Markov Chains for Sampling

Sampling:

- Given complex state space
- Want to sample from it
- Use some Markov Chain
- Run for a long time
- end up “near” stationary distribution
- Reduces sampling to local moves (easier)
- no need for global description of state space
- Allows sample from exponential state space

Formalize: what is “near” and “long time”?

- Stationary distribution π
- arbitrary distribution q
- **relative pointwise distance (r.p.d.)** $\max_j |q_j - \pi_j|/\pi_j$
- Intuitively close.
- Formally, suppose r.p.d. δ.
- Then $(1-\delta)\pi \leq q$
- So can express distribution q as “with probability $1-\delta$, sample from π. Else, do something wierd.
- So if δ small, “as if” sampling from π each time.
- If δ poly small, can do poly samples without goof
- Gives “almost stationary” sample from Markov Chain
- Mixing Time: time to reduce r.p.d to some ϵ

Volume

Outline:

- Describe problem. Membership oracle
- $\sharp P$ hard to volume intersection of half spaces in n dimensions
- In low dimensions, integral.
- even for convex bodies, can't do better than $(n/\log n))^n$ ratio
- what about FPRAS?

Estimating π:

- pick random in unit square
- check if in circle
- gives ratio of square to circle
- Extends to arbitrary shape with "membership oracle"
- Problem: rare events.
- Circle has good easy outer box

Problem: rare events:

- In 2d, long skinny shapes
- In high d, even round shape has exponentially larger bounding box

Solution: "creep up" on volume

- modify P to contain unit sphere B_1 r_1, contined in larger B_2 of radius r, r polynomial
- choose ρ near $1 - 1/d$.
- Consider sequence of bodies $\rho^i rP \cap B_2$
- note for large i, get P
- but for $i = 0$, body contains B_2
- so volume known
- so just need ratios
- At each step, need to random sample from $\rho^i rP \cap B_2$
- Sample method: random walk forbidden to leave
- eigenvalues show rapid mixing
- egienvalues small because body convex: no bottlenecks

Expander Walks

omitted

Another example and application: (n, d, c)-Expanders.

- bipartite
- n vertices, regular degree d
- $|\Gamma(S)| \geq (1 + c(1 - 2|S|/n))|S|$
- factor c more neighbors, at least until S near $n/2$.
- Add self loops (with probability 1/2 to deal with periodicity.
- What is stationary distribution? Uniform.
- Intuition on convergence: because neighborhoods grow, position becomes unpredictable very fast.
- Theorem:
$$\lambda_2 \leq 1 - \frac{c^2}{d(2048 + 4c^2)}$$
- Converse theorem: if $\lambda_2 \leq 1 - \epsilon$, get expander with
$$c \geq 4(\epsilon - \epsilon^2)$$

Gabber-Galil expanders:

- Do expanders exist? Yes! proof: probabilistic method.
- But in this case, can do better deterministically.
  - Gabber Galil expanders.
  - Let $n = 2m^2$. Vertices are (x, y) where $x, y \in Z_m$ (one set per side)
  - 5 neighbors: $(x, y), (x, x + y), (x, x + y + 1), (x + y, y), (x + y + 1, y)$ (add mod m)
  - or 7 neighbors of similar form.
- Theorem: this $d = 5$ graph has $c = (2 - \sqrt{3})/4$, degree 7 has twice the expansion.
- in other words, c and d are constant.
- meaning $\lambda_2 = 1 - \epsilon$ for some **constant** ϵ
- So random walks on this expander mix *very* fast: for polynomially small r.p.d., $O(\log n)$ steps of random walk suffice.
- Note also that n can be huge, since only need to store one vertex ($O(\log n)$ bits).

Application: conserving randomness.

- Consider an BPP algorithm (gives right answer with probability 99/100 (constant irrelevant) using n bits.
- t independent trials with majority rule reduce failure probability to $2^{-O(t)}$ (chernoff), but need tn bits
- in case of RP, used 2-point sampling to get error $O(1/t)$ with $2n$ bits and t trials.
- Use walk instead.
  - vertices are $N = 2^n$ (n-bit) random strings for algorithm.
  - edges as degree-7 expander
  - only 1/100 of vertices are bad.
  - what is probability majority of time spent there?
  - in limit, spend 1/100 of time there
  - how fast converge to limit? How long must we run?
  - Power the markov chain so $\lambda_2^\beta \leq 1/10$ (constant number of steps)
  - use random seeds encountered every β steps.
- number of bits needed:
  - $O(n)$ for stationary starting point
  - 3β more per trial,
- Theorem: after $7k$ samples, probability majority wrong is $1/2^k$. So error $1/2^n$ with $O(n)$ bits!
  - Let B be powered transition matrix
  - let $p^{(i)}$ be distribution of sample i, namely $p^0 B^i$
  - Let W be indicator **matrix** for good witnesses, namely 1 at diagonal i if i is a witness. $\overline{W}$ completmentary set $I - W$.
  - $\|p^i W\|_1$ is probability p^i is witness set. similar for nonwitness.
  - Consider a sequence of $7k$ results "witness or not"
  - represent as matrices $S = (S_1, \ldots, S_{7k}) \in \{W, \overline{W}\}^{7k}$
  - claim
    $$\Pr[S] = \|p^{(0)}(BS_1)(BS_2)\cdots(BS_{7k})\|_1.$$
    (sums prob. of paths through correct sequence of witness/nonwitness)
  - defer: $\|pBW\|_2 \leq \|p\|_2$ and $\|pB\overline{W}\|_2 \leq \frac{1}{5}\|p\|_2$

- deduce if more than $7k/2$ bad witnesses,

$$\begin{aligned}\|p^0 \prod BS_i\|_1 &\le \sqrt{N}\|p^0 \prod BS_i\| \\ &\le \sqrt{N}(\frac{1}{5})^{7k/2}\|p^0\| \\ &\le = (\frac{1}{5})^{7k/2}\end{aligned}$$

- At same time, only 2^{7k} bad sequences, so error prob. $2^{7k}5^{-7k/2} \le 2^{-k}$

- proof of lemma:
  - write $p = \sum c_i e_i$
  - obviously $\|pBW\| \le \|pW\|$ since W jiust zeros some stuff out.
  - write $p = \pi + y$ as before where $y \cdot \pi = 0$
  - argue that $\|\pi B\overline{W}\| \le \|\pi\|/10$ and $yB\overline{W}\| \le \|y\|/10$, done.
  - First π:
    * recall $\pi B = \pi$ is uniform vector, all coords $1/\sqrt{N}$
    * $\overline{W}$ has only $1/100$ of coordintes nonzero, so
    * $\|e_1\overline{W}\| = \sqrt{(N/100)(1/N)} = 1/10$
  - Now y: just note $\|yB\| \le \|y\|/10$ since $\lambda_2 \le 1/10$. Then $\overline{W}$ zeros out.
  - summary: π part unlikely to be in witness set, y part unlikely to be relevant.

Coupling:

Method

- Run two copies of Markov chain X_t, Y_t
- Each considered in isolation is a copy of MC (that is, both have MC distribution)
- **but** they are not independent: they make dependent choices at each step
- in fact, after a while they are almost certainly the **same**
- Start Y_t in stationary distribution, X_t anywhere
- Coupling argument:

$$\begin{aligned}\Pr[X_t = j] &= \Pr[X_t = j \mid X_t = Y_t]\Pr[X_t = Y_t] + \Pr[X_t = j \mid X_t \ne Y_t]\Pr[X_t \ne Y_t] \\ &= \Pr[Y_t = j]\Pr[X_t = Y_t] + \epsilon \Pr[X_t = j \mid X_t \ne Y_t]\end{aligned}$$

  So just need to make ϵ (which is r.p.d.) small enough.

n-bit Hypercube walk: at each step, flip random bit to random value

- At step t, pick a random bit b, random value v
- both chains set but b to value v
- after $O(n \log n)$ steps, probably all bits matched.

Counting k colorings when $k > 2\Delta + 1$

- The reduction from (approximate) uniform generation
  - compute ratio of coloring of G to coloring of $G - e$
  - Recurse counting $G - e$ colorings
  - Base case k^n colorings of empty graph
- Bounding the ratio:
  - note $G - e$ colorings outnumber G colorings
  - By how much? Let L colorings in difference (u and v same color)
  - to make an L coloring a G coloring, change u to one of $k - \Delta = \Delta + 1$ legal colors
  - Each G-coloring arises at most one way from this
  - So each L coloring has at least $\Delta + 1$ neighbors unique to them
  - So L is $1/(\Delta + 1)$ fraction of G.
  - So can estimate ratio with few samples
- The chain:
  - Pick random vertex, random color, try to recolor
  - loops, so aperiodic
  - Chain is time-reversible, so uniform distribution.
- Coupling:
  - choose random vertex v (same for both)
  - based on X_t and Y_t, choose bijection of colors
  - choose random color c
  - apply c to v in X_t (if can), $g(c)$ to v in Y_t (if can).
  - What bijection?
    * Let A be vertices that agree in color, D that disagree.
    * if $v \in D$, let g be identity
    * if $v \in A$, let N be neighbors of v
    * let C_X be colors that N has in X but not Y (X can't use them at v)
    * let C_Y similar, wlog larger than C_X

- * g should swap each C_X with some C_Y, leave other colors fixed. **Result:** if X doesn't change, Y doesn't

- Convergence:
  - Let $d'(v)$ be number of neighbors of v in opposite set, so
$$\sum_{v \in A} d'(v) = \sum_{v \in D} d'(v) = m'$$
  - Let $\delta = |D|$
  - Note at each step, δ changes by $0, \pm 1$
  - When does it increase?
    * v must be in A, but move to D
    * happens if only one MC accepts new color
    * If c not in C_X or C_Y, then $g(c) = c$ and both change
    * If $c \in C_X$, then $g(c) \in C_Y$ so neither moves
    * So must have $c \in C_Y$
    * But $|C_Y| \leq d'(v)$, so probability this happens is
$$\sum_{v \in A} \frac{1}{n} \cdot \frac{d'(v)}{k} = \frac{m'}{kn}$$
  - When does it decrease?
    * must have $v \in D$, only one moves
    * sufficient that pick color not in either neighborhood of v,
    * total neighborhood size 2Δ, but that counts the $d'(v)$ elements of A twice.
    * so Prob.
$$\sum_{v \in D} \frac{1}{n} \cdot \frac{k - (2\Delta - d'(v))}{k} = \frac{k - 2\Delta}{kn}\delta + \frac{m'}{kn}$$
  - Deduce that expected *change* in δ is difference of above, namely
$$-\frac{k - 2\Delta}{kn}\delta = -a\delta.$$
  - So after t steps, $E[\delta_t] \leq (1-a)^t \delta_0 \leq (1-a)^t n$.
  - Thus, probability $\delta > 0$ at most $(1-a)^t n$.
  - But now note $a > 1/n^2$, so $n^2 \log n$ steps reduce to one over polynomial chance.

Note: couple depends on state, but who cares

- From worm's eye view, each chain is random walk
- so, all arguments hold

Counting vs. generating:

- we showed that by generating, can count
- by counting, can generate: