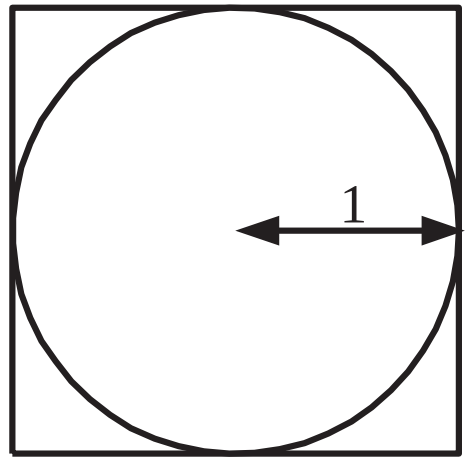
The Monte Carlo Method

- Estimating through sampling (estimating π, p-value, integrals,...)
- The main difficulty - sampling sparse events
- The general sampling to counting reduction
- The Markov Chain Monte Carlo (MCMC) method - Metropolis Algorithm
- Convergence rate
  - Coupling
  - Path coupling
  - Eigenvalues and conductance

The Monte Carlo Method

Example: estimate the value of π.



- Choose X and Y independently and uniformly at random in $[0, 1]$.
- Let

$$Z = \begin{cases} 1 & \text{if } \sqrt{X^2 + Y^2} \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

- $\Pr(Z = 1) = \frac{\pi}{4}$.
- $4\mathbf{E}[Z] = \pi$.

- Let $Z_1, \ldots, Z_m$ be the values of m independent experiments. $W = \sum_{i=1}^{m} Z_i$.
- $$\mathbf{E}[W] = \mathbf{E}\left[\sum_{i=1}^{m} Z_i\right] = \sum_{i=1}^{m} \mathbf{E}[Z_i] = \frac{m\pi}{4},$$
- $W' = \frac{4}{m} W$ is an unbiased estimate for π.
- $$\begin{aligned} \Pr(|W' - \pi| \geq \epsilon\pi) &= \Pr\left(|W - \frac{m\pi}{4}| \geq \frac{\epsilon m\pi}{4}\right) \\ &= \Pr\left(|W - \mathbf{E}[W]| \geq \epsilon \mathbf{E}[W]\right) \\ &\leq 2\mathrm{e}^{-\frac{1}{12} m\pi\epsilon^2}. \end{aligned}$$

(ϵ, δ)-Approximation

Definition

A randomized algorithm gives an (ϵ, δ)-*approximation* for the value V if the output X of the algorithm satisfies

$$\Pr(|X - V| \leq \epsilon V) \geq 1 - \delta.$$

Theorem

Let $X_1, \ldots, X_m$ be independent and identically distributed indicator random variables, with $\mu = E[X_i]$. If $m \geq \frac{3 \ln \frac{2}{\delta}}{\epsilon^2 \mu}$, then

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\mu\right) \leq \delta.$$

That is, m samples provide an (ϵ, δ)-approximation for μ.

Monte Carlo Integration

We want to compute the definite (numeric) integral $\int_a^b f(x)dx$ when the integral does not have a close form.
Let $a = x_0, \ldots, x_N = b$ such that for all i, $x_{i+1} - x_i = \frac{b-a}{N} = \delta(N)$.

$$\int_a^b f(x)dx = \lim_{\delta(N)\to 0} \sum_{i=0}^{N} f(x_i)\delta(N) = \lim_{N\to\infty} \frac{b-a}{N} \sum_{i=0}^{N} f(x_i).$$

We need to estimate

$$\bar{f} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} f(x_i),$$

which is the expected value of $f()$ in $[a, b]$.

We need to estimate

$$\bar{f} = \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N} f(x_i).$$

We choose N independent samples $y_1, \ldots, y_N$ uniformly distributed in $[a, b]$.

$$E[\frac{1}{N} \sum_{i=1}^{N} f(y_i)] = \bar{f}$$

$$Var[\frac{1}{N} \sum_{i=1}^{N} f(y_i)] = \frac{1}{N} Var[f(x)]$$

$$Pr(|\frac{1}{N} \sum_{i=1}^{N} f(y_i) - \bar{f}| \geq \epsilon) \leq \frac{Var[f(x)]}{N\epsilon^2}$$

Approximate Counting

Example counting problems:

1. How many spanning trees in a graph?
2. How many perfect matchings in a graph?
3. How many independent sets in a graph?
4. ....

DNF Counting (Karp, Luby, Madras)

DNF = Disjunctive Normal Form.
Problem: How many satisfying assignments to a DNF formula?
A DNF formula is a disjunction of clauses.
Each clause is a conjunction of literals.

$$(\overline{x_1} \wedge x_2) \vee (x_2 \wedge x_3) \vee (x_1 \wedge x_2 \wedge \overline{x_3} \wedge x_4) \vee (x_3 \wedge \overline{x_4})$$

Compare to CNF.

$$(x_1 \vee x_2) \wedge (x_1 \vee \overline{x_3}) \wedge \cdots$$

m clauses, n variables
Let's first convince ourselves that obvious approaches don't work!

DNF counting is hard

Question: Why?
We can reduce CNF satisfiability to DNF counting.
The negation of a CNF formula is in DNF.

1. CNF formula f
2. get the DNF formula $(\bar{f})$
3. count satisfying assignments to $\bar{f}$
4. If it was 2^n, then f is unsatisfiable.

DNF counting is #P complete

#P is the counting analog of NP.
Any problem in #P can be reduced (in polynomial time) to the DNF counting problem.
Example #P complete problems:

1. How many Hamilton circuits does a graph have?
2. How many satisfying assignments does a CNF formula have?
3. How many perfect matchings in a graph?

What can we do about a hard problem?

(ϵ, δ) FPRAS for DNF counting

n variables, m clauses.
FPRAS = "Fully Polynomial Randomized Approximation Scheme"
Notation:
U: set of all possible assignments to variables
$|U| = 2^n$.
$H \subset U$: set of satisfying assignments
Want to estimate $Y = |H|$
Give $\epsilon > 0, \delta > 0$, find estimate X such that

1. $\Pr[|X - Y| > \epsilon Y] < \delta$
2. Algorithm should be polynomial in $1/\epsilon$, $1/\delta$, n and m.

Monte Carlo method

Here's the obvious scheme.

1. Repeat N times:
    1.1. Sample x randomly from U
    1.2. Count a success if $x \in H$
2. Return "fraction of successes" $\times |U|$.

Question: How large should N be?

We have to evaluate the probability of our estimate being good.

Let $\rho = \frac{|H|}{|U|}$.

$Z_i = 1$ if i-th trial was successful

$$Z_i = \begin{cases} 1 & \text{with probability} \quad \rho \\ 0 & \text{with probability} \quad 1-\rho \end{cases}$$

$Z = \sum_{i=1}^{N} Z_i$ is a binomial r.v

$E[Z] = N\rho$

$X = \frac{Z}{N}|U|$ is our estimate of $|H|$

Probability that our algorithm succeeds

Recall: X denotes our estimate of $|H|$.

$$\begin{aligned}
& \Pr[(1-\epsilon)|H| < X < (1+\epsilon)|H|] \\
= \; & \Pr[(1-\epsilon)|H| < Z|U|/N < (1+\epsilon)|H|] \\
= \; & \Pr[(1-\epsilon)N\rho < Z < (1+\epsilon)N\rho] \\
> \; & 1 - e^{-N\rho\epsilon^2/3} - e^{-N\rho\epsilon^2/2} \\
> \; & 1 - 2e^{-N\rho\epsilon^2/3}
\end{aligned}$$

where we have used Chernoff bounds.
For an (ϵ, δ) approximation, this has to be greater than $1-\delta$,

$$\begin{aligned}
2e^{-N\rho\epsilon^2/3} &< \delta \\
N &> \frac{3}{\rho\epsilon^2} \log \frac{2}{\delta}
\end{aligned}$$

Theorem

Let $\rho = |H|/|U|$. Then the Monte Carlo method is an (ϵ, δ) approximation scheme for estimating $|H|$ provided that $N > \frac{3}{\rho\epsilon^2} \log \frac{2}{\delta}$.

What's wrong?

How large could $\frac{1}{\rho}$ be?

ρ is the fraction of satisfying assignments.

1. The number of possible assignments is 2^n.
2. Maybe there are only a polynomial (in n) number of satisfying assignments.
3. So, $\frac{1}{\rho}$ could be exponential in n.

Question: An example where formula has only a few assignments?

The trick: Change the Sampling Space

Increase the hit rate (ρ)!

Sample from a different universe, ρ is higher, and all elements of H still represented.

What's the new universe?

Notation: H_i set of assignments that satisfy clause i.

$H = H_1 \cup H_2 \cup \ldots H_m$

Define a new universe

$$U = H_1 \uplus H_2 \uplus \ldots \uplus H_m$$

$\uplus$ means *multiset union*.

Element of U is (v, i) where v is an assignment, i is the satisfied clause.

Example - Partition by clauses

$$(\overline{x_1} \wedge x_2) \vee (x_2 \wedge x_3) \vee (x_1 \wedge x_2 \wedge \overline{x_3} \wedge x_4) \vee (x_3 \wedge \overline{x_4})$$

x_1	x_2	x_3	x_4	Clause
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
0	1	1	0	2
0	1	1	1	2
1	1	1	0	2
1	1	1	1	2
1	1	0	1	3
0	0	1	0	4
0	1	1	0	4

More about the universe U

1. Element of U is (v, i) where v is an assignment, i is the satisfied clause.
2. U contains only the satisfying assignments.
3. U contains the same satisfying assignment many times.
   $U = \{(v, i) | v \in H_i\}$
4. Each satisfying assignment v appears in as many clauses as it satisfies.

One way of looking at U

Partition by clauses.

m partitions, partition i contains H_i.

Another way of looking at U

Partition by assignments (one region for each assignment v).
Each partition corresponds to an assignment.
Can we count the different (distinct) assignments?

Example - Partition by assignments

$$(\overline{x_1} \wedge x_2) \vee (x_2 \wedge x_3) \vee (x_1 \wedge x_2 \wedge \overline{x_3} \wedge x_4) \vee (x_3 \wedge \overline{x_4})$$

x_1	x_2	x_3	x_4	Clause
0	0	1	0	4
0	1	0	0	1
0	1	0	1	1
0	1	1	0	1
0	1	1	0	2
0	1	1	0	4
0	1	1	1	1
0	1	1	1	2
1	0	1	0	4
1	1	0	1	3
1	1	1	0	2

Canonical element

Crucial idea: For each assignment group, find a canonical element in U.
An element (v, i) is *canonical* if $f((v, i)) = 1$

$$f((v,i)) = \begin{cases} 1 & \text{if } i = \min\{j : v \in H_j\} \\ 0 & \text{otherwise} \end{cases}$$

For every assignment group, exactly one canonical element.
So, count the number of canonical elements!
Note: could use any other definition as long as exactly one canonical element per assignment

Count canonical elements

Reiterating:

1. Number of satisfying assignments = Number of canonical elements.
2. Count number of canonical elements.
3. Back to old random sampling method for counting!

What is ρ?

Lemma

$\rho \geq \frac{1}{m}$, *(pretty large).*

Proof:

$|H| = |\cup_{i=1}^{m} H_i|$, since H is a normal union.

So $|H_i| \leq |H|$

Recall $U = H_1 \uplus H_2 \uplus \ldots \uplus H_m$

$|U| = \sum_{i=1}^{m} |H_i|$, since U is a multiset union.

$|U| \leq m|H|$

$\rho = \frac{|H|}{|U|} \geq \frac{1}{m}$

How to generate a random element in U?

Look at the partition of U by clauses.

Algorithm Select:

1. Pick a random clause weighted according to the area it occupies.

$$\Pr[i] = \frac{|H_i|}{|U|} = \frac{|H_i|}{\sum_1^m |H_j|}$$

   $|H_i| = 2^{(n-k_i)}$ where k_i is the number of literals in clause i.

2. Choose a random satisfying assignment in H_i.
   - Fix the variables required by clause i.
   - Assign random values to the rest to get v

   (v, i) is the random element.

Running time: $O(n)$.

How to test if canonical assignment?

Or how to evaluate $f((v, i))$?

Algorithm Test:

1. Test every clause to see if v satisfies it.
   $cov(v) = \{(v, j) | v \in H_j\}$
2. If (v, i) the smallest j in $cov(v)$, then $f(v, i) = 1$, else 0.

Running time: $O(nm)$.

Back to random sampling

Algorithm Coverage:

1. $s \leftarrow 0$ (number of successes)
2. Repeat N times:
   - Select (v, i) using **Select**.
   - if $f(v, i) = 1$ (check using **Test**) then success, increment s.
3. Return $s|U|/N$.

Number of samples needed is (from Theorem 4):

$$N = \frac{3}{\epsilon^2 \rho} \ln \frac{2}{\delta} \leq \frac{3m}{\epsilon^2} \ln \frac{2}{\delta}$$

Sampling, testing: polynomial in n and m
We have an FPRAS

Theorem

The Coverage algorithm yields an (ϵ, δ) approximation to $|H|$ provided that the number of samples $N \geq \frac{3m}{\epsilon^2} \log \frac{2}{\delta}$.

Size of Union of Sets

Let $H_1, \ldots, H_k$ be subsets of a finite set S. What is the size of $H = \cup_{i=1}^{k} H_i$?

Theorem

The Coverage algorithm yields an (ϵ, δ) approximation to $|H|$ provided that the number of samples $N \geq \frac{3k}{\epsilon^2} \log \frac{2}{\delta}$.

The Monte-Carlo Markov-Chain (MCMC) Method

Given a graph $G = (V, E)$, an independent set I in G is a set of vertices connected by no edges in G.
$\Omega(G) =$ set of independent sets in G.

$$|V| \leq |\Omega(G)| \leq 2^{|V|}$$

We want to compute an (ϵ, δ)-approximation for $|\Omega(G)|$.

Definition

A randomized algorithm gives an (ϵ, δ)-*approximation* for the value V if the output X of the algorithm satisfies

$$\Pr(|X - V| \leq \epsilon V) \geq 1 - \delta.$$

Simple Monte-Carlo?

Theorem

Let $X_1, \ldots, X_m$ be independent and identically distributed indicator random variables, with $\mu = E[X_i]$. If $m \geq \frac{3 \ln \frac{2}{\delta}}{\epsilon^2 \mu}$, then

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right| \geq \epsilon\mu\right) \leq \delta.$$

That is, m samples provide an (ϵ, δ)-approximation for μ.

Repeat m times: choose a random set of vertices, if independent set $X_i = 1$, else $X_i = 0$.

$$\tilde{\mu} = \frac{1}{m}\sum_{i=1}^{n} X_i \qquad |\Omega(\tilde{G})| = \tilde{\mu} 2^{|V|} \qquad \frac{|V|}{2^{|V|}} \leq \tilde{\mu} \leq 1$$

$\mu = E[\tilde{\mu}]$ can be exponentially small, $\frac{|V|}{2^{|V|}} \leq \mu \leq 1$.
Can we sample from a different domain, such that the corresponding $\mu = \Omega(1)$

Counting Independent Sets

Input: a graph $G = (V, E)$. $|V| = n$, $|E| = m$.
Let $e_1, \ldots, e_m$ be an arbitrary ordering of the edges.

$$G_i = (V, E_i), \quad \text{where} \quad E_i = \{e_1, \ldots, e_i\}$$

$G = G_m$, $G_0 = (V, \emptyset)$ and G_{i-1} is obtained from G_i be removing a single edge.
$\Omega(G_i) =$ the set of independent sets in G_i.

$$|\Omega(G)| = \frac{|\Omega(G_m)|}{|\Omega(G_{m-1})|} \times \frac{|\Omega(G_{m-1})|}{|\Omega(G_{m-2})|} \times \frac{|\Omega(G_{m-2})|}{|\Omega(G_{m-3})|} \times \cdots \times \frac{|\Omega(G_1)|}{|\Omega(G_0)|} \times |\Omega(G_0)|.$$

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}, \qquad |\Omega(G)| = 2^n \prod_{i=1}^{m} r_i$$

Lemma

$r_i \geq 1/2$.

Proof.

$$\Omega(G_i) \subseteq \Omega(G_{i-1}).$$

Suppose that G_{i-1} and G_i differ in the edge $\{u, v\}$.
An independent set in $\Omega(G_{i-1}) \setminus \Omega(G_i)$ contains both u and v. To bound the size of the set $\Omega(G_{i-1}) \setminus \Omega(G_i)$, we associate each $I \in \Omega(G_{i-1}) \setminus \Omega(G_i)$ with an independent set $I \setminus \{v\} \in \Omega(G_i)$. An independent set $I' \in \Omega(G_i)$ is associated with no more than one independent set $I \cup \{v\} \in \Omega(G_{i-1}) \setminus \Omega(G_i)$, and thus $|\Omega(G_{i-1}) \setminus \Omega(G_i)| \leq |\Omega(G_i)|$. It follows that

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} = \frac{|\Omega(G_i)|}{|\Omega(G_i)| + |\Omega(G_{i-1}) \setminus \Omega(G_i)|} \geq 1/2.$$

Estimating r_i
Input: Graphs $G_{i-1} = (V, E_{i-1})$ and $G_i = (V, E_i)$.
Output: $\tilde{r}_i =$ an approximation of r_i.

1. $X \leftarrow 0$.
2. Repeat for $M = 12m^2\epsilon^{-2} \ln \frac{2m}{\delta}$ independent trials:
   1. Generate an uniform sample from $\Omega(G_{i-1})$;
   2. If the sample is an independent set in G_i, let $X \leftarrow X + 1$.
3. Return $\tilde{r}_i \leftarrow \frac{X}{M}$.

Lemma

When $m \geq 1$ and $0 < \epsilon \leq 1$, the procedure for estimating r_i yields an estimate $\tilde{r}_i$ that is $(\epsilon/2m, \delta/m)$-approximation for r_i.

How good is this estimate?

Lemma

When $m \geq 1$ and $0 < \epsilon \leq 1$, the procedure for estimating r_i yields an estimate $\tilde{r}_i$ that is $(\epsilon/2m, \delta/m)$-approximation for r_i.

- Our estimate is $2^n \prod_{i=1}^{m} \tilde{r}_i$
- The true number is $|\Omega(G)| = 2^n \prod_{i=1}^{m} r_i$.
- To evaluate the error in our estimate we need to bound the ratio

$$R = \prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i}.$$

How good is this estimate?

Lemma

Suppose that for all i, $1 \leq i \leq m$, $\tilde{r}_i$ is an $(\epsilon/2m, \delta/m)$-approximation for r_i. Then

$$\Pr(|R - 1| \leq \epsilon) \geq 1 - \delta.$$

For each $1 \leq i \leq m$, we have

$$\Pr\left(|\tilde{r}_i - r_i| \leq \frac{\epsilon}{2m} r_i\right) \geq 1 - \frac{\delta}{m}.$$

Equivalently,

$$\Pr\left(|\tilde{r}_i - r_i| > \frac{\epsilon}{2m} r_i\right) < \frac{\delta}{m}.$$

By the union bound the probability that $|\tilde{r}_i - r_i| > \frac{\epsilon}{2m} r_i$ for any i is at most δ, and hence $|\tilde{r}_i - r_i| \leq \frac{\epsilon}{2m} r_i$ for all i with probability at least $1 - \delta$. Equivalently,

$$1 - \frac{\epsilon}{2m} \leq \frac{\tilde{r}_i}{r_i} \leq 1 + \frac{\epsilon}{2m}$$

holds for all i with probability at least $1 - \delta$. When these bounds hold for all i, we can combine them to obtain

$$1 - \epsilon \leq \left(1 - \frac{\epsilon}{2m}\right)^m \leq \prod_{i=1}^{m} \frac{\tilde{r}_i}{r_i} \leq \left(1 + \frac{\epsilon}{2m}\right)^m \leq (1 + \epsilon),$$

Estimating r_i
Input: Graphs $G_{i-1} = (V, E_{i-1})$ and $G_i = (V, E_i)$.
Output: $\tilde{r}_i =$ an approximation of r_i.

1. $X \leftarrow 0$.
2. Repeat for $M = 12m^2\epsilon^{-2} \ln \frac{2m}{\delta}$ independent trials:
   1. Generate an uniform sample from $\Omega(G_{i-1})$;
   2. If the sample is an independent set in G_i, let $X \leftarrow X + 1$.
3. Return $\tilde{r}_i \leftarrow \frac{X}{M}$.

How do we Generate an (almost) uniform sample from $\Omega(G_{i-1})$?

Definition

Let w be the (random) output of a sampling algorithm for a finite sample space Ω. The sampling algorithm generates an ϵ-*uniform sample* of Ω if, for any subset S of Ω,

$$\left| \Pr(w \in S) - \frac{|S|}{|\Omega|} \right| \leq \epsilon.$$

A sampling algorithm is a *fully polynomial almost uniform sampler (FPAUS)* for a problem if, given an input x and a parameter $\epsilon > 0$, it generates an ϵ-uniform sample of $\Omega(x)$, and it runs in time polynomial in $\ln \epsilon^{-1}$ and the size of the input x.

From Approximate Sampling to Approximate Counting

Theorem

Given a fully polynomial almost uniform sampler (FPAUS) for independent sets in any graph, we can construct a fully polynomial randomized approximation scheme (FPRAS) for the number of independent sets in a graph G with maximum degree at most Δ.

The Markov Chain Monte Carlo Method

Consider a Markov chain whose states are independent sets in a graph $G = (V, E)$:

1. X_0 is an arbitrary independent set in G.
2. To compute X_{i+1}:
   1. Choose a vertex v uniformly at random from V.
   2. If $v \in X_i$ then $X_{i+1} = X_i \setminus \{v\}$;
   3. if $v \notin X_i$, and adding v to X_i still gives an independent set, then $X_{i+1} = X_i \cup \{v\}$;
   4. otherwise, $X_{i+1} = X_i$.

- The chain is irreducible
- The chain is aperiodic
- For $y \neq x$, $P_{x,y} = 1/|V|$ or 0
- $\Rightarrow$ uniform stationary distribution.

Time Reversible Markov Chain

Theorem

Consider a finite, irreducible, and ergodic Markov chain on n states with transition matrix P. If there are non-negative numbers $\bar{\pi} = (\pi_0, \ldots, \pi_n)$ such that $\sum_{i=0}^{n} \pi_i = 1$, and for any pair of states i, j,

$$\pi_i P_{i,j} = \pi_j P_{j,i},$$

then $\bar{\pi}$ is the stationary distribution corresponding to P.

Proof.

$$\sum_{i=0}^{n} \pi_i P_{i,j} = \sum_{i=0}^{n} \pi_j P_{j,i} = \pi_j.$$

Thus $\bar{\pi}$ satisfies $\bar{\pi} = \bar{\pi} P$, and $\sum_{i=0}^{n} \pi_i = 1$, and $\bar{\pi}$ must be the unique stationary distribution of the Markov chain. □

$N(x)-$ set of neighbors of x. Let $M \geq \max_{x \in \Omega} |N(x)|$.

Lemma

Consider a Markov chain where for all x and y with $y \neq x$, $P_{x,y} = \frac{1}{M}$ if $y \in N(x)$, and $P_{x,y} = 0$ otherwise. Also, $P_{x,x} = 1 - \frac{|N(x)|}{M}$. If this chain is irreducible and aperiodic, then the stationary distribution is the uniform distribution.

Proof.

We show that the chain is time-reversible.
For any $x \neq y$, if $\pi_x = \pi_y$, then

$$\pi_x P_{x,y} = \pi_y P_{y,x},$$

since $P_{x,y} = P_{y,x} = 1/M$. It follows that the uniform distribution $\pi_x = 1/|\Omega|$ is the stationary distribution. □

The Metropolis Algorithm

Assuming that we want to sample with non-uniform distribution.
For example, we want the probability of an independent set of size i to be proportional to λ^i.
Consider a Markov chain on independent sets in $G = (V, E)$:

1. X_0 is an arbitrary independent set in G.
2. To compute X_{i+1}:
   1. Choose a vertex v uniformly at random from V.
   2. If $v \in X_i$ then set $X_{i+1} = X_i \setminus \{v\}$ with probability $\min(1, 1/\lambda)$;
   3. if $v \notin X_i$, and adding v to X_i still gives an independent set, then set $X_{i+1} = X_i \cup \{v\}$ with probability $\min(1, \lambda)$;
   4. otherwise, set $X_{i+1} = X_i$.

Lemma

For a finite state space Ω, let $M \geq \max_{x \in \Omega} |N(x)|$. For all $x \in \Omega$, let $\pi_x > 0$ be the desired probability of state x in the stationary distribution. Consider a Markov chain where for all x and y with $y \neq x$,

$$P_{x,y} = \frac{1}{M} \min\left(1, \frac{\pi_y}{\pi_x}\right)$$

if $y \in N(x)$, and $P_{x,y} = 0$ otherwise. Further, $P_{x,x} = 1 - \sum_{y \neq x} P_{x,y}$. Then if this chain is irreducible and aperiodic, the stationary distribution is given by the probabilities π_x.

Proof.

We show the chain is time-reversible. For any $x \neq y$, if $\pi_x \leq \pi_y$, then $P_{x,y} = 1$ and $P_{y,x} = \pi_x/\pi_y$. It follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. Similarly, if $\pi_x > \pi_y$, then $P_{x,y} = \pi_y/\pi_x$ and $P_{y,x} = 1$, and it follows that $\pi_x P_{x,y} = \pi_y P_{y,x}$. □

Note that the Metropolis Algorithm only needs the ratios π_x/π_y's. In our construction, the probability of an independent set of size i is λ^i/B for $B = \sum_x \lambda^{size(x)}$ although we may not know B.

Coupling and MC Convergance

- An Ergodic Markov Chain converges to its stationary distribution.
- How long do we need to run the chain until we sample a state in almost the stationary distribution?
- How do we measure distance between distributions?
- How do we analyze speed of convergence?

Variation Distance

Definition

The *variation distance* between two distributions D_1 and D_2 on a countably finite state space S is given by

$$||D_1 - D_2|| = \frac{1}{2} \sum_{x \in S} |D_1(x) - D_2(x)|.$$

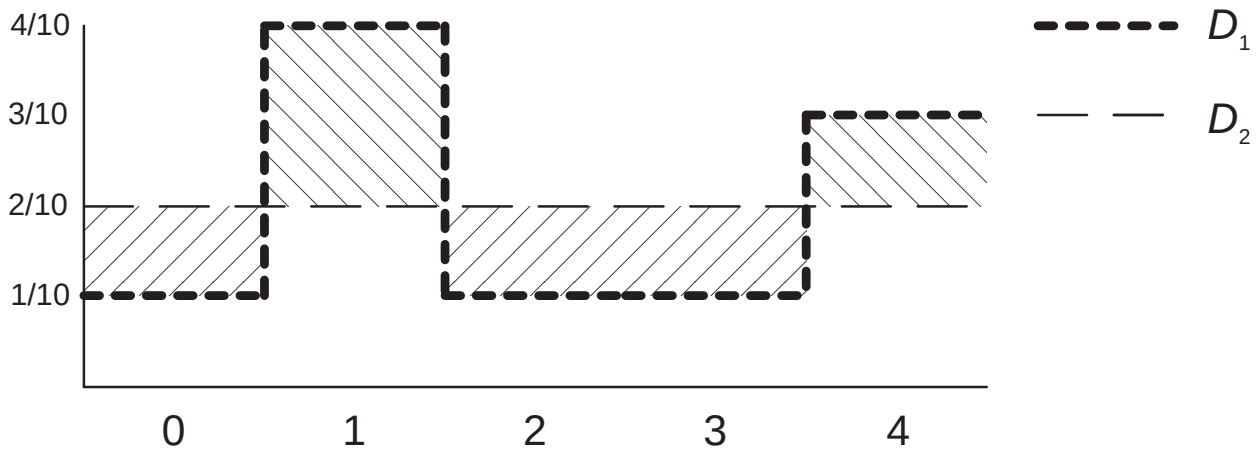


Figure: The total area shaded by upward diagonal lines must equal the total areas shaded by downward diagonal lines, and the variation distance equals one of these two areas.

Lemma

For any $A \subseteq S$, let $D_i(A) = \sum_{x \in A} D_i(x)$, for $i = 1, 2$. Then,

$$||D_1 - D_2|| = \max_{A \subseteq S} |D_1(A) - D_2(A)|.$$

Let $S^+ \subseteq S$ be the set of states such that $D_1(x) \geq D_2(x)$, and $S^- \subseteq S$ be the set of states such that $D_2(x) > D_1(x)$.
Clearly

$$\max_{A \subseteq S} D_1(A) - D_2(A) = D_1(S^+) - D_2(S^+),$$

and

$$\max_{A \subseteq S} D_2(A) - D_1(A) = D_2(S^-) - D_1(S^-).$$

But since $D_1(S) = D_2(S) = 1$, we have

$$D_1(S^+) + D_1(S^-) = D_2(S^+) + D_2(S^-) = 1,$$

which implies that

$$D_1(S^+) - D_2(S^+) = D_2(S^-) - D_1(S^-).$$

$$\max_{A \subseteq S} |D_1(A) - D_2(A)| = |D_1(S^+) - D_2(S^+)| = |D_1(S^-) - D_2(S^-)|.$$

and

$$|D_1(S^+) - D_2(S^+)| + |D_1(S^-) - D_2(S^-)| = \sum_{x \in S} |D_1(x) - D_2(x)|$$

$$= 2||D_1 - D_2||,$$

we have

$$\max_{A \subseteq S} |D_1(A) - D_2(A)| = ||D_1 - D_2||,$$

Rate of Convergence

Definition

Let π be the stationary distribution of a Markov chain with state space S. Let p_x^t represent the distribution of the state of the chain starting at state x after t steps. We define

$$\Delta_x(t) = ||p_x^t - \pi||; \qquad \Delta(t) = \max_{x \in S} \Delta_x(t).$$

That is, $\Delta_x(t)$ is the variation distance between the stationary distribution and p_x^t, and $\Delta(t)$ is the maximum of these values over all states x.
We also define

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon\}; \qquad \tau(\epsilon) = \max_{x \in S} \tau_x(\epsilon).$$

That is, $\tau_x(\epsilon)$ is the first step t at which the variation distance between p_x^t and the stationary distribution is less than ϵ, and $\tau(\epsilon)$ is the maximum of these values over all states x.

Example: Shuffling Cards

Markov chain:

- States: orders of the deck of cards.
- Transitions: at each step choose one card, uniformly at random, and move to the top.
- Uniform stationary distribution (not time reversal, but fully symmetric).

How many transitions until the process is mixing?

Coupling

Definition

A coupling of a Markov chain M with state space S is a Markov chain $Z_t = (X_t, Y_t)$ on the state space $S \times S$ such that

$$\begin{aligned} \Pr(X_{t+1} = x' | Z_t = (x, y)) &= \Pr(X_{t+1} = x' | X_t = x); \\ \Pr(Y_{t+1} = y' | Z_t = (x, y)) &= \Pr(Y_{t+1} = y' | Y_t = y). \end{aligned}$$

The Coupling Lemma

Lemma (Coupling Lemma)

Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain M on a state space S. Suppose that there exists a T so that for every $x, y \in S$,

$$\Pr(X_T \neq Y_T \mid X_0 = x, Y_0 = y) \leq \epsilon.$$

Then

$$\tau(\epsilon) \leq T.$$

That is, for any initial state, the variation distance between the distribution of the state of the chain after T steps and the stationary distribution is at most ϵ.

Proof.

Consider the coupling when Y_0 is chosen according to the stationary distribution and X_0 takes on any arbitrary value. For the given T and ϵ, and for any $A \subseteq S$

$$\begin{aligned}
\Pr(X_T \in A) &\geq \Pr((X_T = Y_T) \cap (Y_T \in A)) \\
&= 1 - \Pr((X_T \neq Y_T) \cup (Y_T \notin A)) \\
&\geq (1 - \Pr(Y_T \notin A)) - \Pr(X_T \neq Y_T) \\
&\geq \Pr(Y_T \in A) - \epsilon \\
&= \pi(A) - \epsilon.
\end{aligned}$$

Similarly,

$$\Pr(X_T \notin A) \geq \pi(S \setminus A) - \epsilon$$

or

$$\Pr(X_T \in A) \leq \pi(A) + \epsilon$$

It follows that

$$\max_{x,A} |p_x^T(A) - \pi(A)| \leq \epsilon,$$

Example: Shuffling Cards

- Markov chain:
  - States: orders of the deck of cards.
  - Transitions: at each step choose one card, uniformly at random, and move to the top.
  - Uniform stationary distribution
- Given two such chains: X_t and Y_t we define the coupling:
  - The first chain chooses a card uniformly at random and move it to the top.
  - The second chain move the same card (it may be in a different location) to the top.
- The probability that any card was not chosen by the first chain in $n \log n + cn$ steps is e^{-c}.
- After $n \log(n/\epsilon)$ steps the variation distance between our chain and the uniform distribution is bounded by ϵ.

$$\tau(\epsilon) \leq n \ln(n/\epsilon).$$

Example: Random Walks on the Hypercube

- Consider n-cube, with $N = 2^n$ nodes., Let $\bar{x} = (x_1, \ldots, x_n)$ be the binary representation of x. Nodes x and y are connected by an edge iff $\bar{x}$ and $\bar{y}$ differ in exactly one bit.
- Markov chain on the n-cube: at each step, choose a coordinate i uniformly at random from $[1, n]$, and set x_i to 0 with probability $1/2$ and 1 with probability $1/2$.
- Coupling: both chains choose the same bit and give it the same value.
- The chains couple when all bits have been chosen.
- By the Coupling Lemma the mixing time satisfies

$$\tau(\epsilon) \leq n \ln(n\epsilon^{-1}).$$

Example: Sampling Independent Sets of a Given Size

Consider a Markov chain whose states are independent sets of size k in a graph $G = (V, E)$:

1. X_0 is an arbitrary independent set of size k in G.
2. To compute X_{i+1}:
   1. Choose uniformly at random $v \in X_t$ and $w \in V$.
   2. if $w \notin X_i$, and $(X_t - \{v\}) \cup \{w\}$ is an independent set, then $X_{t+1} = (X_t - \{v\}) \cup \{w\}$
   3. otherwise, $X_{i+1} = X_i$.

- If the chain is irreducible
- The chain is aperiodic
- For $y \neq x$, $P_{x,y} = 1/|V|$ or 0.
- Uniform stationary distribution

Irreducible

Lemma

Let G be a graph on n vertices with maximum degree $\leq \Delta$. For $k \leq n/(3\Delta + 3)$, the chain is irreducible.

Proof.

Let $N(I)$ be the set of neighbors of nodes in I.
Let I_1 and I_2 be two independent sets of size k. The two independent sets and the neighbors of their nodes cover no more than $2k(\Delta + 1)$ nodes. Thus, there is a third independent set J, such that

$$(J \cup N(J)) \cap (I_1 \cup I_2 \cup N(I_1) \cup N(I_2)) = \emptyset.$$

.
The chain can move from I_1 to I_2 by first moving to J and then to I_2. □

Convergence Time

Theorem

Let G be a graph on n vertices with maximum degree $\leq \Delta$. For $k \leq n/(3\Delta + 3)$,

$$\tau(\epsilon) \leq kn \ln \epsilon^{-1}.$$

Coupling:

1. X_0 and Y_0 are arbitrary independent sets of size k in G.
2. To compute X_{i+1} and Y_{t+1}:
   1. Choose uniformly at random $v \in X_t$ and $w \in V$.
   2. if $w \notin X_i$, and $(X_t - \{v\}) \cup \{w\}$ is an independent set, then $X_{t+1} = (X_t - \{v\}) \cup \{w\}$, otherwise, $X_{i+1} = X_i$.
   3. If $v \notin Y_t$ choose v' uniformly at random from $Y_t - X_t$, else $v' = v$.
   4. if $w \notin Y_i$, and $(Y_t - \{v'\}) \cup \{w\}$ is an independent set, then $Y_{t+1} = (Y_t - \{v'\}) \cup \{w\}$, otherwise, $Y_{t+1} = Y_t$.

Let $d_t = |X_t - Y_t|$,

- $|d_{t+1} - d_t| \leq 1$.
- $d_{t+1} = d_t + 1$ iff $v \in Y_t$ and there is move in only one chain. Either w or its neighbor must be in $(X_t - Y_t) \cup (Y_t - X_t)$

$$\Pr(d_{t+1} = d_t + 1) \leq \frac{k - d_t}{k} \frac{2d_t(\Delta + 1)}{n}.$$

- $d_{t+1} = d_t - 1$ if $v \notin Y_t$ and w and its neighbors are not in $X_t \cup Y_t - \{v, v'\}$. $|X_t \cup Y_t| = k + d_t$

$$\Pr(d_{t+1} = d_t - 1) \geq \frac{d_t}{k} \frac{n - (k + d_t - 2)(\Delta + 1)}{n}.$$

We have for $d_t > 0$,

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t] &= d_t + \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1) \\
&\leq d_t + \frac{k - d_t}{k} \frac{2d_t(\Delta + 1)}{n} - \frac{d_t}{k} \frac{n - (k + d_t - 2)(\Delta + 1)}{n} \\
&= d_t \left(1 - \frac{n - (3k - d_t - 2)(\Delta + 1)}{kn}\right) \\
&\leq d_t \left(1 - \frac{n - (3k - 3)(\Delta + 1)}{kn}\right).
\end{aligned}$$

Once $d_t = 0$, the two chains follow the same path, thus $\mathbf{E}[d_{t+1} \mid d_t = 0] = 0$.

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq \mathbf{E}[d_t] \left(1 - \frac{(n - 3k + 3)(\Delta + 1)}{kn}\right).$$

$$\mathbf{E}[d_t] \leq d_0 \left(1 - \frac{n - (3k - 3)(\Delta + 1)}{kn}\right)^t.$$

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq \mathbf{E}[d_t] \left(1 - \frac{(n-3k+3)(\Delta+1)}{kn}\right).$$

Since $d_0 \leq k$, and d_t is a non-negative integer,

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq k \left(1 - \frac{n-(3k-3)(\Delta+1)}{kn}\right)^t \leq \mathrm{e}^{-t\frac{n-(3k-3)(\Delta+1)}{kn}}.$$

For $k \leq n/(3\Delta+3)$,

$$\tau(\epsilon) \leq \frac{kn \ln \epsilon^{-1}}{n-(3k-3)(\Delta+1)}.$$

In particular, when k and Δ are constants, $\tau(\epsilon) = O(\ln \epsilon^{-1})$.

Approximately Sampling Proper Colorings

- A *proper vertex coloring* of a graph gives each vertex v a color from a set $C = \{1, 2, \ldots, c\}$ such that the two endpoints of every edge are colored by two different colors.
- Any graph with maximum degree Δ can be colored properly with $c = \Delta + 1$ colors.
- We are interested in sampling almost uniformly at random a proper coloring of a graph with a fixed $c \geq \Delta + 1$ colors.

MCMC for Sampling Proper Coloring

Markov chain whose states are proper coloring of a graph $G = (V, E)$ with colors in C:

1. X_0 is an arbitrary proper coloring of G.
2. To compute X_{i+1}:
   1. Choose uniformly at random $v \in V$ and $b \in C$.
   2. if coloring v with b gives a proper coloring then change the color of v to b to obtain X_{t+1}
   3. otherwise, $X_{i+1} = X_i$.

- The chain is irreducible if $c \geq 2\Delta + 1$
- The chain is aperiodic
- Uniform stationary distribution

Easy Result

Theorem

For any graph with n vertices and maximum degree Δ, the mixing time of the graph-coloring Markov chain is

$$\tau(\epsilon) \leq \left\lceil \frac{nc}{c - 4\Delta} \ln(n/\epsilon) \right\rceil,$$

as long as $c \geq 4\Delta + 1$.

Simple coupling: use the same choice of v and c in both chains.

Proof

- $D_t =$ the set of vertices that have different colors in the two chains at time t,
- $d_t = |D_t|$ can change by at most ± 1 in each iteration.
- The probability that $v \in D_t$ and b is not used by the Δ neighbors of v in both chains is

$$\Pr(d_{t+1} = d_t - 1 \mid d_t > 0) \geq \frac{d_t}{n} \frac{c - 2\Delta}{c}.$$

- The probability that $v \in V - D_t$ and it is recolored in only one chain is bounded by the probability that v has a neighbor $w \in D_t$, and we choose one of the colors used by w in the two chains.

$$\Pr(d_{t+1} = d_t + 1) \leq \frac{d_t \Delta}{n} \frac{2}{c}.$$

$$\begin{aligned}
\mathbf{E}[d_{t+1} \mid d_t] &= d_t + \Pr(d_{t+1} = d_t + 1) - \Pr(d_{t+1} = d_t - 1) \\
&\leq d_t + \frac{d_t}{n}\frac{2\Delta}{c} - \frac{d_t}{n}\frac{c - 2\Delta}{c} \\
&\leq d_t \left(1 - \frac{c - 4\Delta}{nc}\right),
\end{aligned}$$

which also holds if $d_t = 0$.
Using the conditional expectation equality, we have

$$\mathbf{E}[d_{t+1}] = \mathbf{E}[\mathbf{E}[d_{t+1} \mid d_t]] \leq \mathbf{E}[d_t] \left(1 - \frac{c - 4\Delta}{nc}\right).$$

By induction, we find

$$\mathbf{E}[d_t] \leq d_0 \left(1 - \frac{c - 4\Delta}{nc}\right)^t .$$

Since $d_0 \leq n$, and d_t is a non-negative integer,

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq n \left(1 - \frac{c - 4\Delta}{nc}\right)^t \leq n\mathrm{e}^{-t(c-4\Delta)/nc}.$$

Hence the variation distance is at most ϵ after

$$t = \left\lceil \frac{nc}{c - 4\Delta} \ln(n/\epsilon) \right\rceil$$

steps.

Stronger result

Theorem

Given an n vertex graph with maximum degree Δ, the mixing time of the graph-coloring Markov chain is

$$\tau(\epsilon) \leq \left\lceil \frac{n(c-\Delta)}{c-2\Delta} \ln(n/\epsilon) \right\rceil,$$

as long as $c \geq 2\Delta + 1$.

Better Coupling

- D_t - vertices with different colors in the two chains.
- $A_t = V - D_t$ - vertices with the same colors in both chains.
- For $v \in A_t$ let $d'(v)$ be the number of neighbors of v in D_t
- For $v \in D_t$ let $d'(v)$ be the number of neighbors of v in A_t
- $\sum_{v \in D_t} d'(v) = \sum_{v \in A_t} d'(v) = m'$

$$\begin{aligned} Pr(d_{t+1} = d_t - 1 \mid d_t > 0) &\geq \frac{1}{n} \sum_{v \in D_t} \frac{c - 2\Delta + d'(v)}{c} \\ &= \frac{1}{cn} \left((c - 2\Delta) d_t + m' \right). \end{aligned}$$

- We want to decrease the probability that a vertex $v \in A_t$ is re-colored in just one chain.
- When $v \in A_t$ let $S_1(v)$ be the set of colors of neighbors of v in the first chain and not in the second chain, $S_2(v)$ in the second chain and not the first.
- When choosing the color in the second chain couple $S_1(v)$ and $S_2(v)$ as much as possible, so when the first chain uses $c \in S_1(v)$ the second chain uses $c' \in S_2(v)$.
- The number of coloring that increase d_t is bounded by $\max(|S_1(v)|, |S_2(v)|) \leq d'(v)$.

$$Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \quad \leq \quad \frac{1}{n} \sum_{v \in A_t} \frac{d'(v)}{c} = \frac{m'}{cn}$$

$$\begin{aligned} E[d_{t+1} \mid d_t] &\leq d_t + \frac{m'}{cn} - \frac{1}{cn}\left((c-2\Delta)d_t + m'\right) \\ &= d_t\left(1 - \frac{c-2\Delta}{nc}\right). \end{aligned}$$

$$\Pr(d_t \geq 1) \leq \mathbf{E}[d_t] \leq n\left(1 - \frac{c-2\Delta}{nc}\right)^t \leq n\mathrm{e}^{-t(c-2\Delta)/nc},$$

and the variation distance is at most ϵ after

$$\tau(\epsilon) = \left\lceil \frac{nc}{c-2\Delta} \ln(n/\epsilon) \right\rceil$$

steps.