Algorithmic Applications of Markov Chains.

Till Now: Basics of Markov Chains.

- Main thm: $\exists$ and unicity of a stat distr.
- Main thm: when distr a limiting distr?
- Notion of Ergodic chain
  $\uparrow$
  irred, aperiodic, pos rec.
- Rate of approach of the stat distr.

  mixing time, spectral gap.

  within of chains that satisfy detailed balance.

All this has to be kept in Mind as a framework for the second part of the course.
$\downarrow$
Appl & Alg.

Algorithmic Question that we will tackle is about <u>SAMPLING</u>. from a distribution.

$$\underline{\underline{\{\pi_i\}}}_{i \in S} \qquad S = \text{state space.}$$

How do you sample efficiently?

<u>Today's program.</u>

1) Motivate this question. (a little). ✓

2) Remind some very classical methods for sampling that work well for "easy" distr. ✓

3) Examples of "<u>hard</u>" to sample distributions. ←

4) <u>Markov Chain Monte Carlo method</u>.

↓

Metropolis-Hasting algo.

In the coming weeks:

Next Lect
1) Metropolis-Hasting to minimization of function or a cost fct.
2) Annealing Algor.

1 weeks
3) Coloring a graph: analyze MCMC to color a graph.

1/2 weeks.
4) Ising Model: paradigm of "hard" to sample distr

2 weeks
5) last two weeks: Propp & Wilson Coupling from the past Method.

special implementation of MCMC.

H.

Motivation for sampling from a distribution π_i, $i \in S$.

- For example you want to compute: $\sum_{i \in S} f(i)\, \pi_i$.

$$= \mathbb{E}(f(X)).$$

X is r.v s.t $\mathbb{P}(X = i) = \pi_i$.

Monte Carlo Method: you draw M samples

$X_1, X_2, \ldots, X_M$ i.i.d from π.

Take the estimator $\frac{1}{M} \sum_{k=1}^{M} f(X_k)$.

By the law of large nbs: $\frac{1}{M} \sum_{k=1}^{M} f(X_k) \xrightarrow[M \to +\infty]{} \mathbb{E}(f(X))$.

Variance $\mathrm{Var}\left\{ \frac{1}{M} \sum_{k=1}^{M} f(X_k) \right\} = \frac{1}{M^2} \sum_{k=1}^{M} \mathrm{Var}(f(X_k))$

$$= \frac{1}{M} \mathrm{Var}(f(X)).$$

$$\frac{1}{M} \sum_{k=1}^{M} f(X_k) = \mathbb{E}(f(X)) + \underbrace{O\left(\frac{1}{\sqrt{M}}\right)}_{\text{error of estimator.}}$$

$\frac{1}{\sqrt{M}} \sqrt{\mathrm{Var}(f(X))} \leftarrow$ estimator.

Classical Sampling Methods: (easy to sample distr).

- Most Naive one but very much used (in practice on computer):

Hyp: Efficient way to generate a $U \sim \text{Unif}[0,1]$.

To sample from $(\pi_i)_{i \in S}$ $S = \{0, 1, 2, \dots\}$.

$$X = \begin{cases} 0 \; ; & 0 \leq U \leq \pi_0 \; ; \quad \mathbb{P}(U=0) = \pi_0 = \mathbb{P}(X=0) \\ 1 \; ; & \pi_0 \leq U \leq \pi_0 + \pi_1 \; ; \quad \mathbb{P}(U=1) = \pi_1 = \mathbb{P}(X=1) \\ \vdots & \\ i \; ; & \sum_{k=0}^{i-1} \pi_k \leq U \leq \sum_{k=0}^{i} \pi_k \; ; \quad \mathbb{P}(U=i) = \pi_i \\ & = \mathbb{P}(X=i). \end{cases}$$

"U Acts a dice"

- Importance Sampling.

Assume we have an efficient method to sample from distr $\{\psi_i\}_{i\in S}$.

<u>Remark</u>: $\underbrace{\sum_{i\in S} f(i)\,\pi_i}_{\mathbb{E}_{X\sim\pi}(f(X))} = \underbrace{\sum_{i\in S} f(i)\, w_i\, \psi_i}_{\mathbb{E}_{X\sim\psi}(f(X)\, W(X))}$ ✓

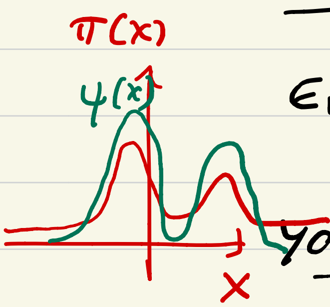
$\boxed{\text{where } w_i = \frac{\pi_i}{\psi_i}}$

<u>Idea</u> is to consider the estimator:

$$\frac{1}{M}\sum_{k=1}^{M} f(X_k)\,W(X_k) \qquad \text{where } W_i = \frac{\pi_i}{\psi_i}$$

and the sum is over sample i.i.d $X_k \sim \psi$.

<u>Variance</u> $= \frac{1}{M}\,\mathrm{Var}(f(X)W(X))$.



Error is order $\frac{1}{\sqrt{M}}\cdot\sqrt{\mathrm{Var}(f(X)W(X))}$ and <u>you could optimize over ψ to make it smaller</u> <u>than</u> for previous method.

• Rejection sampling $\nearrow$ proceed as in importance sampling with some acceptance/rejection prob.

$$\underbrace{\sum_{i \in S} f(i)\, \pi_i}_{\mathbb{E}_\pi(f(x))} = \frac{\sum_{i \in S} f(i)\, \tilde{w}_i\, \psi_i}{1/c} \quad \checkmark$$

$$\tilde{w}_i = \frac{1}{c} \frac{\pi_i}{\psi_i}$$

• easy to sample from distr ψ.

• pick a sample i with probability ψ_i.

• accept the sample i with probability $\tilde{w}_i = \frac{1}{c} \frac{\pi_i}{\psi_i}$

here $c \geq \max_{i \in S} \left(\frac{\pi_i}{\psi_i} \right)$.

Estimator: $\displaystyle \frac{1}{M'} \sum_{k:\, X_k \text{ is accepted}} f(X_k)$ ✓

# of accepted samples

$X_1, \dots, X_M$ are samples from ψ. You accept in total M' of them.

$M > M'$

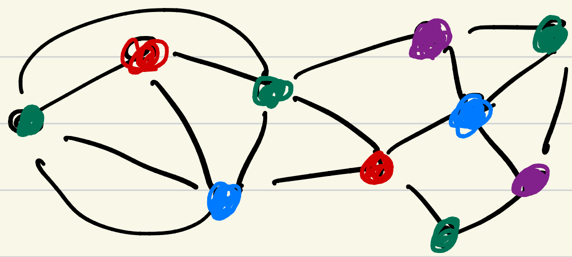
$$\frac{\sum_{i \in S} f(i)\, \psi_i \tilde{w}_i}{\sum_{i \in S} \psi_i \tilde{w}_i} = \frac{\sum_{i \in S} f(i)\, \psi_i \tilde{w}_i}{1/c} = \sum_{i \in S} f(i)\, \pi_i = \mathbb{E}_\pi(f(x))$$

$\psi_i \cdot \tilde{w}_i = \frac{1}{c} \pi_i \implies \sum_{i \in S} \psi_i \cdot \tilde{w}_i = \frac{1}{c}$

3) Examples of hard to sample distributions.

- Graph theory or theoretical computer science:

$\boxed{\text{Coloring an arbitrary large graph.}}$



$G = (V, E)$.

q colors at disposal $\{1, 2, \dots q\}$.

Def: Proper Coloring of G is an assignment of colors to V s.t. if $(a, b) \in E$ then a & b dont have the same color.

Take the distribution: π = unif distr over the set of all possible proper colorings.

S

$$\pi_{\text{proper col}} = \frac{\mathbb{1}(\text{proper col})}{Z}$$

$Z = \#(\text{of proper colorings})$

$=$ number of proper colorings.

- you dont know the set S
- you dont know Z the normalisation factor is unkown. (specially for large G.).

- **Ising Model.** (we will come back to this in a few weeks).

$G = (V, E)$ r.v assigned to $i \in V$: $s_i = \pm 1$

Cost function : $H(s_1 \cdots s_{|V|}) = - \sum_{(i,j) \in E} J_{ij} \, s_i s_j$

sum is over all Edges of G. $\in \mathbb{R}^*$.

distr : Ising Model distr (MRF or Gibbs distr for finite G).

$$\pi(s_1 \ldots s_{|V|}) = \frac{e^{-\beta H(s_1 \cdots s_{|V|})}}{Z}$$

state space = set of all binary assignments $(s_1 \cdots s_V)$

$$Z = \underbrace{\sum_{s_1 \cdots s_{|V|} \in \{\pm 1\}^{|V|}}}_{\text{sum contains } 2^{|V|} \text{ terms.}} e^{-\beta H(s_1 \cdots s_{|V|})}.$$

hard to compute.

Markov chain Monte Carlo (MCMC) Sampling Method.

- Goal is to sample π

Idea: view π as the stationary distribution and even in fact the limiting distribution of a Markov chain!

Given π, we will construct:

- Markov chain s.t π is the unique limiting distribution.

- Convergence rate (mixing time or spectral gap) of MC to π should be fast.

For these reasons almost always one considers MC that is ergodic and satisfies detailed balance

But also the construction should be such that we don't need to known too much about computing whatever is hard in π_i's.

[in the 40's group in Los Alamos found a way to achieve this construction.]

General MCMC algorithm.

(i.e. S state space)

It constructs of MC that has π as limiting distr:

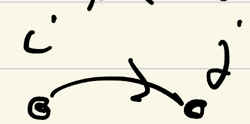
1. Select an easy-to-simulate Markov Chain with transition probabilities ψ_{ij} for $i,j \in S$. <u>We require</u> that the chain ψ is aperiodic & irreducible. <u>We also require</u> that $\psi_{ij} > 0 \iff \psi_{ji} > 0$.

ψ is called the base chain or the proposal chain

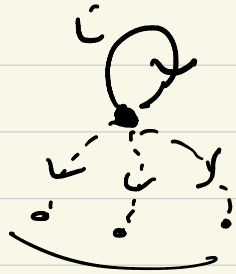
2. Design acceptance probabilities a_{ij} = the prob that the transition $i \to j$ of the base chain or proposal chain is accepted.

3. Construct the new chain (MCMC) with transition probability matrix $\mathbb{P}$:

accepted move.



$$\begin{cases} P_{ij} = \psi_{ij}\, a_{ij} & i \neq j \\ P_{ii} = 1 - \sum_{k \neq i} \psi_{ik}\, a_{ik} \checkmark \end{cases}$$

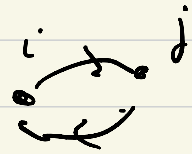


rejected move.

<u>Remark</u>:

$$P_{ii} = \psi_{ii} + \sum_{k \neq i} \psi_{ik}(1 - a_{ik}) = 1 - \sum_{k \neq i} \psi_{ik} a_{ik}.$$

We have not specified a_{ij} yet.



Theorem: [Metropolis-Hasting].

Set $$a_{ij} = \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right) \qquad \circledast$$

Then the chain with matrix P (above) is ergodic with stat distr π.

Proof: q irred and $q_{ij} > 0 \iff q_{ji} > 0$. Thus $a_{ij} > 0$ and also $p_{ij} = q_{ij} a_{ij} > 0$.

$\Rightarrow$ chain P is irred.

Moreover $p_{ii} > 0$ for some i $\Rightarrow$ P aperiodic.

One can check that $\boxed{\text{detailed balance is satisfied}}$ (with Metro Hast rule)

$\Downarrow$

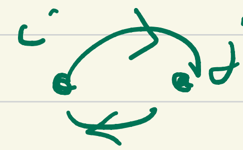
π is the stat distr. (so it exists).

$\Rightarrow$ P is pos recurrent.

Therefore P irred, aper, pos rec [ERGODIC]; limiting stat distr unique; and we know it's π. $\blacksquare$

Check that detailed balance is satisfied:

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



$$\pi_i P_{ij} = \pi_i q_{ij} \min\left(1 ; \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right)$$

Metro-Hast rule

$$= \underbrace{\min\left(\pi_i q_{ij} ; \pi_j q_{ji}\right)}_{\text{symmetric under } i \leftrightarrow j}$$

$$= \pi_j P_{ji}$$

∎

① **Remark**: Why is this a nice rule Metropolis-Hastings and why does it help to sample from "hard" dist.

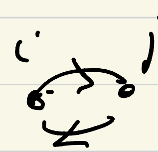
$$\frac{\mathbb{1}(\text{proper coloring})}{Z} \quad ; \quad \frac{e^{-\beta \sum_{ij\in E} J_{ij} s_i s_j}}{Z}$$

in $$a_{ij} = \min\left(1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right)$$

The Z will always simplify in $\frac{\pi_j}{\pi_i}$

we never have here to compute Z.

Also we will see that $\frac{\pi_j}{\pi_i}$ often can be computed very easily. (even by hand) sometimes.



② **Remark**: Interpretation of the rule. Suppose $q_{ij}=q_{ji}$

$$\Rightarrow a_{ij} = \min\left(1; \frac{\pi_j}{\pi_i}\right) = \begin{cases} 1 & \text{if } \pi_j \geq \pi_i \\ \frac{\pi_j}{\pi_i} & \text{if } \pi_j < \pi_i \end{cases}$$

Metropolis original rule.

state j is more probable than i so if we start at i we should certainly go to j.

But sometimes you should move to lower probability state otherwise you could be stuck.