Math 554

Hemanshu Kaul

kaul@iit.edu

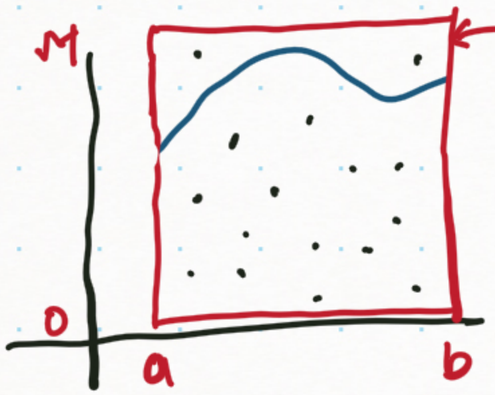
Counting via Markov Chain Monte Carlo (MCMC)

Classical Monte Carlo is a method for estimating quantities that are hard to compute.

$$Z = \mathbb{E}[Z]$$

- Z: quantity we want to compute
- $\mathbb{E}[Z]$: random variable that samples underlying "objects" from a probability space Ω

As long as there is an efficient procedure for sampling from Ω we can take the mean of sufficiently large set of independent samples of Z, to get an approximation of Z.



Generate random points in BBox unit.

Then

$$\text{area under the curve} \propto \left(\frac{\#\text{successful trials}}{\#\text{Total Trials}}\right)(\text{Area of BBox})$$

Counting via Markov Chain Monte Carlo (MCMC)

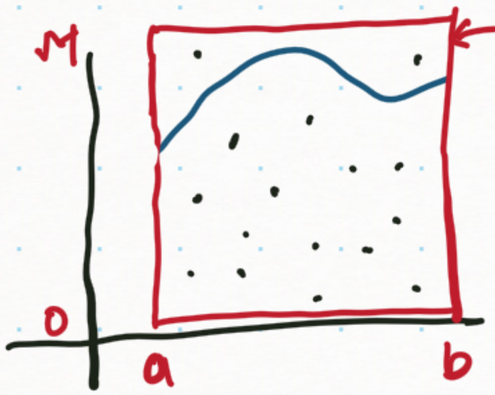
Classical Monte Carlo is a method for estimating quantities that are hard to compute.

$|S|$ where S is a collection of ind. sets of G, etc. $\longrightarrow Z = \mathbb{E}[Z]$

Z: quantity we want to compute

Z: random variable that samples underlying "objects" from a probability space Ω

As long as there is an efficient procedure for sampling from Ω we can take the mean of sufficiently large set of independent samples of Z, to get an approximation of Z.



Generate random points in BBox unit.

Then

$$\text{area under the curve} \propto \left(\frac{\#\text{successful trials}}{\#\text{Total Trials}}\right)(\text{Area of BBox})$$

But what if the "objects" to be sampled are not points but are combinatorial structures?

How can we sample?

Short review of Homogenous Discrete Time Markov Chains

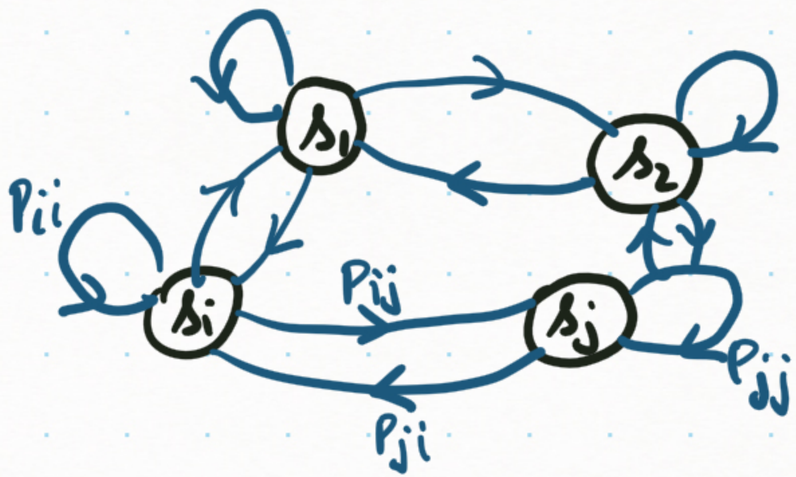
A random process $(X_0, X_1, \dots)$ on a finite space $S = \{s_1, \dots, s_k\}$ is said to be a <u>HDT Markov Chain</u> with transition matrix $P_{k \times k}$ matrix if $\forall n, \forall i,j \in [k], \forall i_0, \dots, i_{n-1} \in [k]$

$$\mathbb{P}[X_{n+1} = s_j \mid X_0 = s_{i_0}, X_1 = s_{i_1}, \dots, X_{n-1} = s_{i_{n-1}}, X_n = s_i]$$

$$= \mathbb{P}[X_{n+1} = s_j \mid X_n = s_i] \leftarrow \text{Memorylessness property / Markov prop.}$$

$$= P_{ij} \leftarrow (i,j) \text{ entry of } P$$

- $P_{ij} \in [0,1]$
- $\sum_{j=1}^{k} P_{ij} = 1 \quad \forall i$



<u>on edge wt.ed digraph on S</u>

State space

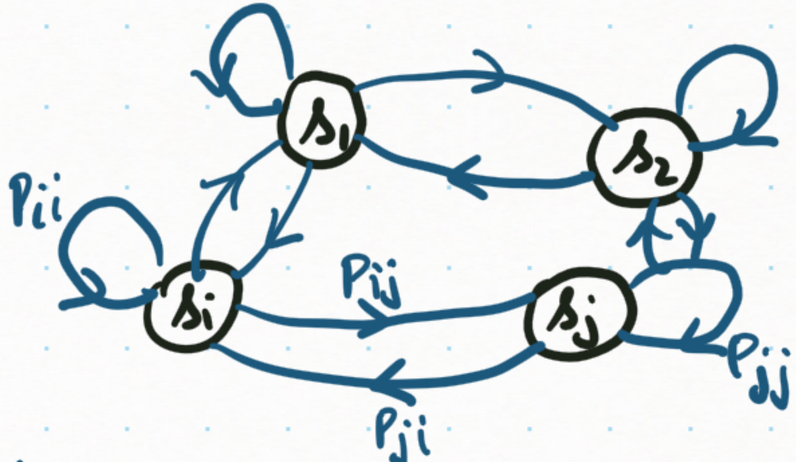
Short review of Homogenous Discrete Time Markov Chains

A random process $(X_0, X_1, \cdots)$ on a finite space $S = \{s_1, \ldots, s_k\}$ is said to be a <u>HDT Markov Chain</u> with transition matrix $P_{k \times k}$ matrix if $\forall n, \forall i,j \in [k], \forall i_0, \ldots, i_{n-1} \in [k]$

$$\mathbb{P}[X_{n+1} = s_j \mid X_0 = s_{i_0}, X_1 = s_{i_1}, \cdots, X_{n-1} = s_{i_{n-1}}, X_n = s_i]$$

$$= \mathbb{P}[X_{n+1} = s_j \mid X_n = s_i] \leftarrow \text{Memorylessness property / Markov prop.}$$

$$= P_{ij} \leftarrow (i,j) \text{ entry of } P$$



<u>on edge wt.ed digraph on S</u>

State space

- $P_{ij} \in [0,1]$
- $\sum_{j=1}^{k} P_{ij} = 1 \quad \forall i$
- <u>Initial distribution</u> $\mu^{(0)} = (\mu_1^{(0)}, \ldots, \mu_k^{(0)}) = (\mathbb{P}[X_0 = s_1], \cdots, \mathbb{P}[X_0 = s_k])$

  <u>After n-steps</u> $\mu^{(n)} = (\mu_1^{(n)}, \ldots, \mu_k^{(n)}) = (\mathbb{P}[X_n = s_1], \cdots, \mathbb{P}[X_n = s_k])$

  $= \mu^{(0)} P^n$

Two fundamental conditions on Markov Chains

① MC is called <u>irreducible</u> if the corresponding transition digraph is strongly connected. (can go from each state to another w.s. positive probab.)

i.e., $\forall i,j \ \exists n \geq 0$ s.t. $\mathbb{P}[X_{m+n} = s_i \mid X_m = s_j] > 0$

s_i communicates with s_j, $s_i \leftrightarrow s_j$ if $\exists n \geq 0$ s.t. $[P^n]_{ij} > 0$

② <u>Period</u> $d(s_i)$ of state s_i is $\gcd\{n \geq 1 : [P^n]_{i,i} > 0\}$

If $d(s_i) = 1 \ \forall s_i$ then MC is called <u>aperiodic</u>.

Two fundamental conditions on Markov Chains

① MC is called <u>irreducible</u> if the corresponding transition digraph is strongly connected. (can go from each state to another w.s. positive probab.)

i.e., $\forall i,j \ \exists n \geq 0$ s.t. $\mathbb{P}[X_{m+n} = s_i \mid X_m = s_j] > 0$

s_i communicates with s_j, $s_i \leftrightarrow s_j$ if $\exists n \geq 0$ s.t. $[P^n]_{ij} > 0$

② <u>Period</u> $d(s_i)$ of state s_i is $\gcd\{n \geq 1 : [P^n]_{i,i} > 0\}$

If $d(s_i) = 1 \ \forall s_i$ then MC is called <u>aperiodic</u>.

irreducible & aperiodic

<u>Theorem</u> Let (X_n) be an <u>Ergodic</u> MC with state space $S = \{s_1, \dots, s_k\}$ and transition matrix P, then $\exists N < \infty$ s.t. $[P^n]_{ij} > 0 \ \forall i,j \in [k]$ & $\forall n \geq N$

How can we ensure our MC is ergodic?

→ strong connectivity → $P_{ii} > 0$

What happens as we run a MC for a long time?

$\pi = (\pi_1, \ldots, \pi_k)$ is s.t.b. a **stationary distribution** for the MC if

- $\pi_i \geq 0 \ \forall i$ and $\sum_{i=1}^{k} \pi_i = 1$
- $\pi P = \pi$, ie., $\sum_{i=1}^{k} \pi_i P_{ij} = \pi_j \ \forall j$

left eigenvector of P w.r.t. eigenvalue 1.

Note that if $\mu^{(0)} = \pi$ then $\mu^{(1)} = \mu^{(0)} P = \pi P = \pi$ & so on $\mu^{(n)} = \pi$.

Theorem For any Ergodic MC, $\exists$ unique stationary distribution

What happens as we run a MC for a long time?

$\pi = (\pi_1, \ldots, \pi_k)$ is s.t.b. a <u>stationary distribution</u> for the MC if

- $\pi_i \geq 0 \ \forall i$ and $\sum_{i=1}^{k} \pi_i = 1$
- $\pi P = \pi$, ie., $\sum_{i=1}^{k} \pi_i P_{ij} = \pi_j \ \forall j$

<u>$\pi P = \pi$</u>

left eigenvector of P w.r.t. eigenvalue 1.

Note that if $\mu^{(0)} = \pi$ then $\mu^{(1)} = \mu^{(0)} P = \pi P = \pi$ & so on $\mu^{(n)} = \pi$.

<u>Fundamental Theorem of Markov Chains</u> For any Ergodic MC, $\exists$ unique stationary distribution π

For any Ergodic MC starting from an arbitrary initial distribution $\mu^{(0)}$, $\mu^{(n)} \longrightarrow \pi$ as $n \to \infty$

What does this mean? How is "close to π" defined

Let $\nu = (\nu_1, \ldots, \nu_k)$ & $\mu = (\mu_1, \ldots, \mu_k)$ be two probability distributions on S then <u>Total variation distance</u> between ν and μ is:

$$\|\nu - \mu\|_{TV} = \frac{1}{2} \sum_{i=1}^{k} |\nu_i - \mu_i| = \max_{A \subseteq S} |\nu(A) - \mu(A)|$$

How to find the stationary distribution?

How to design a MC with appropriate stationary distribution?

A probability distribution π on S is said to be **reversible for P**

if $\forall i,j \in [k]$, $\pi_i P_{ij} = \pi_j P_{ji}$ ← at each step of MC $\mathbb{P}[i \to j] = \mathbb{P}[j \to i]$

MC looks same whether time runs forward or backward

MC is s.t.b. **reversible** if $\exists$ a reversible distribution for it.

How to find the stationary distribution?

How to design a MC with appropriate stationary distribution?

A probability distribution π on S is said to be <u>reversible for P</u>

if $\forall i,j \in [k]$, $\pi_i P_{ij} = \pi_j P_{ji}$ ← at each step of MC $\mathbb{P}[i \to j] = \mathbb{P}[j \to i]$

MC looks same whether time runs forward or backward

MC is s.t.b. <u>reversible</u> if $\exists$ a reversible distribution for it.

<u>Observation</u> If π is a reversible distribution for a MC, then it is also a stationary distribution for that MC.

$$\sum_{i=1}^{k} \pi_i P_{ij} = \sum_{i=1}^{k} \pi_j P_{ji} = \pi_j \sum_{i=1}^{k} P_{ji} = \pi_j \text{, i.e., } \pi P = \pi$$

If we design an Ergodic MC with a reversible distribution then we know the unique stationary distribution of the MC, and we can sample its states with that distribution.

Random walk on a finite graph

$S = V(G) = \{v_1, \dots, v_k\}$

Move from v_i to one of its neighboss uniformly at random $(d_i = \deg(v_i))$

$$P_{ij} = \begin{cases} 1/d_i & \text{if } j \in N(v_i) \\ 0 & \text{otherwise} \end{cases}$$

- P is irreducible $\Leftrightarrow$
- P is aperiodic $\Leftrightarrow$

Random walk on a finite graph

$S = V(G) = \{v_1, \dots, v_k\}$

Move from v_i to one of its neighboss uniformly at random $(d_i = \deg(v_i))$

$$P_{ij} = \begin{cases} 1/d_i & \text{if } j \in N(v_i) \\ 0 & \text{otherwise} \end{cases}$$

- P is irreducible $\Leftrightarrow$ G is connected
- P is aperiodic $\Leftrightarrow$ G is not bipartite
- P is reversible with distribution $\pi_i = d_i / 2|E(G)|$

Pf. $v_i v_j \notin E(G) \Rightarrow P_{ij} = P_{ji} = 0$

$$v_i v_j \in E(G) \Rightarrow \pi_i P_{ij} = \frac{d_i}{2|E(G)|} \frac{1}{d_i} = \frac{1}{2|E(G)|} = \frac{d_j}{2|E(G)|} \frac{1}{d_j} = \pi_j P_{ji}$$

For a non-bipartite connected graph G, this MC has a unique stationary distribution given by $\pi_i = \frac{d_i}{2|E(G)|} = \frac{d_i}{\sum_{j=1}^{k} d_j}$

What does it mean to run this MC on such graph?

Given a probability distribution π on $S = \{s_1, \dots, s_k\}$,
how do we sample a random object in S with distribution π?

Think of S as - set of all independent sets in a graph G
- set of all matchings in a graph G
- set of all q-colorings of a graph G, etc.

How can we design an ergodic MC that samples uniformly from S?

Given a probability distribution π on $S = \{s_1, \dots, s_k\}$,
how do we sample a random object in S with distribution π?

Think of S as - set of all independent sets in a graph G
- set of all matchings in a graph G
- set of all q-colorings of a graph G, etc.

How can we design an ergodic MC that samples uniformly from S?

Let $\mathcal{G}$ be the transition graph on S with

$$P_{ij} = \begin{cases} 1/\Delta(\mathcal{G}) & \text{if } v_iv_j \in E(\mathcal{G}) \\ 1 - d_i/\Delta(\mathcal{G}) & \text{if } v_i = v_j \\ 0 & \text{otherwise} \end{cases}$$

Given a probability distribution π on $S = \{s_1, \dots, s_k\}$,
how do we sample a random object in S with distribution π?

Think of S as - set of all independent sets in a graph G
- set of all matchings in a graph G
- set of all q-colorings of a graph G, etc.

How can we design an ergodic MC that samples uniformly from S?

Let $\mathcal{G}$ be the transition graph on S with

$$P_{ij} = \begin{cases} 1/\Delta(\mathcal{G}) & \text{if } v_iv_j \in E(\mathcal{G}) \\ 1 - d_i/\Delta(\mathcal{G}) & \text{if } v_i = v_j \\ 0 & \text{otherwise} \end{cases}$$

This MC is aperiodic, irreducible
and $\pi = \left(\frac{1}{|V(\mathcal{G})|}, \dots, \frac{1}{|V(\mathcal{G})|}\right)$ is a reversible distribution

$\therefore$ π = uniform distribution is the unique stationary distribution

(In fact, any symmetric P will give a unif. stationary dist.)

Lemma Let $S = \{s_1, \ldots, s_k\}$ be a finite state space and $\mathcal{G}$ be the transition graph for a Markov chain on S.
Let $\Delta(\mathcal{G}) = \max$ #nbhrs of a state $= \max_{i=1,\ldots,k} \{d_i\}$.
Let $M \geq \Delta(\mathcal{G})$. If $P_{ij} = \begin{cases} 1/M & \text{if } i \neq j \text{ and } s_i s_j \in E(\mathcal{G}) \\ 0 & \text{if } i \neq j \text{ and } s_i s_j \notin E(\mathcal{G}) \\ 1 - d_i/M & \text{if } i = j \end{cases}$
and this chain is irreducible and aperiodic,
then it has a unique stationary distribution which is uniform.

Pf.. Check $\pi_i P_{ij} = \pi_j P_{ji}$

Since $P_{ij} = P_{ji}$, any π with $\pi_i = \pi_j$ will give reversibility.
Let $\pi_i = \frac{1}{|S|} = \frac{1}{k}$. This works.
And by FTMC ($\because$ MC is ergodic), this π is unique.

Let $\mathcal{I}(G)$ = set of all independent sets of a graph G.

We want a MC over $\mathcal{I}(G)$, $(X_0, X_1, \ldots)$

Let $\mathcal{I}(G)$ = set of all independent sets of a graph G.

We want a MC over $\mathcal{I}(G)$, $(X_0, X_1, \dots)$

Let X_0 = arbitrary ind. set (say $\emptyset$ or single vertex)

At each time $n+1$

→ choose $v \in V(G)$ unif. at random

→ if $v \in X_n$ then $X_{n+1} = X_n \setminus \{v\}$

→ if $v \notin X_n$ and no nbhrs of v are in X_n, then $X_{n+1} = X_n \cup \{v\}$

→ Otherwise $X_{n+1} = X_n$

Let $\mathcal{I}(G)$ = set of all independent sets of a graph G.

We want a MC over $\mathcal{I}(G)$, $(X_0, X_1, \dots)$

Let X_0 = arbitrary ind. set (say $\emptyset$ or single vertex)

At each time $n+1$

→ choose $v \in V(G)$ unif. at random

→ if $v \in X_n$ then $X_{n+1} = X_n \setminus \{v\}$

→ if $v \notin X_n$ and no nbhrs of v are in X_n, then $X_{n+1} = X_n \cup \{v\}$

→ Otherwise $X_{n+1} = X_n$

Transition probabilities are defined implicitly by this randomized algorithmic process.

- In the transition graph, neighbors of a state I are all ind. sets I' s.t. I' differ from I in just 1 vertex.

Let $\mathcal{I}(G)$ = set of all independent sets of a graph G.

We want a MC over $\mathcal{I}(G)$, $(X_0, X_1, \ldots)$

Let X_0 = arbitrary ind. set (say ϕ or single vertex)

At each time $n+1$

→ choose $v \in V(G)$ unif. at random

→ if $v \in X_n$ then $X_{n+1} = X_n \setminus \{v\}$

→ if $v \notin X_n$ and no nbhrs of v are in X_n, then $X_{n+1} = X_n \cup \{v\}$

→ Otherwise $X_{n+1} = X_n$

- Irreducible: each state can reach and is reachable from $I = \phi$.
- Aperiodic: with positive probability pick a vertex $v \notin I$ and $v \in \bigcup_{w \in I} N(w)$ then $X_{n+1} = X_n = I$.
- P_{I_1, I_2} is $\neq 0$ if $I_1 = I_2$; $= 0$ if I_1 & I_2 differ by 2 or more vertices;
  $= 1/|V(G)|$ if I_1 & I_2 differ by exactly 1 vertex, say v.
  (so, v was picked with prob. $1/|V(G)|$ & added to I_2)

∴ P is symmetric & MC is reversible. Hence ! stationary distribution is uniform.

The previous MC on $\mathcal{I}(G)$ was an example of Gibbs Sampler.

Gibbs Samplers

To simulate probability distribution π on a state space of the form S^V i.e., functions $V \to S$ where S, V are finite sets.

(For us V = vertex set (or edge set), S = values attainable at each element of V,
e.g. $V = V(G)$ and $S = \{0,1\}$ can model any family of subsets of vertices, like $\mathcal{I}(G)$, etc.)

The previous MC on $\mathcal{G}(G)$ was an example of Gibbs Sampler.

Gibbs Samplers

To simulate probability distribution π on a state space of the form S^V i.e., functions $V \to S$ where S, V are finite sets.

(For us V = vertex set (or edge set), S = values attainable at each element of V,
e.g. $V = V(G)$ and $S = \{0,1\}$ can model any family of subsets of vertices, like $\mathcal{G}(G)$, etc.)

At time $n+1$:

1. Pick $v \in V$ (uniformly) at random
2. Pick $X_{n+1}(v)$ according to conditional π-distribution of the value at v given that all other vertices take values according to X_n.
3. Let $X_{n+1}(w) = X_n(w)$ for all $w \neq v$.

The previous MC on $\mathcal{G}(G)$ was an example of Gibbs Samplers.

Gibbs Samplers

To simulate probability distribution π on a state space of the form S^V i.e., functions $V \to S$ where S, V are finite sets.

(For us V = vertex set (or edge set), S = values attainable at each element of V,
e.g. $V = V(G)$ and $S = \{0,1\}$ can model any family of subsets of vertices, like $\mathcal{G}(G)$, etc.)

At time $n+1$:

1. Pick $v \in V$ (uniformly) at random
2. Pick $X_{n+1}(v)$ according to conditional π-distribution of the value at v given that all other vertices take values according to X_n.
3. Let $X_{n+1}(w) = X_n(w)$ for all $w \neq v$.

It can be shown: its aperiodic & has π as reversible distribution.
If it is also irreducible then π is the unique stationary distribution.

Gibbs sampler for proper q-colorings of graph G?

G be a graph, q ($\geq \chi(G)$) integer.

$\mathcal{C}_{G,q}$ = set of all proper q-colorings of G.

Let $V = V(G)$ and $S = \{1, 2, \cdots, q\}$. (MC on $\mathcal{C}_{G,q}$ takes values in S^V)

Gibbs sampler for proper q-colorings of graph G?

G be a graph, q ($\geq \chi(G)$) integer.

$\mathcal{C}_{G,q}$ = set of all proper q-colorings of G.

Let $V = V(G)$ and $S = \{1, 2, \dots, q\}$. (MC on $\mathcal{C}_{G,q}$ takes values in S^V)

At time $n+1$:

① Pick $v \in V$ unif. at random

② Pick $X_{n+1}(v)$

Gibbs sampler for proper q-colorings of graph G?

G be a graph, q ($\geq \chi(G)$) integer.

$\mathcal{C}_{G,q}$ = set of all proper q-colorings of G.

Let $V = V(G)$ and $S = \{1, 2, \dots, q\}$. (MC on $\mathcal{C}_{G,q}$ takes values in S^V)

At time $n+1$:

① Pick $v \in V$ unif. at random

② Pick $X_{n+1}(v)$ according to unif. dist. over colors that are not assigned to the neighbors of v.

③ $X_{n+1}(w) = X_n(w)$ $\forall w \neq v$.

- Color of <u>atmost</u> one vertex (v) changes between X_n & X_{n+1}
- Aperiodic: positive probability coloring remains unchanged. (Why?)
- $P_{\sigma_1,\sigma_2} = 1/|V|$ if $\sigma_1 \neq \sigma_2$ (proper q-colorings) and $|\sigma_1 - \sigma_2| = 1$

∴ MC is reversible and stationary dist. is uniform on $\mathcal{C}_{G,q}$.

Irreducibility? depends on G & q, and non-trivial to determine. Known if $q \geq \Delta(G)+2$.

Metropolis Process

Other standard way to create MC with desired stationary distribution.

Want distribution π on state space S with $\pi(s) = w(s)/Z$ (Z is the normalizing constant — often unknown)

- Define a "neighborhood structure" on S, an undirected connected graph G on S.
- For each $s \in S$, define a "proposal distribution" $K(s, \cdot)$ where $K(s,t) = \mathbb{P}[\text{choose a neighbor } t \text{ of } s]$, e.g. $K(s,t) = \frac{1}{d(s)}$ $\forall t \in N(s)$, with $K(s,t) = K(t,s)$; $K(s,s) = 1 - \sum_{t \in N(s)} K(s,t)$; $K(s,t) = 0$ if $d(s,t) > 1$.

Metropolis Process

Other standard way to create MC with desired stationary distribution

Want distribution π on state space S with $\pi(s) = w(s)/Z$ (Z is the normalizing constant — often unknown)

- Define a "neighborhood structure" on S, an undirected connected graph G on S.
- For each $s \in S$, define a "proposal distribution" $K(s, \cdot)$ where $K(s,t) = \mathbb{P}[\text{choose a neighbor } t \text{ of } s]$, e.g. $K(s,t) = \frac{1}{d(s)}$ $\forall t \in N(s)$, with $K(s,t) = K(t,s)$; $K(s,s) = 1 - \sum_{t \in N(s)} K(s,t)$; $K(s,t) = 0$ if $d(s,t) > 1$.

At each step (current state is s):

① "Proposal step" Pick a neighbor t of s with probab. $K(s,t)$

② "move" Move to t with probability $\min\left(1, \frac{w(t)}{w(s)}\right)$, else stay at s

Metropolis Process

Other standard way to create MC with desired stationary distribution.

Want distribution π on state space S with $\pi(s) = w(s)/Z$ (Z is the normalizing constant — often unknown)

- Define a "neighborhood structure" on S, an undirected connected graph G on S.
- For each $s \in S$, define a "proposal distribution" $K(s, \cdot)$ where $K(s,t) = \mathbb{P}[\text{choose a neighbor } t \text{ of } s]$, e.g. $K(s,t) = \frac{1}{d(s)}$ $\forall t \in N(s)$, with $K(s,t) = K(t,s)$; $K(s,s) = 1 - \sum_{t \in N(s)} K(s,t)$; $K(s,t) = 0$ if $d(s,t) > 1$.

At each step (current state is s):

① "Proposal step" Pick a neighbor t of s with probab. $K(s,t)$

② "move" Move to t with probability $\min\left(1, \frac{w(t)}{w(s)}\right)$, else stay at s

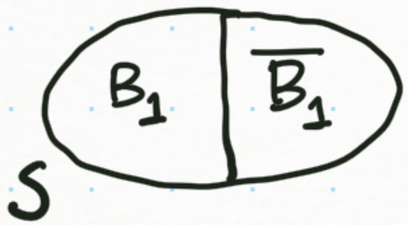
- Irreducible (G is connected)
- Aperiodic (ensure $K(s,s) > 0$)
- reversible with $\pi(s) = w(s)/Z$

Key Don't need to know Z, only need $\lambda = w(t)/w(s)$.

Simulated Annealing: λ varies with time $\left(\frac{w(t)}{w(s)}\right)^{1/\tau}$ where τ is "temperature" that gradually dec. to 0.

How does sampling with these MC help in counting?

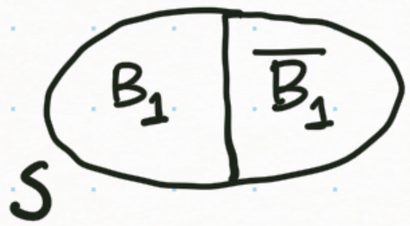
To count S, split S into 2 (roughly equal) parts based on some easy to verify property (say, S = population, property $\equiv$ black hair or not)



Pick a uniform random sample from S, ← Markov Chain
to get an estimate $\hat{p}_1$ of the ratio $\frac{|B_1|}{|S|} = p_1$ ← Monte Carlo

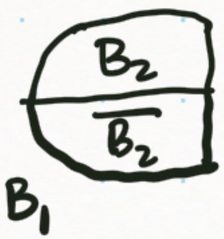
How does sampling with these MC help in counting?

To count S, split S into 2 (roughly equal) parts based on some easy to verify property (say, S = population, property $\equiv$ black hair or not)



Pick a uniform random sample from S, ← Markov Chain
to get an estimate $\hat{p}_1$ of the ratio $\frac{|B_1|}{|S|} = p_1$ ← Monte Carlo

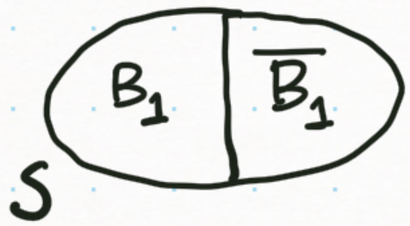
then



Pick a uniform random sample from B_1,
to get an estimate $\hat{p}_2$ of the ratio $\frac{|B_2|}{|B_1|} = p_2$
& so on.....

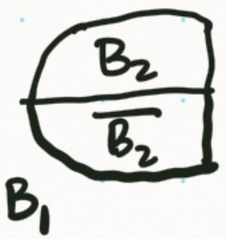
How does sampling with these MC help in counting?

To count S, split S into 2 (roughly equal) parts based on some easy to verify property (say, S = population, property $\equiv$ black hair or not)



Pick a uniform random sample from S, ← Markov Chain
to get an estimate $\hat{P}_1$ of the ratio $\frac{|B_1|}{|S|} = P_1$ ← Monte Carlo

then



Pick a uniform random sample from B_1,
to get an estimate $\hat{P}_2$ of the ratio $\frac{|B_2|}{|B_1|} = P_2$
& so on.....

$$|S| = \frac{|S|}{|B_1|}\frac{|B_1|}{|B_2|}\frac{|B_2|}{|B_3|}\cdots\frac{|B_{R-1}|}{|B_R|}|B_R| = |B_R|/P_1P_2\cdots P_R \simeq |B_R|/\hat{P}_1\hat{P}_2\cdots\hat{P}_R$$

where $R \simeq \log|S|$ if we assume each criteria splits the population. (R = # stages)

Two Questions

① How "quickly" can we sample using a Markov Chain?

② How can we formalize the counting algo based on Monte Carlo?

Approximation (counting) Algorithms

A counting problem can be viewed as computing a function $f : \Sigma^* \to \mathbb{N}$ (where Σ^* is set of all instances based on finite alphabet Σ used to encode problem instances)

Fully Polynomial Randomized Approximation Scheme (FPRAS)

Given input $x \in \Sigma^*$, error parameter $\epsilon > 0$, a randomized approx. scheme for f is a randomized algorithm that outputs z s.t. $\mathbb{P}[(1-\epsilon)f(x) \leq z \leq (1+\epsilon)f(x)] \geq 3/4$

It is called fully polynomial if its running time is a polynomial in $|x|$, and $1/\epsilon$.

Approximation (counting) Algorithms

A counting problem can be viewed as computing a function $f : \Sigma^* \to \mathbb{N}$ (where Σ^* is set of all instances based on finite alphabet Σ used to encode problem instances)

Fully Polynomial Randomized Approximation Scheme (FPRAS)

Given input $x \in \Sigma^*$, error parameter $\epsilon > 0$, a randomized approx. scheme for f is a randomized algorithm that outputs z s.t. $\mathbb{P}[(1-\epsilon)f(x) \leq z \leq (1+\epsilon)f(x)] \geq 3/4$

It is called fully polynomial if its running time is a polynomial in $|x|$, and $1/\epsilon$.

Lemma Any FPRAS can be extended to an algorithm that take input $\delta \in (0,1]$ s.t. $\mathbb{P}[z$ is within $(1 \pm \epsilon)f(x)] \geq 1-\delta$ with running time inc. by factor $O(\log 1/\delta)$

Proof Take $k = 16 \log(3/\delta)$ trials of original scheme and output the median.
$\mathbb{P}[$median outside $(1 \pm \epsilon)f(x)] \leq \ldots \leq 2e^{-k^2/16k} \leq \delta$ using C-H bounds.

Approximate Uniform Sampling

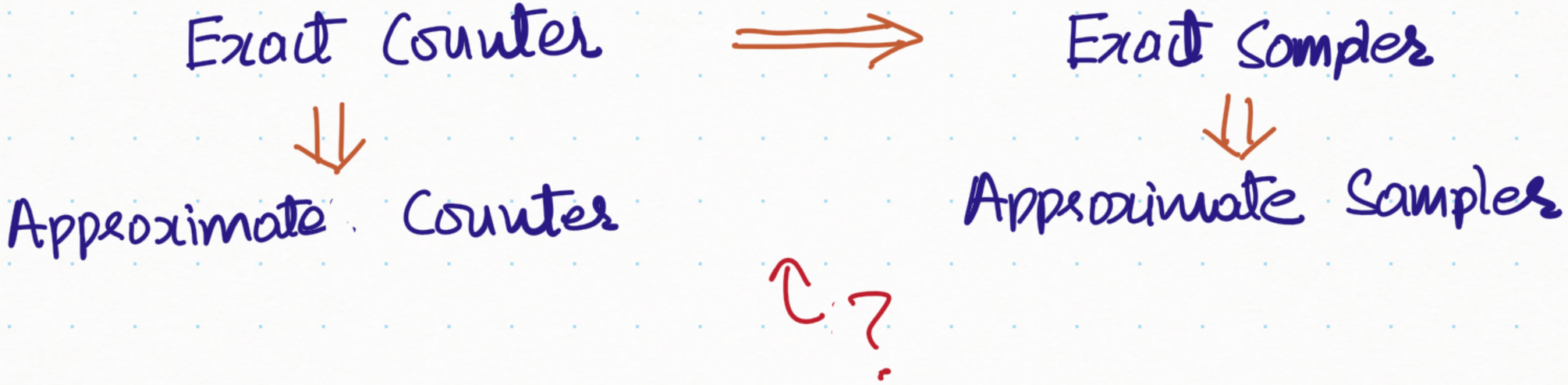
Given an instance $x \in \Sigma^*$, a sampling problem wants to output from a distribution π over $\Omega(x)$, the solutions to x.

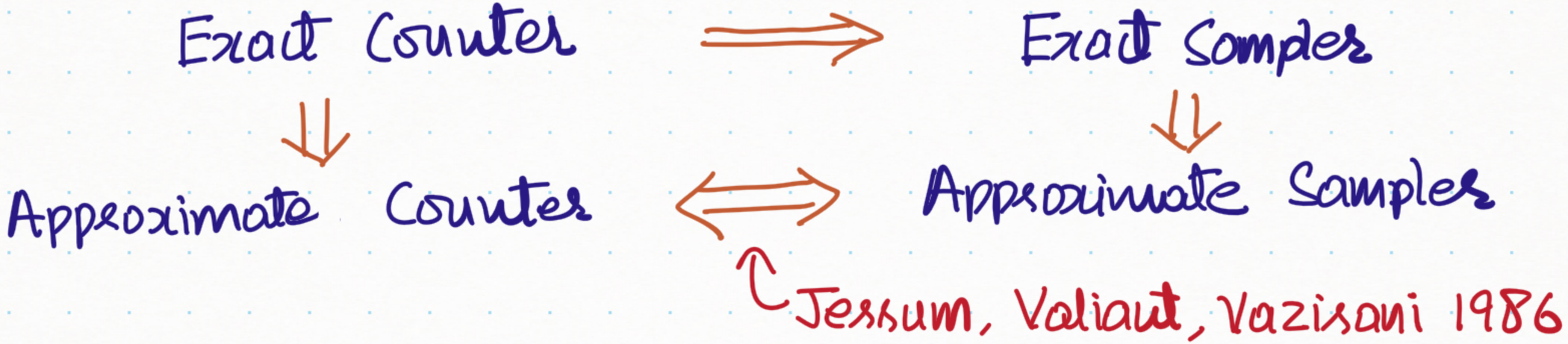
A sampling algorithm gives an ϵ-uniform sample from Ω if

$$\left| \mathbb{P}(\omega \in S) - \frac{|S|}{|\Omega|} \right| \leq \epsilon \qquad \forall S \subseteq \Omega$$

A sampling algorithm is **Fully Polynomial Almost Uniform Sampler** (FPAUS) if given an input x and $\epsilon \in (0,1)$, it generates an ϵ-uniform sample of $\Omega(x)$, sample space for x, in time that is polynomial in $|x|$ and $\log(1/\epsilon)$.

compare to running time for FPRAS


Exact Counter
Exact Sampler
Approximate Counter
Approximate Samplers
?



Theorem [Jerrum, Valiant, Vazirani, 1986]

For any "self-reducible" counting problem, there exists a FPRAS iff there exists a FPAUS.

A self-reducible problem is one whose set of feasible solutions can be put in one-to-one correspondence with the disjoint union of the solution sets of at most polynomially many sub-problems of smaller size. Most counting problems satisfy this.

Using a FPAUS for proper q-colorings of G, we get a FPRAS

Let $\Omega(G)$ = set of all proper q-colorings of G

Let $E(G) = \{e_1, \dots, e_m\}$

Define $G_i = (V(G), \{e_1, \dots, e_i\})$ for $i = 0, 1, \dots, m$

$$|\Omega(G)| = |\Omega(G_m)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \cdots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|$$

$|\Omega(G_0)| = ?$

Using a FPAUS for proper q-colorings of G, we get a FPRAS

Let $\Omega(G)$ = set of all proper q-colorings of G

Let $E(G) = \{e_1, \dots, e_m\}$

Define $G_i = (V(G), \{e_1, \dots, e_i\})$ for $i = 0, 1, \dots, m$

$$|\Omega(G)| = |\Omega(G_m)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \cdots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|$$

$|\Omega(G_0)| = q^n$, G_0 is empty graph so all colorings are proper.

We could estimate $|\Omega(G)|$ in poly-time if we can estimate each $\rho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$, $i = 1, \dots, m$ ($\leq n^2$) (& then $|\Omega(G)| = \rho q^n$ where $\rho = \prod_{i=1}^{m} \rho_i$)

Using a FPAUS for proper q-colorings of G, we get a FPRAS

Let $\Omega(G)$ = set of all proper q-colorings of G

Let $E(G) = \{e_1, \dots, e_m\}$

Define $G_i = (V(G), \{e_1, \dots, e_i\})$ for $i = 0, 1, \dots, m$

$$|\Omega(G)| = |\Omega(G_m)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \cdots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|$$

$|\Omega(G_0)| = q^n$, G_0 is empty graph so all colorings are proper.

We could estimate $|\Omega(G)|$ in poly-time if we can estimate each $\rho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$, $i = 1, \dots, m$ ($\leq n^2$) (& then $|\Omega(G)| = \rho q^n$ where $\rho = \prod_{i=1}^{m} \rho_i$)

Note $\Omega(G_i) \subseteq \Omega(G_{i-1})$, so ρ_i can be estimated by sampling from $\Omega(G_{i-1})$ and calculating the proportion of samples that fall in $\Omega(G_i)$.

Sampling from $\Omega(G_{i-1})$ To generate each sample,
1 Run the Markov chain (Gibbs samples for proper q-coloring) until its within ϵ-TV-distance of its stationary distribution (which is uniform distribution).

Using the theory of mixing-time of Markov Chains, it can be shown: This MC samples q-coloring almost uniformly at random in time $O(n \log n + n \log 1/\epsilon)$ where n = #vertices, ϵ = TV distance

Monte Carlo estimate of $\rho_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}$ where $\Omega(G_i) \subseteq \Omega(G_{i-1})$

$\hat{\rho}_i = \frac{\#\text{samples from } \Omega(G_{i-1}) \text{ that fall in } \Omega(G_i)}{\#\text{total samples from } \Omega(G_{i-1})}$ estimates ρ_i

Easy way to check the count in the numerator:

let $e_i = uv$ be the edge in G_i that is not in G_{i-1}, for each sample from $\Omega(G_{i-1})$, we only need to check whether u & v are colored differently.

Using Chebyshev ineq. (or Chernoff bound) and the union bound, it can be shown that (when $q \geq 2D(n)+1$)

$$\mathbb{P}\left[\left|\hat{\rho}_i - \rho_i\right| \geq \frac{\epsilon}{2m}\rho_i\right] \leq \frac{1}{4m} \quad \text{if} \quad t \geq \frac{16m^3}{3\epsilon^2}$$

, the #samples needed to achieve this accuracy per ratio ρ_i

This gives a total running time of

$$O\left(\underbrace{m}_{\#\text{ratios}} \times \underbrace{\frac{m^3}{\epsilon^2}}_{\#\text{samples needed per ratio}} \times \underbrace{\left(n \log n + n \log 1/\epsilon\right)}_{\text{sampling time of the Markov Ch. per sample.}}\right)$$

Outline of an MCMC algorithm

Step1 Create a Markov Chain over the state space $\Omega(G)$

Two standard ways are Gibbs sampler and Metropolis process.

Ensure your MC is Ergodic.

Aperiodicity is easy to achieve by adding a loop $P_{ii} > 0$

Irreducibility might be trickier (e.g. for the Gibbs sampler for proper q-colorings)

And your MC will be reversible, ensuring a unique stationary distribution that is uniform on $\Omega(G)$.

Outline of an MCMC algorithm

Step1 Create a Markov Chain over the state space $\Omega(G)$

Step 2. Analyze the MC from step1 to show that it only needs polynomial number of steps (in $n=|V(G)|$ and $1/\epsilon$) to be within ϵ-TV-distance of the stationary distribution.

This requires use of techniques from MC Theory like:

— stochastic methods like coupling, coupling from past.

— linear algebraic methods like spectral (eigenvalue) bounds.

— combinatorial methods like Congestion / Expansion.

Outline of an MCMC algorithm

Step 1 Create a Markov Chain over the state space $\Omega(G)$

Step 2. Analyze the MC from step 1 to show that it only needs polynomial number of steps (in $n = |V(G)|$ and $1/\epsilon$) to be within ϵ-TV-distance of the stationary distribution.

Step 3 Set up the FPRAS (the approx counting algorithm) by describing $|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \cdots \frac{|\Omega(G_2)|}{|\Omega(G_1)|} |\Omega(G_1)|$

a telescoping product whose last term is easily known.

and the total #factors (m) must be polynomial in n (size of problem)

This may not always be easy.

Outline of an MCMC algorithm

Step1 Create a Markov Chain over the state space $\Omega(G)$

Step 2. Analyze the MC from step1 to show that it only needs polynomial numbers of steps (in $n=|V(G)|$ and $1/\epsilon$) to be within ϵ-TV-distance of the stationary distribution.

Step3 Set up the FPRAS (the approx counting algorithm) by describing $|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \cdots \frac{|\Omega(G_1)|}{|\Omega(G_0)|} |\Omega(G_0)|$

a telescoping product whose last term is easily known.

and the total #factors (m) must be polynomial in n (size of problem)

Step4 Approximate $|\Omega(G)|$ by estimating each ratio using the Monte Carlo principle with samples generated by Markov Chain in step 1. Combine the total running time from previous steps using Chebyshev or Chernoff.

Most papers on MCMC published in Statistical Physics and in Combinatorics/Theoretical Computer Sc.

Useful to have a dictionary of terms between the two fields

Statistical Physics	Discrete Math
• monomer–dimer coverings	matchings
• dimer coverings	perfect matchings
• hardcore lattice gas model	independent sets
• spin	bit
• ground states of the Potts model	vertex colorings
• partition function	normalizing constant
• connectivity	degree
• activity / fugacity	vertex weight
• interaction	edge weight
• ferromagnetism	positive correlation
• polynomial time mixing	rapid mixing
• rapid mixing	$O(n \log n)$ time mixing

Mixing Time of Markov Chains

$\Delta_{s_0}(n) = \| M_{s_0}^{(n)} - \pi \|_{TV}$ distance of distribution $M_{s_0}^{(n)}$ (n steps after starting from $s_0 \in S$) from the stationary distribution π

$\Delta(n) = \max_{s_0 \in S} \Delta_{s_0}(n)$ worst possible distance from π over all starting pts

$\tau_{s_0}(\epsilon) = \min \{ n : \Delta_{s_0}^{(n)} \leq \epsilon \}$ min time after which $M_{s_0}^{(n)}$ is ϵ-close to π

Mixing time: $\tau(\epsilon) = \max_{s_0 \in S} \tau_{s_0}(\epsilon)$, worst time to be ϵ-close to π

Definition $\tau_{mix} = \tau(1/2e)$

↑ any $\epsilon < 1/2$ would work

Lemma $\tau(\epsilon) \leq \tau_{mix} \lceil \log 1/\epsilon \rceil$

How to bound τ_{mix}? Running-time for FPAUS based on a Markov Chain

Spectral Bound on Mixing time

Let P be the transition matrix of a MC on Ω with $|\Omega| = N$ and stationary distribution π.

Perron-Frobenius Theory of non-neg. matrices

- If P is <u>reversible</u> then all its eigenvalues are real

- If P is reversible & <u>irreducible</u> then its eigenvalues are

$$1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} \geq -1$$

If P is also <u>aperiodic</u> then $\lambda_{N-1} > -1$

- If P is <u>lazy</u>, i.e., $P_{i,i} \geq 1/2 \ \forall i$, then $\lambda_{N-1} \geq 0$

<u>Theorem</u> Let $\pi_{min} = \min_{s_i \in \Omega} \pi(s_i)$

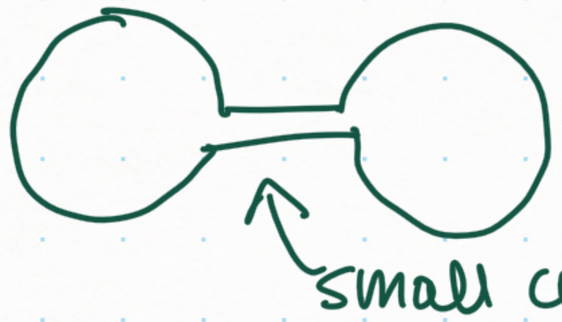
For all $\epsilon > 0$,

$$\frac{1}{1-|\lambda_1|} \log\left(\frac{1}{\epsilon \pi_{min}}\right) \geq \tau(\epsilon) \geq \frac{|\lambda_1|}{2(1-|\lambda_1|)} \log\left(\frac{1}{2\epsilon}\right)$$

Hard to use because N is exponential in size of problem & P is $N \times N$.

~

Conductance



Transition graph $\mathcal{G}$ on the state space

Intuition $\exists$ small cut in the $\mathcal{G} \iff$ Slow mixing time

Made precise using:

Conductance $\Phi = \min_{S \subseteq \Omega} \Phi(S)$

where $\Phi(S)$ is the conditional probab. that MC in equilibrium escapes from S in one step, given it's in S.

"large" Φ means there is "no small" cut in $\mathcal{G}$.

Theorem $\tau(\epsilon) \leq \frac{2}{\Phi^2} \log\left(\frac{1}{\pi_{min}\epsilon}\right)$

Need good lower bounds on Φ

Cheeger bounds $\frac{\Phi^2}{2} \leq 1 - |\lambda_1| \leq 2\Phi$