15. Approximate Counting

Decision Problem in NP - π is in NP if for any YES instance I, $\exists$ a proof that I is a YES instance that can be verified in polytime.
Counting Problem - We define a counting problem for a decision problem π in NP, where given an instance I of π, we have to produce an output the number of solutions for the instance.

Polynomial Approximation Scheme

A deterministic algorithm A for a counting problem P such that it takes an instance I as input, and $\epsilon \in \mathbb{R}_{>0}$, and in polynomial time wrt $n = |I|$ produces an output $A(I)$ such that

$$(1-\epsilon)\#(I) \leq A(I) \leq (1+\epsilon)\#(I)$$

A fully polynomial approximation scheme (FPAS) is a PAS such that it runs in time $\operatorname{poly}(n, \frac{1}{\epsilon})$

Polynomial Randomized Approximation Scheme

A randomized algorithm A for a counting problem P that takes instance I as input, and $\epsilon \in \mathbb{R}_{>0}$, in time $\operatorname{poly}(\mathrm{n})$ produces $A(I)$ such that

$$\mathbb{P}[(1-\epsilon)\#(I) \leq A(I) \leq (1+\epsilon)\#(I)] \geq \frac{3}{4}$$

A fully polynomial randomized approximation scheme (FPRAS) is a PRAS such that it runs in time $\operatorname{poly}(n, \frac{1}{\epsilon})$

An $(\epsilon, \delta) - \mathbf{FPRAS}$ for a counting problem is an FPRAS that takes as input an instance I and computes an $\epsilon-$approximation to $\#(I)$ with probability $\geq 1-\delta$ in time $\operatorname{poly}(n, \frac{1}{e}, \frac{1}{\delta})$

Abstract Problem

We have U a finite set of known size, and a function $f : U \to \{0, 1\}$, using which we define $G = \{u \in U | f(u) = 1\}$. We have to estimate $|G|$.

Attempt 1

Sample $u_1, u_2, \ldots, u_N$ from U independently.
For all $i \in [N]$

$$Y_i = \begin{cases} 1 & \text{if } f(u_i) = 1 \\ 0 & \text{otherwise} \end{cases}$$

And we define Z as a random variable such that

$$Z = |U| \cdot \sum_{i=1}^{N} \frac{Y_i}{N}$$

Essentially, Z is an estimation of G.

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}\left[\sum Y_i\right] \cdot \frac{|U|}{N} \\ &= \sum_{i=1}^{N} \mathbb{E}[Y_i] \cdot \frac{|U|}{N} \\ &= \frac{|U|}{N} \sum_{i=1}^{N} \mathbb{P}[Y_i = 1] \\ &= \frac{|U|}{N} \cdot N \cdot \frac{|G|}{|U|} = |G| \end{aligned}$$

Estimator Theorem

Let $\rho = \frac{|G|}{|U|}$, then the Monte Carlo method yields an $\epsilon-$approximation to $|G|$ with probability $\geq 1 - \delta$ provided

$$N \geq \frac{4}{\epsilon^2 \rho} \ln \frac{2}{\delta}$$

Let $Y = \sum_{i=1}^{N} Y_i$. This allows us to rewrite $Z = \frac{|U| \cdot Y}{N}$

Computing probability it's an $\epsilon-$approximation,

$$\begin{aligned} \mathbb{P}[(1-\epsilon)|G| \leq Z \leq (1+\epsilon)|G|] &= \mathbb{P}[(1-\epsilon)N\rho \leq Y \leq (1+\epsilon)N\rho] \\ &\geq 1 - \mathbb{P}[Y > (1+\epsilon)N\rho] - \mathbb{P}[Y < (1-\epsilon)N\rho] \\ &= 1 - 2 \cdot e^{-N\rho\frac{\epsilon^2}{4}} \\ &\geq 1 - \delta \end{aligned}$$

Hence,

$$\begin{aligned}
\delta &\geq 2 \cdot e^{-N\rho\frac{\epsilon^2}{4}} \\
\implies \frac{\delta}{2} &\geq e^{-N\rho\frac{\epsilon^2}{4}} \\
\implies \frac{\delta}{2} &\geq e^{-N\rho\frac{\epsilon^2}{4}} \\
\implies \frac{N\rho\epsilon^2}{4} &\geq \ln\frac{2}{\delta} \\
\implies N &\geq \frac{4}{\epsilon^2\rho}\ln\frac{2}{\delta}
\end{aligned}$$

However, N is dependent on $\rho = \frac{|G|}{|U|}$, hence N can be exponential if ρ is exponentially small. Next, we look at an example of how we can reduce the universal set size to get a better ρ.

Example of DNF Counting

DNFs are the opposite of CNF, OR of ANDS

$$F = T_1 \vee T_2 \vee \cdots \vee T_m, T_i = L_1 \wedge L_2 \wedge \cdots \wedge L_k$$

The universe is $U = \{T, F\}^n$
We need to find the number of satisfying solutions to F

$$G = \{u \in U | F(u) = T\}$$

To count this, we make use of *biased sampling*, which is essentially reducing our universe so that we can get a more accurate estimation with lesser trials.

We define H_i as the subset of assignments that satisfy term T_i, hence we have $H_1, H_2, \ldots, H_m$
We know
$|H_i| = 2^{n-r_i}$, as there is exactly one assignment of the r_i literals in term T_i to make it satisfy, and all other $n - r_i$ literals can take any value.

$$\left| H = \bigcup_{i=1}^{m} H_i \right|$$

is the count we want, instead, we first work with a multiset union of these.
Definition of the Universal Set

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

$$|U| = \sum_{i=1}^{m} |H_i| \geq |H|$$

To make things easier, we define U as

$$U = \{(v, i) | v \in H_i\}$$

To distinguish between duplicates and where they came from.

Coverage Set

$$\mathrm{cov}(v) = \{i | (v, i) \in U\}$$

This essentially captures all the terms that would be satisfied by this assignment. A trivial bound on this is $|\mathrm{cov}(v)| \leq m$ as an assignment can't satisfy more terms than those that exist.

$$|U| = \sum_{v \in H} |\mathrm{cov}(v)|$$

Function to define G in terms of U

$f : U \to \{0, 1\}$

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

$G = \{(v, j) | f((v, j)) = 1\}$

By defining in terms of this, we essentially just take the minimum index occurrence of an assignment in case of duplicates, hence avoiding duplicates in G.

We can see that $|G| = |H|$

Lemma -

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

Proof -

$$
\begin{aligned}
|U| &= \sum_{v \in H} |\text{cov}(v)| \\
&\leq \sum_{v \in H} m \\
&= m|H| \\
&= m|G|
\end{aligned}
$$

$$
\implies \frac{|G|}{|U|} \geq \frac{1}{m}
$$

Hence, we've modified U such that the ratio has a much tighter bound in this specific counting problem.

Using the Estimator theorem, we can say that if $\rho \geq \frac{1}{m}$, then the monte carlo method yields an $\epsilon-$approximation to $|G|$ with probability $\geq 1-\delta$, provided

$$
N \geq \frac{4m}{\epsilon^2} \ln \frac{2}{\delta}
$$

Randomized algorithm for DNF

We still need to ensure that the queries $u_1, u_2, \ldots, u_n$ are still randomly sampled from U_i. The algorithm to do this is given as follows.

1. Sample $i \in [m]$ with probability $\frac{|H_i|}{|U|}$
2. Sample from $|H_i|$ uniformly. Fix the r_i bits that are there in the term, and for the rest of the bits you can just flip a coin to decide each of their values.

$$
\begin{aligned}
\mathbb{P}[(v, i) \text{ is sampled}] &= \mathbb{P}[i \text{ is sampled}] \cdot \mathbb{P}[v \text{ is sampled from } H_i | i \text{ is sampled}] \\
&= \frac{|H_i|}{|U|} \cdot \frac{1}{|H_i|} = \frac{1}{|U|}
\end{aligned}
$$

Hence, every element in U is sampled with uniform probability.