Lower Bounds on Sample Complexity

Lecture 12
28/3/2021

Theorem.

Let $\mathcal{H}$ be any hypothesis class over a domain $\mathcal{X}$ with a finite VC-dim d. For every $\varepsilon, \delta \in (0,1)$, for any sampling distribution D and any true labelling $f: X \to \{+1,-1\}$,

$$\mathbb{P}_{S\sim D^n}\left[\exists h \in \mathcal{H} : \mathcal{L}_S(h,f) = 0 \wedge \mathcal{L}_D(h,f) > \varepsilon\right] \leq \delta$$

whenever

$$n \geq \frac{8d}{\varepsilon}\log\left(\frac{8e}{\varepsilon}\right) + \frac{4}{\varepsilon}\log\left(\frac{2}{\delta}\right).$$

Defn.

A hypothesis class $\mathcal{H}$ over a domain X is said to be PAC-learnable with a sample complexity $s_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ if

for any $\varepsilon, \delta \in (0,1)$,
for any prob. dist D on X, and
any true labelling $f: X \to \{+1,-1\}$,

$$\mathbb{P}_{S\sim D^n}\left[\exists h \in \mathcal{H} : \mathcal{L}_S(h,f) = 0 \wedge \mathcal{L}_D(h,f) > \varepsilon\right] \leq \delta$$

whenever $n \geq s_{\mathcal{H}}(\varepsilon,\delta)$.

Note: Some definitions insist that $s_{\mathcal{H}}(\varepsilon,\delta)$ is polynomially bounded in $1/\varepsilon$ and $1/\delta$.

THEOREM (equivalently)

A hypothesis class $\mathcal{H}$ with a finite VC-dimension d is PAC-learnable with sample complexity

$$S_{\mathcal{H}}(\varepsilon, \delta) \leq \frac{8d}{\varepsilon} \log\left(\frac{8e}{\varepsilon}\right) + \frac{4}{\varepsilon} \log\left(\frac{2}{\delta}\right)$$

$$\in O\left(\frac{1}{\varepsilon}\left(d \log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right)\right)$$

$$\boxed{\begin{array}{c} \text{VC-dim}(\mathcal{H}) < \infty \\ \Longrightarrow \\ \mathcal{H} \text{ is PAC learnable} \end{array}}$$

CONVERSE

$$\text{VC-dim}(\mathcal{H}) = \infty \overset{?}{\Longrightarrow} \mathcal{H} \text{ is Not PAC-learnable}$$

TIGHTNESS

For VC-dim finite, can we reduce the sample complexity?

VC Dim$(\mathcal{H}) = \infty \Rightarrow$ NOT PAC learnable

Given $\mathcal{H}$. $\forall \mathcal{D}, \forall f, \forall \varepsilon, \forall \delta$

Domain (X)

Ans: YES
VC-dim $(\mathcal{H})$ is finite
$\iff$ $\mathcal{H}$ is PAC-learnable.

CLAIM:

$$S_{\mathcal{H}}(\overset{\varepsilon}{1/2}, \overset{\delta}{1/2}) \geq \frac{1}{2} \text{VC Dim}(\mathcal{H})$$

Given $\mathcal{H}$: Any hypothesis class of VC-dim d

Given X : Domain of $\mathcal{H}$

$f(x) = +1 \ \forall x \in X$

$T = \{x_1, \ldots, x_d\} \subseteq X$: A set shattered by $\mathcal{H}$

$1/d, \ldots, 1/d$

Clever Choice:

$$\mathcal{D} : \mathcal{P}(x) = \begin{cases} 1/d & \forall x \in T \\ 0 & \forall x \in X \setminus T \end{cases}$$

$\varepsilon, \delta = 1/2$

Let $n < d/2$ and $S \sim \mathcal{D}^n$

$h(x) = 0 \ \forall x \in T \setminus S$

$\Rightarrow$ At least half the points in T are not in S ← Training data

$\Rightarrow \exists h : L_S(h,f) = 0 \wedge L_D(h,f) > 1/2$

$\Rightarrow \mathcal{P}_{S \sim D^n}[\exists h : L_S(h,f) = 0 \wedge L_D(h,f) > \varepsilon] = 1$

$> \delta = 1/2$

Hence $S_H(1/2, 1/2) \geq d/2$.

(Note: S_H is decreasing in ε and δ so it is no better for smaller ε and δ)

Can we have a smaller algorithm?

Let 'A' be the small algo.

$h^* = A(S)$ (Deterministic)

"Best choice" among

$\{h \in \mathcal{H} : L_S(h, f) = 0\}$

We will beat it with same D but different f's.

Let $f(x) = \begin{cases} +1, & x \in X \setminus T \\ \begin{cases} +1; & \text{w.p } 1/2 \\ -1, & \text{w.p } 1/2 \end{cases}, & x \in T. \end{cases}$

("Probabilistic Method")

$f \sim F$

$n < d/2$

For any set S of at most n elements from T (At least 1/2 of T is outside S) $|T \setminus S| \geq \frac{1}{2}|T|$.

$$\mathbb{E}_{f \sim F}\left[L_D(A(S), f) \right] = \frac{1}{2}|T\setminus S|$$
$$> \frac{1}{4}$$

Hence $\exists f$ s.t

$\forall S \quad L_D(A(S), f) > \frac{1}{4} = \varepsilon.$

$\Rightarrow P_{S \sim D^n}\left[L_D(A(S), f) > 1/4 \right] = 1$

Can a randomised algorithm work?

Ans: No.

FINITE VC-dimension

- Smaller Sample Complexity?

Final Answer:

$$S_H(\varepsilon, \delta) = \Omega\left(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$$

so tight up to constants.

$\log 1/\varepsilon$ term can be removed if we make "smart algorithms"

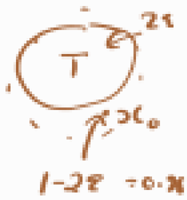
$S_H(1/8, 1/2) \geq d/2$

Idea: $S_H(\varepsilon, \delta) \geq \Omega(d/\varepsilon)$.

$T = \{x_1, \dots, x_d\}$ shattered by H

$x_0 \in X \setminus T$

$$D : P(x) = \begin{cases} 0, & x \in X \setminus (T \cup \{x_0\}) \\ 1 - 2\varepsilon, & x = x_0 \\ \frac{2\varepsilon}{d}, & x \in T. \end{cases}$$



$1-2\varepsilon$

Let $n < d/8\varepsilon$

$$\mathbb{E}_{S \sim D^n}\left[|S \cap T|\right] = n P_{x \sim D}[x \in T]$$
$$= n\, 2\varepsilon$$
$$< \frac{d}{8\varepsilon} \cdot 2\varepsilon$$
$$= d/4$$

"linearity of expectation"

$x > 0$

$P\{x > k\,\mathbb{E}x\} \leq 1/k$

$\therefore P_{S \sim D^n}\left[|S \cap T| \geq \frac{d}{2}\right] \leq 1/2$ (Markov Inequality)

$P_{S \sim D^n}\left\{|S \cap T| < \frac{d}{2}\right\} > 1/2$

$|S \cap T| < \frac{d}{2} \implies \exists h : L_S(h, f) = 0 \wedge L_D(h, f) > \varepsilon$

Hence $P_{S \sim D^n}\left[\exists h : L_S(h, f) = 0 \wedge L_D(h, f) > \varepsilon\right] > 1/2$

i.e. $\quad S_H(\varepsilon, 1/2) \geq d/8\varepsilon$

$$= \Omega(d/\varepsilon)$$

$$\frac{d}{8\varepsilon} \leq S_H(\varepsilon, 1/2) \leq \frac{8d}{\varepsilon}\log(1/\varepsilon) + \frac{4}{\varepsilon}\log\left(\frac{2}{1/2}\right)$$

$$= \frac{8d}{\varepsilon}\log(1/\varepsilon) + \frac{8}{\varepsilon}$$

($\approx$> 64-factor gap ")

Rule of thumb : d/ε

UNIFORM CONVERGENCE

LECTURE 13
4/3/2021

Making full use of PAC learning

We proved that

$\mathcal{H}$ is PAC-learnable

if and only if

VC-dimension of $\mathcal{H}$ is finite

What more can one ask for?

1. Can we tolerate a non-zero in-sample error?

$\exists h \in H : (L_S(h, f) = 0) \wedge (L_D(h, f) > \varepsilon)$

2. Can we use a combination of simple (small VC-dim) classifiers?

3. Anything b/n d and $d+1$?

4. Boosting

Non-Zero In-Sample Error

Goal:

$$\mathbb{P}_{S\sim D^n}\left[\exists h\in\mathcal{H} : \underbrace{|E_D(h,f)-E_S(h,f)|>\varepsilon}_{C_h}\right]\le\delta$$

$$\equiv \mathbb{P}_{S\sim D^n}\left[\bigcup_{h\in\mathcal{H}} C_h\right]\le\delta, \text{ where}$$

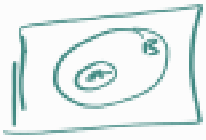
$$C_h = |E_D(h,f)-E_S(h,f)|>\varepsilon$$

$$|\ \mu_X \quad - \quad X\ |$$

$\mathcal{H}$: hypo class
X : domain of $\mathcal{H}$
f : true labelling
D : sampling dist

$\exists h\in\mathcal{H} : E_S(h,f)=0 \ \wedge$
$E_D(h,f)>\varepsilon.$

"mis leading"
Those training data (S) which permit even one $h\in\mathcal{H}$ which has a very different in-sample and out-of-sample error.



Step 1. Fix an arbitrary $h\in\mathcal{H}$.

Let $\mu = E_D[h,f]$ (a deterministic quantity)

$X = E_S[h,f]$ (random variable)

$$= \frac{1}{n}\left|\{x\in S : h(x)\ne f(x)\}\right|$$

$$= \frac{1}{n}(X_1+X_2+\cdots+X_n),$$

where $X_i = \begin{cases}1 & \text{if } h(x_i)\ne f(x_i)\\ 0, & \text{o/w}\end{cases}$

$(S = \{x_1,\ldots,x_n\})$

So X is the average of n independent Bernoulli random variables.

Moreover $\mathbb{E}X = \frac{1}{n}\sum \mathbb{E}X_i$ (linearity of expectation)

$$= \frac{1}{n}\sum \underline{\mathbb{P}_{x_i\sim D}\{h(x_i)\ne f(x_i)\}}$$

$$\boxed{\mathbb{E}X = E_D(h,f)} = \frac{1}{n}\sum E_D(h,f)$$

$$= \frac{1}{n}\sum \mu$$

$$= \mu. \ //$$

If X is a $\{0,1\}$-R.V., then
$\mathbb{E}X = P(X_i=1)\times 1 + P(X_i=0)\times 0$
$= P(X_i=1).$
$= p$

Hence $C_h = |X - \mu_i| > \varepsilon$

C_h : Event that X deviates from its mean by more than ε.

Hoeffding Bounds

Let $X_1, X_2, \dots, X_n$ be independent $\{0,1\}$-random variables with $P(X_i = 1) = p \; (\forall i)$.

Let $X = \frac{1}{n}(X_1 + \dots + X_n)$. Then

$EX = \frac{1}{n} \sum E X_i = p$

(a) $P(X > p + \varepsilon) \leq e^{-2\varepsilon^2 n}$

(b) $P(X < p - \varepsilon) \leq e^{-2\varepsilon^2 n}$

$$\boxed{P\{|X - p| > \varepsilon\} \leq 2e^{-2\varepsilon^2 n}}$$

Hence $\underset{S \sim D^n}{P}[C_h] \leq 2e^{-2\varepsilon^2 n}$ (Exercise: Finite H)

Recall: $P_{S \sim D^n}[A_h] \leq (1-\varepsilon)^n \leq e^{-\varepsilon n}$

$A_h : (E_S(h,f) = 0) \wedge (E_D(h,f) > \varepsilon)$

$1 + x \leq e^x$

$A_h = (E_S(h,f) = 0) \wedge (E_D(h,f) > \varepsilon)$;

$B = \bigcup_{h \in H} B_h$

$B_h = (E_S(h,f) = 0) \wedge (E_{S'}(h,f) > \varepsilon/2)$.

$P(B_h / A_h) \geq \frac{1}{2}$

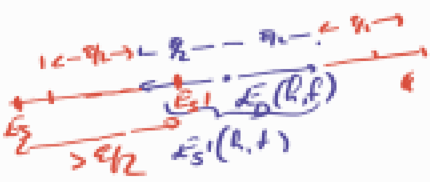
$P(B) \leq \delta/2$

$\Rightarrow P(A) \leq \delta$

$C_h = |E_D(h,f) - E_S(h,f)| > \varepsilon$

$D_h = |E_{S'}(h,f) - E_S(h,f)| > \varepsilon/2$

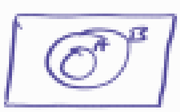
✓ CLAIM. $P(D_h / C_h) \geq 1/2$



PROOF: Let $D'_h = |E_{S'}(h,f) - E_D(h,f)| \leq \varepsilon/2$

Given C_h, $D'_h/C_h \Rightarrow D_h/C_h$

If $A \Rightarrow B$, $P(A) \leq P(B)$.



$$P(D_h/C_h) \geq P(D'_h/C_h)$$

$$P\{|X-\mu_x| > \tfrac{\varepsilon}{2}\} \leq 2e^{-2\cdot\frac{\varepsilon^2}{4}\cdot n}$$

$$\geq |X - \mu_x| \leq \varepsilon/2$$

Average of n Bernoulli's

$$\geq 1 - 2e^{-\frac{\varepsilon^2}{2}n}$$

$$\geq \tfrac{1}{2} \quad \text{if } \underline{n > 4/\varepsilon^2}$$

Hence $P(D/C) \geq 1/2$ when — (1)

$$D = \bigcup_{h \in H} D_h \quad \text{and} \quad C = \bigcup_{h \in H} C_h$$

$S, S' \sim D^n$

New random expt.

$$X \xrightarrow{D^{2n}} T \begin{cases} \nearrow S \\ \text{uniform equipartition} \\ \searrow S' \end{cases}$$

For any h, $P_{S,S' \sim D^n}[D_h]$

$$= P_{\substack{T \sim D^{2n} \\ (S,S') \sim \text{u.e.} T}} [D_h]$$

$$= P_{S,S'}\Big[\,|\underbrace{E_S(h,f)}_{X} - \underbrace{E_{S'}(h,f)}_{X'}| > \varepsilon/2\Big]$$

$$\leq P_{S,S'}\Big[(|E_S(h,f) - \mu| > \varepsilon/4) \vee (|E_{S'}(h,f) - \mu| > \varepsilon/4)\Big]$$

$$\leq 2 \cdot 2e^{-2\frac{\varepsilon^2}{16}n} \qquad \text{(Hoeffding)}$$

$$= 4 \cdot e^{-\frac{\varepsilon^2}{8}n} \qquad \text{— }(*)$$

$|X - X'| > \varepsilon/2$

$EX = EX' = \mu = E_D$

$|X - \mu| > \varepsilon/4$

$|X' - \mu| > \varepsilon/4$

Putting it all together

$$C_h : |L_S(h,f) - L_D(h,f)| > \varepsilon$$

CLAIM

$$D_h : |L_S(h,f) - L_{S'}(h,f)| > \varepsilon/2$$

$$\mathbb{P}_{S\sim D^n}\left[\bigcup_{h\in H} C_h\right] \leq \delta$$

$\Uparrow$ $P(D/C) \geq 1/2$

$$\mathbb{P}_{S,S'\sim D^n}\left[\bigcup_{h\in H} D_h\right] \leq \delta/2$$

$\Uparrow$ Smartest Idea

$h_1 \sim h_2$ if $h_1|_T = h_2|_T$ $(T = S\cup S')$

$$\mathbb{P}_{\substack{T\sim D^n \\ S,S'\sim T}}\left[\bigcup_{[h]_T\in H} D_h\right] \leq \delta/2$$

$\Uparrow$ Shattering

$$g_H(2n)\,\mathbb{P}_{S,S'}[D_h] \leq \delta/2$$

$\Uparrow$ (*)

$$g_H(2n)\, 4 e^{-\frac{\varepsilon^2}{8}n} \leq \delta/2$$

$\Uparrow$ Sauer's Lemma

$$\binom{2n}{\leq d} e^{-\frac{\varepsilon^2}{8}n} \leq \delta/8$$

(VC dim(H) = d)

$\left(\frac{en}{d}\right)^d$ $\Uparrow$

$$n \geq \frac{c}{\varepsilon^2}\left(d\ln(1/\varepsilon) + \ln(1/\delta)\right)$$

THEOREM

VC DIMENSION ($\mathcal{H}$) finite

$\iff$ $\mathcal{H}$ is UNIFORM PAC LEARNABLE

($\mathcal{H}$ has Uniform Convergence Property).

Sample Complexity: $O\left(\frac{1}{\varepsilon^2}\left(d\log 1/\varepsilon + \log 1/\delta\right)\right)$

Converse?

$VC\dim(\mathcal{H}) = \infty$

$\Rightarrow$ $\mathcal{H}$ is not PAC-learnable

$\Rightarrow$ $\mathcal{H}$ is not uniform PAC-learnable.

$(\mathbb{P}(A) \leq \mathbb{P}(C))$.

LECTURE 14
4/3/2021

1. Approximation vs. Generalisation Tradeoff
2. Combination of Concepts.

$\mathcal{L}_S = 0$

$\downarrow$

$|\mathcal{L}_D - \mathcal{L}_S| < \varepsilon$

$\mathcal{L}_D - \mathcal{L}_S < \varepsilon$

$\mathcal{L}_D < \mathcal{L}_S + \varepsilon$

$\uparrow$

Care about

We know

$$P_{S\sim D^n}\left[\exists h \; |\mathcal{L}_S(h,f) - \mathcal{L}_D(h,f)| > \varepsilon\right]$$

$$= P_{S\sim D^n}[C]$$

$$\leq 2P_{S,S'\sim D^n}[D]$$

$$\leq 2 \cdot g_H(2n) P_{S,S'\sim D^n}[D_h]$$

$$\leq 2 g_H(2n) \; 4 e^{-\frac{\varepsilon^2}{8}n}$$

$$\leq 8(2n)^d e^{-\frac{\varepsilon^2}{8}n} \quad \text{--- (1)}$$

Earlier we solved for n so that $(1) \leq \delta$.

We can also solve for ε assuming n is given.

$$8(2n)^d e^{-\frac{\varepsilon^2}{8}n} \leq \delta$$

$$e^{\frac{\varepsilon^2}{8}n} \geq \frac{8}{\delta}(2n)^d$$

$$\varepsilon^2 \geq \frac{8}{n}\left[\ln\left(\frac{8}{\delta}\right) + d\ln 2n\right]$$

$$\varepsilon \geq \sqrt{\frac{8}{n}\left[d\ln 2n + \ln 8/\delta\right]}$$

Hence $= \Omega$. $\approx \tilde{O}\left(\sqrt{d/n}\right)$

$$|\mathcal{L}_n - \mathcal{L}_D| \leq \Omega \quad \text{w.p. } 1-\delta$$

(Generalisation error)

$$\mathcal{E}_D \leq \underbrace{\mathcal{E}_{in}}_{\text{Approximation Error}} + \underbrace{\Omega}_{\text{Generalisation error}}$$

"Occam's Razor"

Ω increases with increase in d/n ✓

$\mathcal{E}_{in}$ is likely to decrease with d/n. ✓

Hence choice of d (i.e. choice of $\mathcal{H}$) is a trade off.

"Amateur ML engineers" tend to choose a larger d/n so that $\mathcal{E}_{in}$ (which is staring at their face) is low.

"Overfitting"

"Professional ML engineers" are more careful about Ω

(Prof. Abu-Mostafa, CALTECH)

Finding the sweet spot?

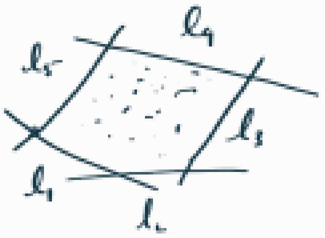
- Bias-variance trade off
- Regularisation
- Validation

(Not in this course).

Combining Hypothesis Classes

Scenarios

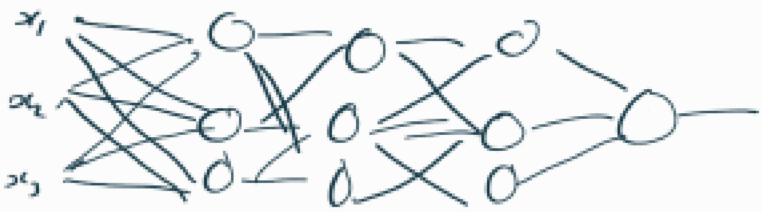
1.



$G = G_1 \cap G_2 \cap G_3 \cap B_4 \cap B_5$

$f: x_1 \wedge x_2 \wedge x_3 \wedge \overline{x_4} \wedge \overline{x_5}$

2



"Perception n/w"

Let H be a hypothesis class that is obtained as a combination of k hypothesis classes $H_1, \ldots H_k$.

ie $\quad h(x) = f(h_1(x), \ldots, h_k(x))$

where $h_i \in H_i$ and

$f: \{0,1\}^k \to \{0,1\}$ is any Boolean function.

Then $\quad$ VC-dim$(H) = ?$

Will depend on f $\quad (2^{2^k}\ f's)$

An upper bound independent of f? $\quad$ Yes

Simplification: $\quad \forall i,\ $ VC-dim $(H_i) \leq d$.

Claim: $\text{VC-dim}(\mathcal{H}) \leq 2kd \ln kd$

<u>Proof.</u>

Let S be a largest shattered by $\mathcal{H}$ $\quad (\text{VC-dim}(\mathcal{H}) = |S| = n)$

Let $|S| = n$

$\Rightarrow |\mathcal{H}|_S| = 2^n \quad \text{--- (1)}$

$h, h' \in \mathcal{H}$

$h|_S \neq h'|_S$

Consider $\mathcal{H}^* = \mathcal{H}_1 \times \mathcal{H}_2 \times \dots \times \mathcal{H}_k$

$= \{(h_1, \dots, h_k) : h_i \in \mathcal{H}_i\}$

$h(x) = f(h_1(x), \dots, h_k(x))$

$h'(x) = f(h_1'(x), \dots, h_k'(x))$

$h'(x) \neq h_2(x)$ needs atleast one $h_i(x) \neq h_i'(x)$

Obs 1. $|\mathcal{H}|_S| \leq |\mathcal{H}^*|_S|$

<u>Obs 2</u> $\mathcal{H}^*|_S = \mathcal{H}_1|_S \times \mathcal{H}_2|_S \times \dots \times \mathcal{H}_k|_S$

$$|\mathcal{H}^*|_S| = |\mathcal{H}_1|_S| \times \dots \times |\mathcal{H}_k|_S|$$

$$\leq \binom{n}{\leq d_1} \times \dots \times \binom{n}{d_k}$$

$$\leq n^{d_1} \times \dots \times n^{d_k}$$

$$\leq n^{kd}$$

Hence $2^n \leq n^{kd}$

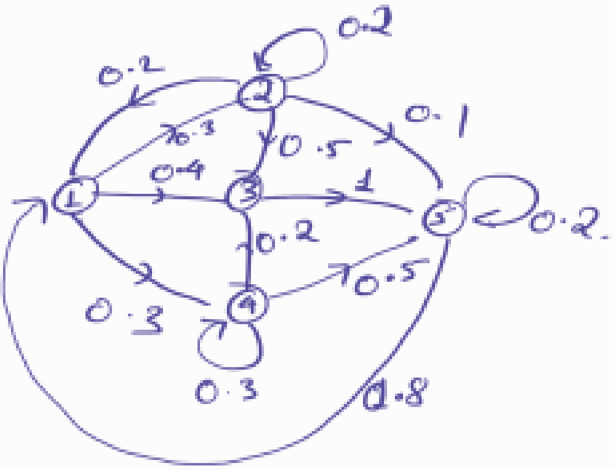
$\Updownarrow$

$n \leq 2kd \log kd$ ∎

<u>Fact</u> Perceptron networks (neural networks with hard thresholds) have VC-dim $O(m \log m)$, where m is the no. of edges.

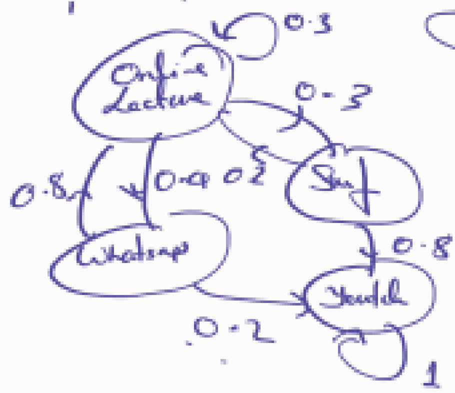
Intro to Random Walks on Graphs.

Lecture 15
7/3/2021



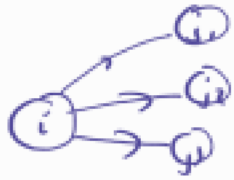
Examples:

- (Web Pages, Hyperlinks)
- (Activities, Switching)



Edge weight = transition probabilities

P_{ij} = P[Next State = j / Current State = i]



Hence $\forall i$, $\sum_{j \in \Gamma(i)} P_{ij} = 1$ — (P1)

(outarrows from i).

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ P_{21} & \cdots & P_{2n} \\ \vdots & & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}_{n \times n}$$

Transition Probability Matrix (TPM).

"Stochastic Matrix"

$M_{ij} \geq 0$

$\sum_{j=1}^{n} M_{ij} = 1$ (row sum)

P1 ≡ Every row of P sums to 1.

$$\begin{bmatrix} P_{11} & \cdots & P_{1n} \\ & & \\ P_{n1} & \cdots & P_{nn} \end{bmatrix}_{n \times n} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Hence $\mathbb{1} = [1,1,\ldots,1]^T$ is a (right) eigen vector of P with eigen-value 1.

In fact, If P is a real non-negative matrix then

P is a stochastic Matrix $\iff$ $P\mathbb{1} = \mathbb{1}$ (P2)

Corollary: If P_1 & P_2 are $n \times n$ SMs, so is $P_1 P_2$.

$P = P_1 P_2$

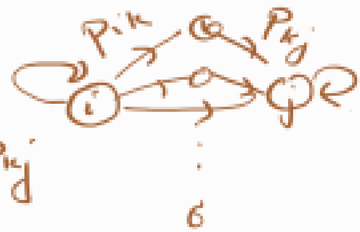
$P[i,j] \geq 0$

$$P_1 P_2 \mathbb{1} = P_1 (P_2 \mathbb{1}) = P_1 \mathbb{1} = \mathbb{1}.$$

Corollary: If P is a SM, $\forall k \in \mathbb{N}$, P^k is a S.M.

$$P^k[i,j] = P[\text{State after } k \text{ steps} = j / \text{Current State} = i]$$

(Proof: Exercise)



$$P[\text{Next to Next state} = j / \text{Current state} = i] = \sum_{k=1}^{n} p_{ik} p_{kj} = P^2[i,j]$$

Prop 3.

All eigen values of P have magnitude ≤ 1

i.e. $Px = \lambda x \Rightarrow |\lambda| \leq 1$

Proof: Let $x = (x_1, \dots, x_n)$ be an eigen vector corresponding to λ s.t. some $x_i = 1$ and $|x_j| \leq 1 \ \forall j$ (Scaling)

$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ $x^T = [x_1 \dots x_n]$

We have $Px = \lambda x$

$\langle (i^{th} \text{ row of } P), x \rangle = \lambda x_i$

$$\sum_{j=1}^{n} P_{ij} x_j = \lambda 1.$$

$$\lambda = \sum_{j=1}^{n} P_{ij} x_j$$

$$|\lambda| \leq \sum_{j=1}^{n} P_{ij} |x_j| \quad \text{--- (1)}$$

$$\leq \sum_j P_{ij} 1.$$

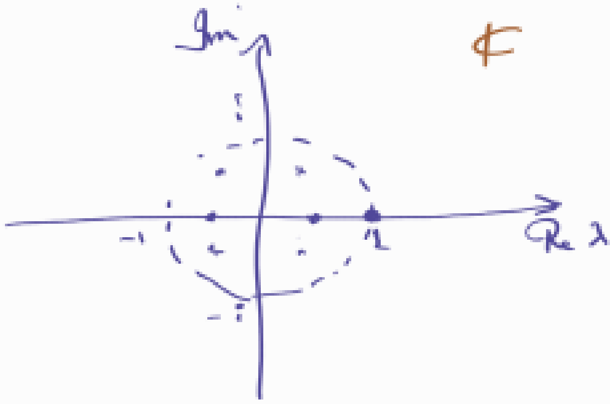
$$= 1.$$

$|z_1 + z_2| \leq |z_1| + |z_2|$

(tight $\Rightarrow |x_j| = 1 \ \forall j$

$\Rightarrow x_j = 1$

$\forall j : P_{ij} > 0$



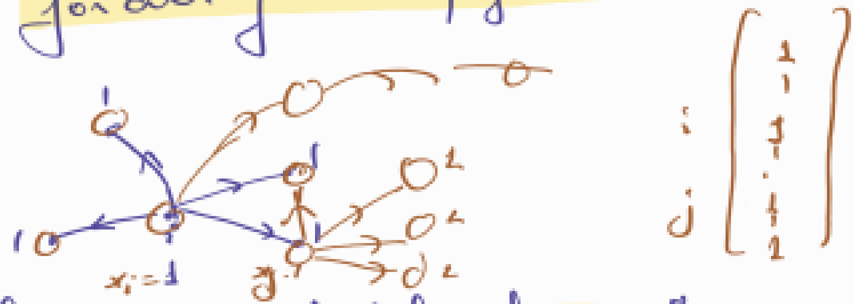
$\mathbb{C}$

Next Question:

What is the multiplicity of $\lambda = 1$?

If $\lambda = 1$, then eqn(1) above has to be an equality.

i.e. $x_j = 1$ for all j s.t. $P_{ij} \neq 0$



If every vertex is reachable from i, then $x_j = 1 \ \forall j \Rightarrow x = \mathbb{1}$

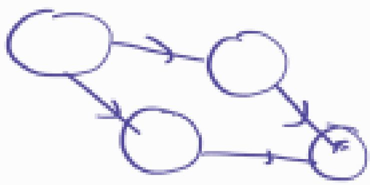
** Prop 4.

If the underlying graph G is strongly connected then the eigenvalue 1 has multiplicity 1.

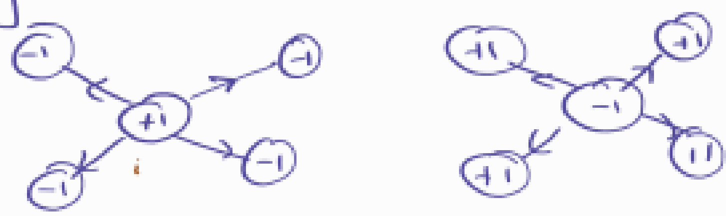
$\equiv$ $\mathbb{1}$ is the unique eigenvector (upto scaling) for $\lambda = 1$.

Converse? Exercise:

(Hint. Sink component).



When do we get $\lambda = -1$.



iff G is "bipartite".
Strongly connected $\Rightarrow$ -1 also has multiplicity 1.

Qn. Is P full rank?

Not necessarily

$$\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Interestingly, the $n \times (n+1)$ matrix
$P - I : \mathbb{1}$ has rank n.

(Exercise).

Summary:

Let P be the TPM of a random walk on a finite directed graph G.

(1) $\sum_{j=1}^{n} P_{ij} = 1$ (TPM is a SM)

(2) $P\mathbb{1} = \mathbb{1}$, and hence $\lambda = 1$ is an eigen value

(3) All eigen values of P lie in the closed unit disk around 0.

(4) $\lambda = 1$ has multiplicity 1 iff G is strongly connected.

(5) -1 is an eigen value of P iff G is bipartite.

Note:

All these apply to undirected graph as well.

If G is undirected P is symmetric.

$\Rightarrow$ Eigen values are real

$\Rightarrow \in [-1, 1]$.

Now that we have learned so much about P, let's move on and study

P^T

P^T is more important than P.

$(x_1, \dots x_n)$ is called a prob. vec if $x_i \geq 0$ & $\sum x_i = 1$

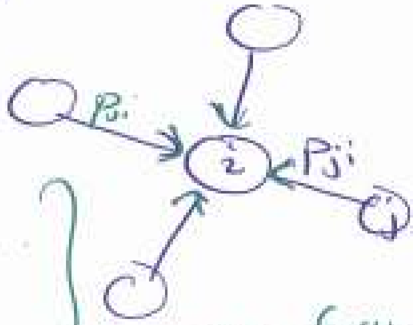
Why

Let $p = (p_1, \dots, p_n)$ be the node probabilities at current step. ← Distribution at time t

Then

$p' = P^T p$ gives the node probability at the next step. ← Distribution at time $t+1$.

$$p_i' = \sum_{j=1}^{n} P_{ji}\, p_j = [P^T p]_i$$



$p(t) = [p_1(t), \dots\ p_n(t)]$

$p(t+1) = P^T p(t)$ (if $p(t)$ is written as a col vector)

$p(t+1) = p(t)\, P$ (if $p(t)$ is written as a row vector)

P^T captures "Evolution of the random walk".

$p(0) = [0, 0, 1, 0, 0 \dots 0]$

↑ start node

$P^T p(0) = p(1) = (0, 1/3, 1/3, 1/3, 0 \dots)$

$P^T p(1) = p(2) = \dots$

* Distribution at time t
  prob. vec $p(t)$

* Evolution
  $p(t+1) = P^T p(t)$.

Properties of P^T

(1) Eigen values (and multiplicities) of P and P^T are the same.

(Proof: Determinant
$\det(P^T - \lambda I) = \det((P-\lambda I)^T)$
$= \det(P - \lambda I)$.

$x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is called a probability vector if
(i) $x_i \geq 0 \ \forall i$
(ii) $\sum_{i=1}^{n} x_i = 1$.

(2) x is a prob. vec $\Longrightarrow$ $P^T x$ is a prob. vec.

Proof: Let $y = P^T x$
$y_i \geq 0$, obvious.

$$\sum_{i=1}^{n} y_i = \mathbb{1}^T y$$
$$= \mathbb{1}^T P^T x$$
$$= (P\mathbb{1})^T x$$
$$= \mathbb{1}^T x$$
$$= 1.$$

$(1, \dots 1)\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = 1$.

If G is strongly connected, then
(3) P^T also has a unique eigen vector (upto scaling) for $\lambda = 1$.
But is it a prob. vec?

$P^T p = p$
"Stationary distributi

STATIONARY DISTRIBUTIONS of RANDOM WALKS

LECTURE 16
11/4/2021

Defn. $\pi \in \mathbb{R}^n$ is called a *stationary distribution* of a random walk with transition prob. matrix P if

(i) π is a prob. vec., and

(ii) $P^T\pi = \pi$.

P^TP

P^TP

$P^Tx = 1x$

Can we have x to be prob. vec?

$P^TP = P$

Theorem 1. (Fundamental theorem of finite Markov Chains)

** A random walk on every finite graph has a stationary distbn.

Obs. Finiteness is necessary.

$① \xrightarrow{1} ② \xrightarrow{1} ③ \cdots$

$V(G) = \mathbb{N}$
$E(G) = \{(i, i+1) : \omega t\ 1\}$

P^TP

(Many important infinite cases can be handled - but not in this course)

Standard proofs.

Key idea "Fixed point theorems"

Let $f: X \to X$, then a point $x \in X$ s.t. $f(x) = x$ is called a *fixed point* of f.

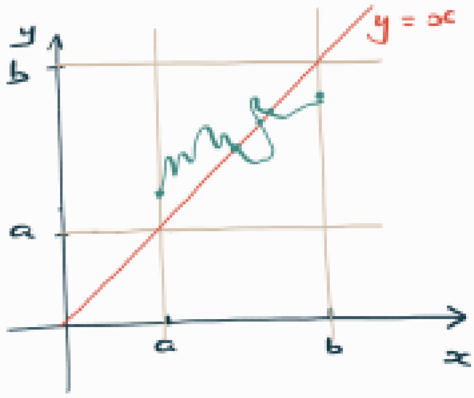
Domain X
$f: X \to X$

$X = \{1, 2, 3\}$
$f(x) = 4 - x$
$1 \mapsto 3$
$2 \mapsto 2$
$3 \mapsto 1$

Example: Any continuous fn from a closed interval $[a,b]$ to itself has a fixed point.

$X = [a,b]$
$f: X \to X$, continuous

Let $f: [a,b] \to [a,b]$, cont



Non-example

$f: [n] \to [n]$, derangement (cyclic shift).

$1 \mapsto 2$
$2 \mapsto 3$
$3 \mapsto 1$

Fact 1 If X is a compact ~~convex~~ set and f is continuous then f has a fixed point (Brouwer's FP theorem, 1909).

$X \subseteq \mathbb{R}^n$

Compact = Closed & Bounded

Fact 2 Closed and bounded sets in $\mathbb{R}^n$ are compact.

$[a,b]$ closed
$(a,b]$ not closed

$a_n = a + \frac{1}{n}$
$\to a$ as $n \to \infty$

What's our X?

X = set of all prob. vecs in $\mathbb{R}^n$ $(n = |V(G)|)$

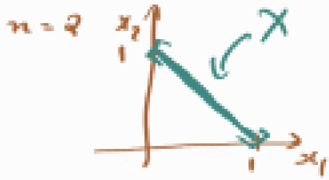
$= \{(p_1, \dots, p_n) : p_i \geq 0, \sum p_i = 1\}$

$f: X \to X$
$p \mapsto P^T p$
$p \in X \Rightarrow P^T p \in X$

Verify : X is ✓(i) convex
✓(ii) bounded
✓(iii) closed

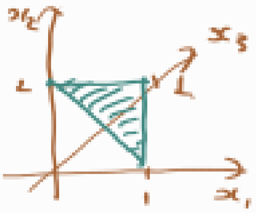
$\exists\, x \in X$ s.t $f(x) = x$,

(i) $p, q \in X$
$\lambda p + (1-\lambda) q \in X$
(Exercise)



(ii) X is bounded
$\|x\|_\infty$: max x_i

Then $f : X \to X$
$x \mapsto P^T x$



$\{\|x\|_\infty : x \in X\} \le B$
$\|$
1

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

$\|x\|_2 = \sqrt{\sum_{i=1}^{n} (x_i)^2}$

Verify - f is continuous.

$x,\ x + \Delta x$
$f(x+\Delta x) - f(x) = f(\Delta x) \to 0$

Done! f has a fixed point in X.

That is your stationary distbn.

Direct Proof.

Let x be any prob. vec.
Consider the sequence of prob. vectors $(x, P^T x, (P^T)^2 x, \ldots)$

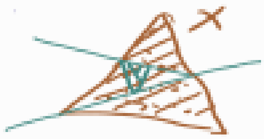
$x = [p_1, p_2, \ldots, p_n]$ $\quad x,\ xP,\ xP^2,\ xP^3, \ldots\ldots$

(evolution of the random walk
$x(0),\ x(1),\ \ldots\ \ldots,$ where
$x(t) = xP^t$)

Let $a(t) = \dfrac{1}{t}\big(x(0) + x(1) + \cdots + x(t-1)\big)$
$= \dfrac{1}{t}\big(x + xP + xP^2 + \cdots\ xP^{t-1}\big)$

"long-term average"

Claim 1. $\forall t$, $a(t)$ is a prob. vec



$\therefore\ a(1), a(2), \ldots \in X$

X bounded $\Rightarrow$ $a(1), \ldots$ contains a convergent subseq $a(t_1), a(t_2), \ldots$
$\qquad\qquad 3 \qquad 7$

(Proof: Pigeonhole principle).

X closed $\Rightarrow \lim\limits_{n\to\infty} a(t_n) = a \in X$,

(i.e. a is a prob. vec.)

$\forall t$,

$$a(t)P - a(t) = \frac{1}{t}(xP + xP^2 + \cdots \quad xP^t) - \frac{1}{t}(x + xP + \cdots \quad xP^{t-1})$$

$$= \frac{1}{t}(\underset{\in X}{xP^t} - \underset{\in X}{x})$$

"P-q"

$(\|x\|_\infty = \max\limits_i |x_i|)$

$$\|a(t)P - a(t)\|_\infty \le 1/t \quad (*)$$

$$\|a(t_n)P - a(t_n)\|_\infty \le \frac{1}{t_n} \to 0 \text{ as } n\to\infty$$

$$\Rightarrow \|aP - a\|_\infty = 0$$

$$aP = a.$$

Hence a is a stationary distribution.

Summary.

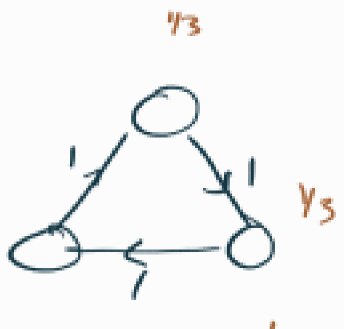
1. Every random walk on a (every) finite graph } has a stationary distribution π $(\pi P = \pi)$

2. If the graph is strongly connected } the stationary distribution is unique, and $a(t) \to \pi$ as $t \to \infty$

Qn. If $p(0)$ is an arb. starting distbn does $p(n) = p(0)P^n \to \pi$ as $n \to \infty$?

Ans: Not always, but Yes in most cases.

Some no cases.



$p(0) = 1 \quad 0$
$p(1) \quad 0 \quad 1$
$p(2) \quad 1 \quad 0$

$1/4 \quad 3/4$
$3/4 \quad 1/4$

$1, 0, 0 \to 0, 1, 0 \to 0, 0, 1$

Yes iff gcd of lengths of all directed cycles in G is 1!

Summary (Once again)

1. Every random walk on a finite graph } has a stationary distribution π

2. If the graph is strongly connected } the stationary distribution is unique. and $a(t) \to \pi$ as $t \to \infty$

3. If the graph G is strongly connected and gcd(cycle lengths) = 1 } For any prob. vec $p(0)$, $p(t) \to \pi$ as $t \to \infty$

Applications

Analysing natural phenomena

- Brownian motion
- Financial markets
- Text/Speech
- Genetics
- Web

Generating samples according to some target distribution ($x \sim D$)

- Card shuffling

→ Metropolis-Hastings algo

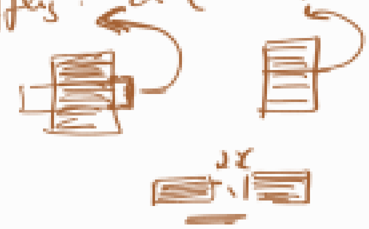
→ Gibbs sampling

$(p_1, p_2 \dots \quad p_n)$ $n = $ #webpages

$\uparrow$ $1/n$

52 cards

Shuffles : $\mathcal{U}(52!)$



Markov Chain Monte Carlo Methods

Lecture 17
17/4/2021

Goal: Sample a point x according to a distribution D on X.

$$x \sim D$$

E.g.: σ is a permutation of $[n]$ chosen uniformly at random.

$x \in \mathbb{R}^n$ is sampled from an n-dimensional gaussian

Key Idea: (π : distbn over X)

Let the domain X be finite ($|X| = n$)

Design a directed graph G and transition probabilities P s.t π is a stationary distribution of P

π will be a n-length prob-vec
$\pi = (\pi_1, \ldots, \pi_n)$
s.t $\pi_i \geq 0$, $\sum_{i=1}^{n} \pi_i = 1$

Lec 16: $P \rightarrow \pi$
Lec 17 $\pi \rightarrow P$

Desirable properties:

(1) G is strongly connected. : unique st. dist

(2) gcd (cycle lengths) = 1. : $p(t) \rightarrow \pi$

(3) Low degrees

(4) Symmetries (regular for example)

(5) Rapid mixing (convergence to stationary distribution).

Easy Case: π is uniform on X. $\pi_i = 1/n$.

G : undirected, connected non-bipartite graph with $V(G) = X$

: k-regular $(k > 2)$

: $P_{ij} = 1/k \quad \forall i, j$

: Expander.

$P^T = P$

$\mathbb{1}$ is eigen vec of P & P^T (since $P^T = P$)

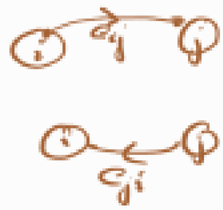
$\therefore \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = \pi$ is the st dist of P

Non-uniform π

A simple sufficient condition.

If π is a prob. vec. & P is a st. mat s.t.

$$\forall i, j. \quad \pi_i P_{ij} = \pi_j P_{ji}, \qquad — (*)$$



then π is a stationary distbn of P.

$\overline{P^T \pi = \pi}$

Proof:

Let $\sigma = P^T \pi$

Then
$$\sigma_i = \sum_{j=1}^{n} P_{ji} \pi_j$$
$$= \sum_{j=1}^{n} P_{ij} \pi_i \qquad (\text{by } *)$$
$$= \pi_i$$

$\blacksquare$

P^T

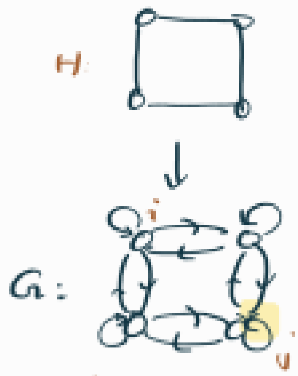
$$\begin{array}{cc} i\text{th row} & \begin{bmatrix} \longleftrightarrow \end{bmatrix} \end{array} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Metropolis-Hasting

Input: π, a prob. dist on $[n]$
(an n-length prob. vec)

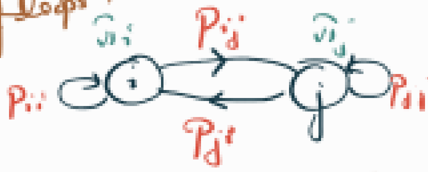
Design of G

Pick a "good" connected undirected graph H and replace each undirected edge with 2 opposite arcs and add a self loop at each node to get G.



Obs. G is strongly connected,
gcd(cycle lengths) = 1.

Automatically satisfied for missing edges of H and self loops.

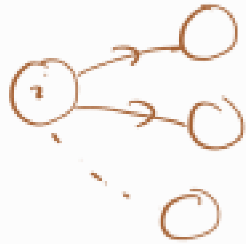


Design of P

(Aim: $\forall i,j \quad \pi_i P_{ij} = \pi_j P_{ji}$)

Attempt 1.

$$P_{ij} = \pi_j \quad \forall j \in N^+(i)$$



Then, $\pi_i P_{ij} = \pi_i \pi_j$:) ✓
$\pi_j P_{ji} = \pi_j \pi_i$

But is $\sum_{j=1}^{n} P_{ij} = 1$?

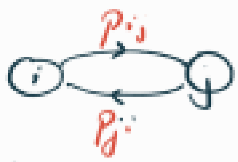
$$\sum_{j=1}^{n} P_{ij} = \sum_{j \in N^+(i)} \pi_j \ll 1 \text{ if } H \text{ is sparse}$$

We can fix this by changing P_{ii}

$$P_{ij} = \begin{cases} \pi_j & , j \in N^+(i) \setminus \{i\} \\ 1 - \sum_{k \in N^+(i) \setminus \{i\}} \pi_k & , j = i \end{cases}$$

Issue: too slow a walk.

Why not scale?



Subject to

1. Equal scaling for P_{ij} and P_{ji}
2. $\sum_{j\neq i} P_{ij} \le 1$.

   (Deficit $\rightarrow P_{ii}$).

Let σ = maximum out-degree of G not counting the self loop.

(2) is ensured if $\forall i,j \quad P_{ij} \le 1/\sigma \quad (i \neq j)$ — $1/\alpha$

Scaling for $P_{ij} = \dfrac{1/\sigma}{\max\{P_{ij}, P_{ji}\}}$

α — β —

Scale factor $\frac{1/\alpha}{\beta}$

Hence

$$\forall j \in N^+(i)\setminus\{i\},\ P_{ij} = \boxed{\frac{1/\sigma}{\max\{\pi_j, \pi_i\}} \times \pi_j}$$

$$= \frac{1}{\sigma} \times \min\left\{\frac{1}{\pi_j}, \frac{1}{\pi_i}\right\} \times \pi_j$$

$$= \frac{1}{\sigma} \min\{1, \pi_j/\pi_i\}$$

Metropolis-Hasting $(G \rightarrow H)$

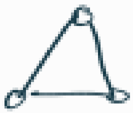
$$\forall i \quad P_{ij} = \begin{cases} \frac{1}{\sigma}\min\{1, \pi_j/\pi_i\} & , j \in N^+(i)\setminus\{i\} \\ 1 - \sum_{j \in N^+(i)\setminus\{i\}} P_{ij} & , j = i \end{cases}$$

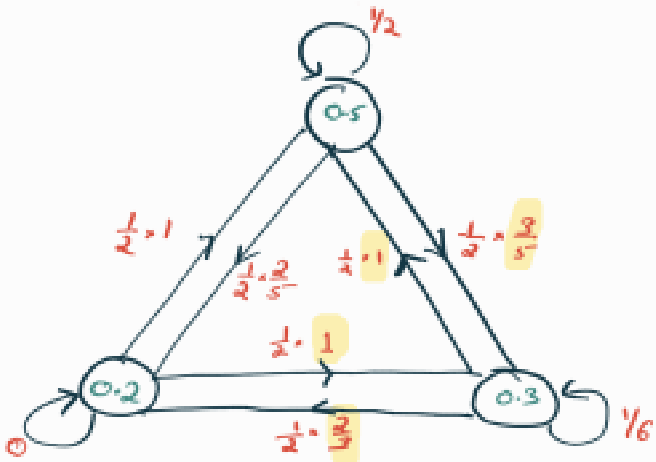
Sanity check.

$$\pi_i P_{ij} = \frac{\pi_i}{r}\min\{1, \pi_j/\pi_i\}$$
$$= \frac{1}{r}\min\{\pi_i, \pi_j\}$$
$$= \frac{\pi_j}{r}\min\left\{1, \frac{\pi_i}{\pi_j}\right\}$$
$$= \pi_j P_{ji} \quad ✓$$

$$\forall i, \quad \sum_{j \in N^+(i)} P_{ij} = 1. \quad ✓$$

Example: $\pi = (0.5, 0.3, 0.2)$

H :  G:  $r = 2$.



Walker's rule

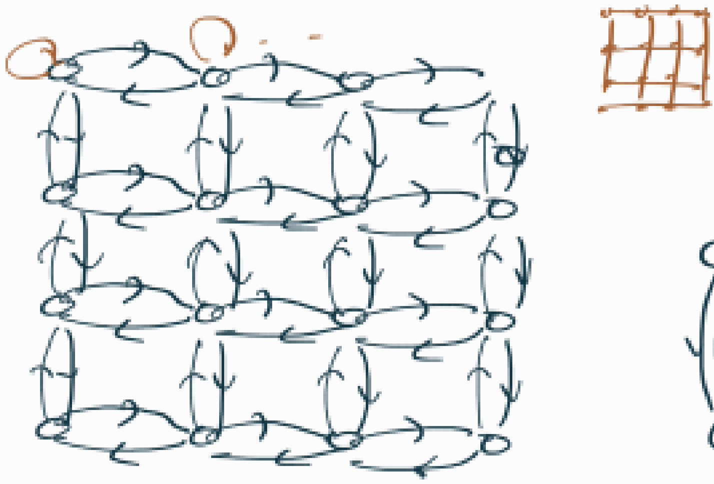
When at node i

- Select each out edge (i,j) w.p. $1/r$ (i,j)
- If $(\pi_j \geq \pi_i)$
  move to node j (w.p. 1)

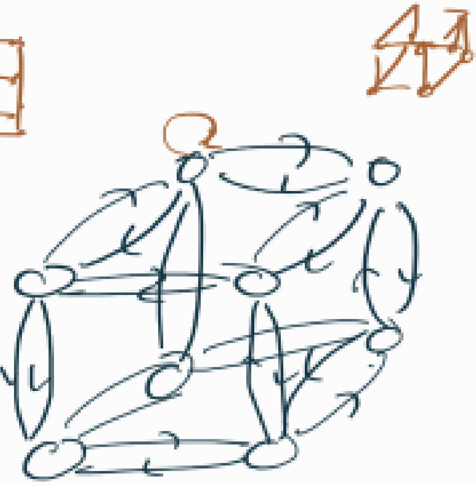
  Else
  move to node j w.p. π_j/π_i

Usually,

Graph: d-dimensional lattice $[m]^d$



$d=2$, $m=4$



$d=3$, $m=2$

- $n = m^d \geq |X|$
- Almost $2d$-regular (except boundary vertices), $r = 2d$.

Gibbs Sampling

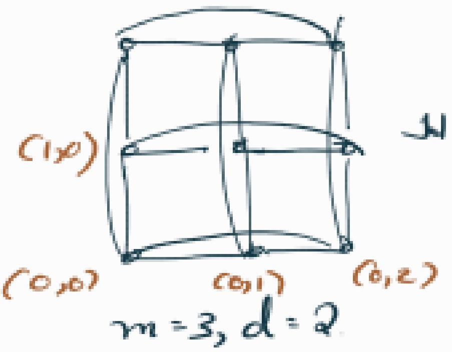
Domain: $\mathcal{X} = [m]^d$
$= \{(x_1, \dots, x_d) : x_i \in [m]\}$

Target: π, a distbn on $\mathcal{X}$

——— × ———

$\mathcal{H} = \underbrace{K_m \square K_m \square \cdots \square K_m}_{d \text{ times}}$

(Hamming graph)



$V(\mathcal{H}) = \mathcal{X}$

$\{x,y\}$ is an edge of $\mathcal{H}$ iff x & y differ in exactly one co-ordinate.

Obs. Much denser than lattice (M-H).

Let $x = (x_1, x_2, \dots, x_d)$
$y = (y_1, x_2, \dots, x_d)$, $y_1 \neq x_1$
(x and y are adjacent in $\mathcal{H}$)

$$P_{xy} = \frac{1}{d}\pi(y_1 / x_2, \dots x_d).$$

$$\left[\pi(y_1/x_1, \dots x_d) = \frac{\pi(y_1, x_2, \dots, x_d)}{\sum_{z=1}^{m} \pi(z, x_2, \dots x_d)} \right]$$

Hence $\pi(x) P_{xy} = \pi(y) P_{yx} \quad \forall x, y.$

$$= \frac{\pi(x_1, x_2 \dots x_d)\, \pi(y_1, x_2, \dots x_d)}{\sum_{z=1}^{m} \pi(z, x_2, \dots x_d)}.$$

$$\sum_{y \in N^+(x)} P_{xy} = \underbrace{\frac{1}{d} + \frac{1}{d} + \cdots \frac{1}{d}}_{d \text{ times}}$$

$$= 1.$$

<u>Walker's Rule</u>

- When at node $(x_1, \ldots, x_d)$
  - Pick a dim $k \in [d]$ w.p $1/d$ each
  - (Follow attempt 1 on dimension k)
    Move to $(x_1, \ldots x_{k-1} y_k, x_{k+1}, x_d)$
    w.p. $\pi(y_k / x_1, \ldots x_{k-1}, x_{k-2}, \ldots x_d)$

Example: