldots, x(d))$ on set $\mathcal{X} = \Lambda^d$, Λ countable, we first choose I u.a.r on $[d]$, and set $x(I) \longrightarrow y$ wp $\pi(y \mid x(1), \ldots, x(I-1), x(I+1), \ldots, x(d))$

Eg - Given $\mathcal{X}$, and 'energy function' (or inverse fitness fn) $h: \mathcal{X} \to \mathbb{R}$ we want $\pi(x) = e^{-h(x)}/Z$, where $Z = \sum_{x \in \mathcal{X}} e^{-h(x')} \equiv$ partition function

- Suppose Q has stationary dist $\pi(x) = 1/|\mathcal{X}|$ (For example, choose Q to be doubly stochastic)

- Metropolis : $\alpha(x,y) = \min\left[1, e^{-(h(y)-h(x))}\right]$

$$\Rightarrow \begin{array}{l} P(x,y)\,\pi(x) = Q(x,y) \min\left[1, e^{-(h(y)-h(x))}\right] \cdot e^{-h(x)} \\ P(y,x)\,\pi(y) = Q(y,x) \min\left[1, e^{-(h(x)-h(y))}\right] \cdot e^{-h(y)} \end{array}, \quad \frac{Q(x,y)}{|\mathcal{X}|} = \frac{Q(y,x)}{|\mathcal{X}|}$$

- Barker : $\alpha(x,y) = \dfrac{e^{-h(y)}}{e^{-h(x)} + e^{-h(y)}}$

$$\Rightarrow \begin{array}{l} P(x,y)\,\pi(x) = Q(x,y) \dfrac{e^{-h(y)}}{e^{-h(x)} + e^{-h(y)}} \cdot e^{-h(x)} \\ P(y,x)\,\pi(y) = Q(y,x) \dfrac{e^{-h(x)}}{e^{-h(x)} + e^{-h(y)}} \cdot e^{-h(y)} \end{array}, \quad \frac{Q(x,y)}{|\mathcal{X}|} = \frac{Q(y,x)}{|\mathcal{X}|}$$

- For any reversible Q with known stationary distr $\tilde{\pi}$ we can modify these to get desired π

$$\frac{Q(x,y)}{Q(y,x)} = \frac{\tilde{\pi}(y)}{\tilde{\pi}(x)}$$

Eg (Glauber dynamics / Gibbs sampler)

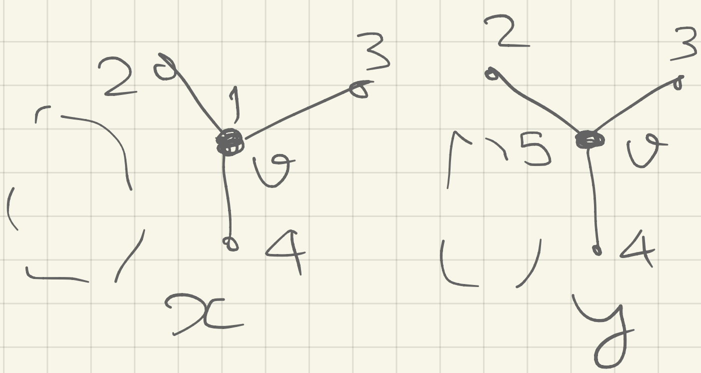
For undirected graph $G(V,E)$, a proper q-coloring is a function $x: V \to [q]$ (ie, element of $[q]^V$) s.t $x(v) \neq x(w)$ $\forall (v,w) \in E$. We also denote $N(v) = \{w \mid (v,w) \in E\}$ (neighborhood of v)

- Aim: Select proper q-coloring u.a.r.
- Gibb's sampler - select $V_t \in V$ u.a.r
  - select color $J_t \in [q]$ u.a.r from 'allowable colors' $A_v(x) = \{j : x(w) \neq j \ \forall w \in N(v)\}$
  - set $X_{t+1}(V_t) = J_t$
- $P(x,y) = \frac{1}{|V|} \cdot \frac{1}{|A_v(x)|}$

Note though that $\forall$ 'adjacent' configs $x, y \in [q]^V$ we have $|A_v(x)| = |A_v(y)|$.

Thus $P(x,y) = P(y,x)$ $\forall$ adjacent $x,y \Rightarrow \pi(x)$ is uniform

• x, y are adjacent if $x(v) = y(v)$ $\forall v$ except one



$q = 7$

$A_v(x) = \{1, 5, 6, 7\}$

$A_v(y) = \{1, 5, 6, 7\}$

Perfect Sampling

- The MCMC method thus gives us a way to sample any π via a MC. However, how do we know when to stop?

- Strong Doeblin condition: For any MC P on finite $\mathcal{X}$

$$\alpha = \sum_{y\in\mathcal{X}} \underbrace{\min_{x\in\mathcal{X}} \left[P(x,y)\right]}_{\alpha_y}$$

- Now we can write

$$P = \underbrace{\begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \vdots & & & \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \end{pmatrix}}_{Q} + (P - Q) = \alpha \underbrace{\begin{pmatrix} \alpha_1/\alpha & \alpha_2/\alpha & \cdots & \alpha_n/\alpha \\ & \vdots & & \end{pmatrix}}_{1^T\theta} + (1-\alpha)\underset{\text{stochastic}}{R}$$

$$\Rightarrow \quad P = \alpha(1^T\theta) + (1-\alpha)R, \quad R \equiv \text{stochastic}, \quad 1^T\theta = \begin{pmatrix} \alpha_1/\alpha & \alpha_2/\alpha & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Now $\pi^T P = \pi^T \Rightarrow \alpha\left(\pi^T \left(1^T\theta\right)_\alpha\right) + \left((1-\alpha)\pi^T R\right) = \alpha\,\theta^T + (1-\alpha)\pi^T R$

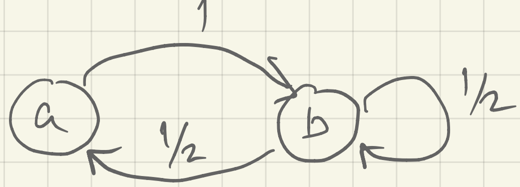
$$\Rightarrow \quad \pi^T = \alpha\theta^T\left(I - (1-\alpha)R\right)^{-1} = \sum_{t=0}^{\infty} (1-\alpha)^t \alpha\, \theta^T R^t \quad (\text{as } |\lambda(R)| \le 1-\alpha)$$

$$= \mathbb{E}_{N,Y}\left[R^{N-1}(Y,\cdot)\right], \quad N \sim \mathrm{Geom}(\alpha),\ Y \sim \theta, \text{ independent}$$

- Algo (Strong Doeblin Sampler) – Sample $X_0 \sim \theta$, $N \sim \mathrm{Geom}(\alpha)$, independent
  – Output X_{N-1}

- $\boxed{\text{Thm}}$ – $X_{N-1} \sim \pi$ (ie., X_{N-1} is a perfect sample from π)

Eg -



$P = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}, \quad \Pi = \begin{pmatrix} 1/3 \\ 2/3 \end{pmatrix}$

$\min_x P(x,y) : 0 \quad 1/2$

- For this example - $\theta = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\alpha = 1/2$,

  $\Rightarrow$ Set $X_0 = \{b\}$, $N \sim \text{Geom}(1/2)$, output $X_{N-1} \sim \Pi$!

- Eg - PageRank MC (Brin, Page, Motwani, Winograd, J. Kleinberg - HITS)
  - Given graph $G(V,E)$, set $\theta = 1^T \, 1/n$, $\alpha \in [0,1)$ (Typically 0.2)

    $|V| = n$
  - $X_{t+1} = \begin{cases} \text{wp } \alpha, \text{ sample } v \in V \text{ u.a.r} \\ \text{else move to random neighbor of } X_t \end{cases}$

    $$P = \alpha \, 1^T \theta + (1-\alpha) \underbrace{D^{-1} A}_{}$$

    Adjacency matrix

    $\text{diag}(1/d_v \ldots)$

    $W \equiv$ random walk matrix
  - Can set any θ. Eg - $\theta^T = e_i^T = \mathbb{1}\{v = i\} \equiv$ Personalized PageRank for node i

---

- Problem - $\boxed{\text{Need } \alpha > 0}$, Not true for many MC.
- Idea behind CFTP - Use $X_t = f(X_{t-1}, Z_t)$ and the grand coupling

  i.e. - Choose $Z_1, Z_2, \ldots$ s.t $\underbrace{f(f \ldots f(f(f(x_0, Z_1), Z_2), Z_3) \ldots Z_\tau)}_{f \circ f \circ f \cdots \circ f(x_0)}$ is equal for all x_0

  τ times

- A more convenient way to write the 'random function' representation is as $X_{t+1} = G_t(X_t)$, where $G_t(\cdot)$ is a random function s.t $G_t \sim \{g_i(\cdot) \text{ wp } p_i\}$.

Eg. for $P = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$, let $g_1(x) = \begin{cases} b \ ; & x=a \\ b \ ; & x=b \end{cases}$, $g_2(x) = \begin{cases} b \ ; & x=a \\ a \ ; & x=b \end{cases}$

$$X_{t+1} = \begin{cases} g_1(X_t) \text{ wp } 1/2 \\ g_2(X_t) \text{ wp } 1/2 \end{cases}$$ is a random fn representation for X_t

---

- Now fix a labelling of time $(-\infty, \dots, -2, -1, 0, 1, 2, \dots \infty)$, with some arbitrary 'current time' 0. Let G_i^j be the composite fn $G_{j-1} \circ G_{j-1} \circ \dots \circ G_i(\cdot)$, ie, the evolution of the MC from $X_i \to X_j$. In particular
  - $G_0^t = G_{t-1} \circ G_{t-2} \circ \dots \circ G_1 \circ G_0$ (forward simulation)

    $G_{-t}^0 = G_{-1} \circ G_{-2} \circ \dots \circ G_{-t+1} \circ G_{-t}$ (backward simulation)

- Similarly we can define two 'coalescence times'

(Forward coalescence) $\tau_F = \inf\{t \geq 0 \mid G_0^t(x) = G_0^t(y) \ \forall x, y \in \mathcal{X}\}$

(Reverse Coalescence) $\tau_R = \inf\{t \geq 0 \mid G_{-t}^0(x) = G_{-t}^0(y) \ \forall x, y \in \mathcal{X}\}$

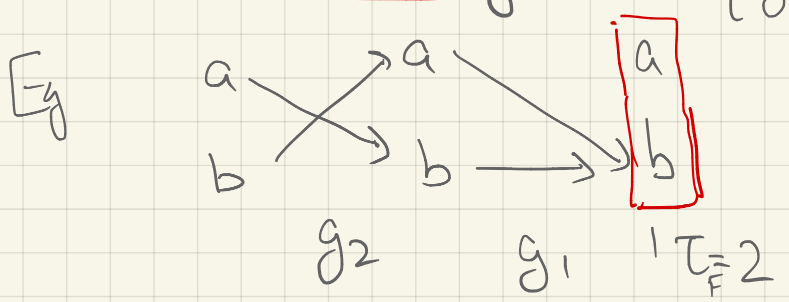
  - In other words, τ_F is the stopping time when all initial states coalesce in the forward simulation; τ_R is the stopping time when all states at time $-\tau_R$ coalesce at 0 in the reverse simulation. We see an example below.

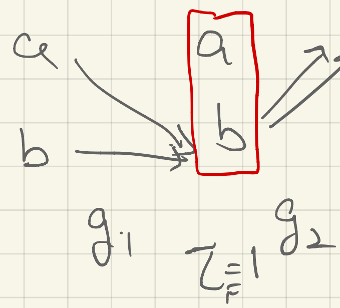
- Note $\tau_F \sim \tau_R$ (ie, they have the same distr^n)

Eg - Consider $P = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$, $g_1(x) = \begin{cases} b \; ; x=a \\ b \; ; x=a \end{cases}$, $g_2(x) = \begin{cases} b \; ; x=a \\ a \; ; x=b \end{cases}$

- Forward coupling - $G_t = \begin{cases} g_1 \text{ wp } 1/2 \\ g_2 \text{ wp } 1/2 \end{cases}$ -

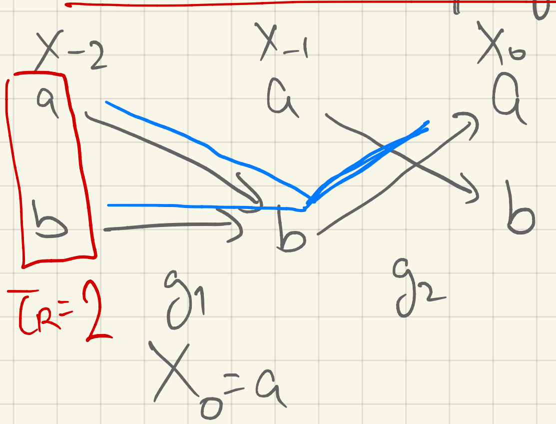
Eg

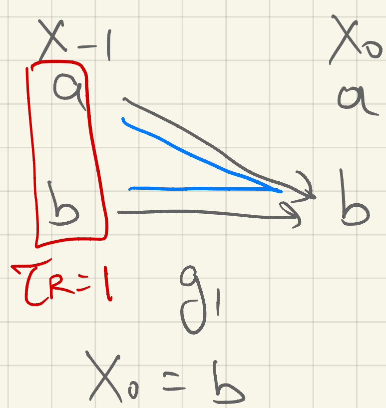




- $\tau_F \sim \text{Geom}(1/2)$, $X_{\tau_F} = b$

- Backward coupling





Again $\tau_R \sim \text{Geom}(1/2)$. However, $X_0 = \begin{cases} a \; ; \tau_R \text{ even} \\ b \; ; \tau_R \text{ odd} \end{cases}$

$$\Rightarrow \mathbb{P}[X_0 = b] = 1/2 + 1/2^3 + \ldots = \frac{1/2}{1 - 1/4} = \frac{2}{3} = \pi(b)!$$

Thus we observe -

i) $\tau_F \sim \tau_R$

ii) $G_0^{\tau_F}(x) = b$ (not $\sim \pi$), $G_{-\tau_R}^{0}(x) \sim \pi$

This procedure to compute $G_{-\tau_R}^{0}$ is called coupling from the past (CFTP)

$\boxed{\text{Thm (CFTP - Propp \& Wilson)}}$ - Assuming τ_R is finite wp 1, then the constant value $Z^0_{-\infty} = G^0_{-\tau_R}(x)$ has distribution

$$Z^0_{-\infty} \sim \pi$$

Pf - Let $\tau_R = \inf_{t \geq 0} \{ G^0_{-t}(x) = G^0_{-t}(x') \ \forall x, x' \in \mathcal{X} \}$

$$\tau^1_R = \inf_{t \geq 0} \{ G^1_{-t}(x) = G^1_{-t}(x) \ \forall x, x' \in \mathcal{X} \}$$

Since τ_R is finite wp 1 $\Rightarrow$ τ^1_R is also finite wp 1, and $Z^0_{-\infty}$ and $Z^1_{-\infty}$ are well defined.

- Now we couple G^0_{-t} and $\hat{G}^1_{-t}$ to use the same $G_k \forall k$, i.e. $\hat{G}^1_{-t} = G_0 \circ G_{-1} \circ G_{-2} \circ \cdots \circ G_{-t+1} \circ G_{-t}$

$$G^0_{-t} = G_{-1} \circ G_{-2} \circ \cdots G_{-t+1} \circ G_{-t}$$

- Let $\hat{Z}^1_{-\infty} = \hat{G}^1_{-\tau^1_R}(x)$, $Z^0_{-\infty} = G^0_{-\tau_R}(x)$

Then $\hat{Z}^1_{-\infty} \sim Z^0_{-\infty}$ (by coupling), and $\hat{Z}^1_{-\infty} = G_0(Z^0_{-\infty})$

$\therefore$ G is a random fn representation for P, this means $\hat{Z}^1_{-\infty} \sim Z^0_{-\infty} \sim \pi$ ($\because$ π is the unique distr s.t if $X \sim \pi$, then $G(X) \sim \pi$). ∎