Approximate Counting By Sampling

CompSci 590.02
Instructor: AshwinMachanavajjhala

Duke UNIVERSITY

Recap

Till now we saw …

- Efficient sampling techniques to get uniformly random samples
  - Reservoir sampling
  - Sampling using a tree index
  - Sampling using a nearest neighbor index

Today's class

- Use sampling for approximate counting.

Counting Problems

- Given a decision problem S, compute the number of feasible solutions to S (denoted by #S).

Example:

- #DNF: Count the number of satisfying assignments of a boolean formula in DNF
  - E.g., $(x_1 \Lambda \bar{x}_2 \Lambda \bar{x}_3 \Lambda x_4) \mathrm{V} (x_3 \Lambda \bar{x}_5 \Lambda x_6)$
  - Let n = number of variables
  - Let m = number of disjuncts

- Counting the number of triangles in a graph

Applications of DNF counting

- Advertising
  - Contracts are of the following form:
    Need 1 million impressions [Males, 15-25, CA] OR [Males, 15-35, TX]
  - Use historical data to estimate whether such a contract can be fulfilled.

- Web Search
  - Given a keyword query q = (k1, k2, …, km)
    Find the number of documents that contain at least one keyword.

DNF Counting is Hard

- Checking whether a DNF formula is unsatisfiable is NP-hard
- #DNF ε **#P**
- #P is the class of all problems for which there exist a non-deterministic polynomial time algorithm A such that for any instance I, the number of accepting computations is #I.
  - i.e., we can verify in polynomial time whether #I > 1.

FPRAS

- Our goal is design an *fully polynomial randomized approximation scheme* (FPRAS).

- For every input DNF, error parameter $\varepsilon > 0$, and confidence parameter $0 < \delta < 1$, the algorithm must output a value C' s.t.

$$\mathbf{P[(1-\varepsilon)\, C < C' < (1+\varepsilon)\, C] > 1-\delta}$$

where C is the true number of satisfying assignments, in time polynomial in the input DNF, $1/\varepsilon$ and $\log(1/\delta)$

FPRAS

- Sometimes, FPRAS are defined without the δ …

- For every input DNF, error parameter $\varepsilon > 0$, the algorithm must output a value C' s.t.

$$\mathbf{P[(1-\varepsilon)\, C < C' < (1+\varepsilon)\, C] > 3/4}$$

where C is the true number of satisfying assignments, in time polynomial in the input DNF, and $1/\varepsilon$

- Exercise: The two definitions are equivalent.

Monte Carlo Method

- Suppose U is a universe of elements
  - In DNF counting, U = set of all assignments from $\{0,1\}^n$
- Let G be a subset of interest in U
  - In DNF counting, G = set of all satisfying assignments.

For i = 1 to N

- Choose u ε U, uniformly at random
- Check whether u ε G ?
- Let $X_i = 1$ if u ε G, $X_i = 0$ otherwise

Return $\hat{C} = |U| \cdot \frac{\sum_i X_i}{N}$

Monte Carlo Method

When should you use it?

- Easy to uniformly sample from U
- Easy to check whether sample is in G
- N is polynomial in the size of the input.

$$Theorem:$$

$$\forall\, 0 < \varepsilon < 1.5, 0 < \delta < 1, if\ N > \frac{|U|}{|G|} \cdot \frac{3}{\varepsilon^2} \cdot \ln \frac{2}{\delta}$$

$$then, P\left[(1-\varepsilon)|G| \leq\ \hat{C} \leq (1+\varepsilon)|G|\right] \geq 1 - \delta$$

Chernoff Bound

Theorem:

$$If\ X_1, X_2, \ldots, X_n\ are\ independent\ binary\ random\ variables, Y_n = \sum_{i=1}^{n} X_i, E[Y_n] = \mu.\ Then, \forall \varepsilon \geq 0,$$

$$P[Y_n \geq (1+\varepsilon)\mu] \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu}$$

$$Moreover, \forall\ 0 \leq \varepsilon \leq 1,$$

$$P[Y_n \leq (1-\varepsilon)\mu] \leq \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{(1-\varepsilon)}}\right)^{\mu}$$

Upper Chernoff Bound Proof

$$P[Y_n \geq (1+\varepsilon)\mu] = P\left[e^{-t\cdot Y_n} \geq e^{-t(1+\varepsilon)\mu}\right], \forall t > 0$$

$$\leq \frac{E[e^{t\cdot Y_n}]}{e^{t(1+\varepsilon)\mu}} \qquad (Markov\ inequality)$$

$$= \frac{\prod_i E[e^{t\cdot X_i}]}{e^{t(1+\varepsilon)\mu}}$$

$$= \frac{\prod_i (p_i e^t + 1 - p_i)}{e^{t(1+\varepsilon)\mu}} = \frac{\prod_i (p_i(e^t - 1) + 1)}{e^{t(1+\varepsilon)\mu}}$$

$$\leq \frac{\prod_i e^{p_i(e^t-1)}}{e^{t(1+\varepsilon)\mu}} = \frac{e^{\mu(e^t-1)}}{e^{t(1+\varepsilon)\mu}}, \forall t > 0$$

$$RHS\ is\ minimized\ when\ t = \ln(1+\varepsilon)$$

Simpler Upper Tail Bound

$$P[Y_n \geq (1+\varepsilon)\mu] \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right)^{\mu}$$

$$\ln(1+\varepsilon) = \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \frac{\varepsilon^4}{4} + \cdots$$

$$(1+\varepsilon)\ln(1+\varepsilon) = \varepsilon + \frac{\varepsilon^2}{3} + \textit{positive terms}$$

$$(1+\varepsilon)^{(1+\varepsilon)} > e^{\left(-\varepsilon+\frac{\varepsilon^2}{3}\right)}$$

$$P[Y_n \geq (1+\varepsilon)\mu] \leq e^{-\varepsilon^2\mu/3}$$

Simpler Lower Tail Bound

$$P[Y_n \leq (1-\varepsilon)\mu] \leq \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{(1-\varepsilon)}}\right)^{\mu}$$

$$\ln(1-\varepsilon) = -\varepsilon - \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} - \frac{\varepsilon^4}{4} - \cdots$$

$$(1-\varepsilon)\ln(1-\varepsilon) = -\varepsilon + \frac{\varepsilon^2}{2} + \textit{positive terms}$$

$$(1-\varepsilon)^{(1-\varepsilon)} > e^{\left(-\varepsilon+\frac{\varepsilon^2}{2}\right)}$$

$$P[Y_n \leq (1-\varepsilon)\mu] \leq e^{-\varepsilon^2\mu/2}$$

DNF Counting

- $|U| = 2^n$
- $|G|$ can be *exponentially* smaller than $|U|$

Example: $(x_1 \wedge x_2) \vee (x_1 \wedge \bar{x}_2) \vee (x_1 \wedge x_3) \vee (x_1 \wedge \bar{x}_3) \ldots$

- Every satisfying assignment must contain $x_1 = 1$
- $|G| = 2^{n/2}$
- Large $|U|/|G|$ leads to an exponential number of samples for convergence.

Importance Sampling

- Set U' = {(*u, i*) | *u* is an assignment that satisfies disjunct *i* }
- Set G' = {(*u, i*) | *u* is an assignment that satisfies disjunct *i* but does not satisfy any disjunct *j* < *i* }

- |G'| = |G|
  - Each assignment appears exactly once.
- Easy to check if sample is in G'
- |U'| / |G'| ≤ m
  - Each assignment appears at most m times in U'

- We are done if we can sample uniformly from U'

Importance Sampling

- Given a DNF formula, it is easy to construct a satisfying assignment.
  - E.g., $(x_1 \Lambda \bar{x}_2 \Lambda \bar{x}_3 \Lambda x_4) \mathrm{V} (x_3 \Lambda \bar{x}_5 \Lambda x_6)$
  - Pick a clause (e.g. 1st)
  - Create a satisfying assignment for variables in that clause (e.g, 1001)
  - Randomly choose 0 or 1 for the remaining variables.

- If a disjunct i has k_i literals, there are 2^{n-ki} satisfying assignments (u,i)

- $|U'| = \sum_i 2^{n-ki}$

Importance Sampling

For i = 1 to N

- Choose a disjunct i, with probability $2^{n-ki}/|U'|$
- Generate a random assignment satisfying disjunct i
- Check whether u ε G ?
- Let $X_i = 1$ if u ε G, $X_i = 0$ otherwise

Return $\hat{C} = |U'| \cdot \frac{\sum_i X_i}{N}$

Theorem: The above algorithm is an (ε,δ) FPRAS if

$$N > m \cdot \frac{3}{\varepsilon^2} \cdot \ln \frac{2}{\delta}$$

Summary of DNF Counting

- #DNF is a #P-hard problem
- Monte Carlo method can result in a (ε,δ) FPRAS if
  - Can sample from U in PTIME
  - Can check membership in G PTIME
  - |G| is not very small compared to |U|
- Monte Carlo on a modified domain results in a (ε,δ) FPRAS for #DNF

Applications of Triangle Counting

- Measures of homophily
  - If A-B and B-C are edges, what is the probability that A-C is also an edge
- Clustering Coefficient: 3 x # triangles / # connected triples
- Transitivity Ratio: # triangles / # connected triples

Triangle Counting is "Easy"

- Naïve method: $O(n^3)$
- Well known methods that take $O(d_{max}^2 n)$ and $O(m^{1.5})$
- Still not efficient for a very large graph
  - Twitter in 2009
  - 54,981,152 nodes
  - 1,963,263,821 edges
  - Max degree > 3 million
  - Clustering Coefficient ~ 0.1

Is there an FPRAS?

- Exercise

References

- R. Karp, M. Luby, N. Madras, "Monte Carlo Estimation Algorithm for Enumeration Problems", Journal of Algorithms 10(3) 1989