CSE 203A Monte Carlo Method

Professor Ramamohan Paturi
Scribe: Zack Chase Lipton

February 11th, 13th, 18th

1 Introduction

Often, we would like to estimate values that may be computationally infeasible to explicitly compute. The Monte Carlo method describes a set of techniques for estimating these values via sampling. One simple example is estimating the value of π. In more complex examples, sampling from the set of all possible assignments can be inefficient and one must cleverly construct sample spaces in order to produce a fast and reliable estimates of desired quantities. These notes describe the application of Monte Carlo methods to estimate the value of π, approximate the number of satisfying assignments to a DNF formula, and to count the number of independent sets in a graph. The material in these notes follows the lectures of Professor Ramamohan Paturi in CSE203A at the University of California, San Diego and additionally the details contained in chapter 10 of [1].

2 Basic Example: Estimating π

Consider the estimation of the value of π. We draw a circle about the origin with radius 1 and a square centered at the origin with edge-length 2 and edges parallel to the x and y axis. Thus any point with x value on the interval $[-1, 1]$ and y value on the interval $[-1, 1]$ is inside the square.

We choose choose points (x, y) uniformly randomly in these intervals and determine whether or not they lie inside the circle with the following algorithm.

The probability that each point lands inside the circle is equal to the proportion between the area of the circle (π), and the area of the square. Thus, because each point is drawn independently, $\mathbb{E}(Z) = \frac{m\pi}{4}$. We let $Z' = 4Z/m$. Thus $\mathbb{E}(Z') = \pi$.

2.1 Chernoff Bounds for Estimating the Value of π

However, knowing $\mathbb{E}(Z')$ to be equal to π provides no guarantee that for any given run of the algorithm, Z' takes value that is actually close to π. To analyze the concentration of Z' about its mean requires the use of Chernoff bounds.

Algorithm 1 Estimating π

```
Z ← 0
for i = 1 : m do
    (x, y) ← random()
    if x^2 + y^2 ≤ 1 then
        Z++
    end if
end for
return Z
```

We would like to know the probability that our estimate Z' is within some factor ϵ of π, thus we want to bound the probability that it lies outside that range.

$$\begin{aligned} Pr(|Z' - \pi| \geq \epsilon\pi) &= Pr\left(\left|Z - \frac{m\pi}{4}\right| \geq \frac{\epsilon m\pi}{4}\right) \\ &= Pr(|Z - \mathbb{E}(Z)| \geq \epsilon\mathbb{E}(Z)) \\ &\leq 2e^{-m\pi\epsilon^2/12} \end{aligned}$$

Thus for a given confidence parameter δ, we must choose m large enough so that $2e^{-m\pi\epsilon^2/12} \leq \delta$:

$$m \geq \frac{12\ln(2/\delta)}{\pi\epsilon^2}$$

3 Approximate-Uniform Sampling

Let x_i ($1 \leq i \leq n$) be iid 0-1 random variables such that $Pr(x_i) = p$. Let $X = \sum_1^n x_i$ Then $\mathbb{E}(X) = pn$ Applying Chernoff bound:

$$Pr(x \geq (1+\epsilon)pn) \leq e^{-\epsilon^2 pn/3}$$

$$Pr(x \leq (1-\epsilon)pn) \leq e^{-\epsilon^2 pn/2}$$

A sampling algorithm generates an ϵ-uniform sample w from Ω if $\forall S \subseteq \Omega, |Pr(w \in S) - \frac{|S|}{|\Omega|}| \leq \epsilon$.

4 The DNF Counting Problem

Definition 4.1. *$A(x)$ is a fully polynomial randomized approximation scheme for a function $P(x)$ given parameters ϵ, δ and some input x, where $\epsilon > 0$ and $\delta < 1$ returns an (ϵ, δ) approximation to $P(x)$ in time polynomial to $|x|$, $\frac{1}{\epsilon}$, $\ln(\frac{1}{\delta})$.*

We want to count the number of satisfying assignments of of a Boolean formula in disjunctive normal form (DNF). A DNF formula is a disjunction of clauses $(C_1, ..., C_l)$ where each clause is a conjunction of literals. For example:

$$F(\boldsymbol{x}) = (x_1 \wedge \bar{x}_2 \wedge x_3) \vee (x_2 \wedge x_4) \vee (\bar{x}_1 \wedge x_3 \wedge x_4)$$

We define $\#F$ to be the number of satisfying assignments of our DNF formula. It seems sensible to sample random assignments from some sample space and infer from the fraction of such assignments that satisfy the formula, the value of $\#F$. A naive choice of sample space would be set of all possible assignments, $\Omega = \{0,1\}^n$ with cardinality $|\Omega| = 2^n$.

4.1 Simplistic Approach

To estimate π, we drew samples from the set of all points in the feasible interval and based our estimate on the fraction of points that fell inside the circle. To estimate the number of satisfying assignments of the DNF counting problem, we could proceed analogously, drawing samples uniformly randomly from the set of all assignments.

Algorithm 2 Estimating the Number of Satisfying Assignments

```
Generate a random sample x = (x_1, ..., x_n).
if F(x) is true then
    y_i = 1
else
    y_i = 0.
end if
y = \sum_{i=1}^m y_i
y' = 2^n y / m
```

Clearly, $\mathbb{E}(y_i) = \frac{\#F}{|\Omega|} = \frac{\#F}{2^n}$. Because the expectation of the sum equals the sum of the expectations, $\mathbb{E}(y) = \sum_{i=1}^m \mathbb{E}(y_i) = \frac{m\#F}{2^n}$. Therefore, $\mathbb{E}(y') = \#F$. However it is insufficient to show that the expectation of y' is equal to the expectation of $\#F$. As with the estimation of π we must provide a (ϵ, δ) approximation.

4.2 Chernoff Bounds for Simplistic Approach to DNF

We want to lower bound the probability that our approximation y' for $\#F$ is correct within a factor ϵ.

$$Pr[|y' - \#F| \geq \epsilon\#F] = Pr\left[\left|y - \frac{\#F}{2^n}\right| \geq \epsilon m \frac{\#F}{2^n}\right]$$

Applying union bound:

$$\leq 2e^{\frac{-\epsilon^2 m\#F/2^n}{3}}$$

Thus, to provide an (ϵ, δ) approximation, we must choose $m = 3/\epsilon^2 \cdot ln\frac{2}{\delta}\frac{2^n}{\#F}$. This would require exponentially many samples from ω with respect to the size of n. Thus to achieve an FPRAS, we must construct a different sample space.

4.3 FPRAS for DNF Counting

To produce a fully polynomial randomized approximation scheme we must consider other sample spaces. Specifically, we must find one such sample space Ω' so that $\frac{\#F}{|\Omega'|}$ is sufficiently large.

Our new sampling procedure considers that every clause can be satisfied, i.e., does not include both a variable and its negation. Any such clause is never satisfiable and thus can be removed from the formula. If clause C_i has l_i literals, there must be 2^{n-l_i} assignments which satisfy C_i. Letting SC_i denote this set of assignments which satisfy C_i and

We let $\Omega' = \{(i, x)|1 \leq i \leq n, x \in SC_i\}$. We want to estimate $\#F = \bigcup_{i=1}^{t} SC_i$. $|\Omega'|$ may be larger than $\#F$ because any given assignment might satisfy more than one clause. Thus we define a subset $S \subseteq \Omega'$ by selecting only one pair for each assignment. We do this by choosing the pair (i, x) which occurs first, i.e., for which i is lowest.

$$S = \{(i, x)|x \in \{0, 1\}^n 1 \leq i \leq l\}$$

We estimate $\#F$ with the following algorithm([1]).

Algorithm 3 Estimating the Number of Satisfying Assignments for DNF

$c \leftarrow 0$.
for $k = 1 : m$ **do** with probability $|SC_i|/\sum_{i=1}^{t}|SC_i|$, choose uniformly at random an assignment $x \in SC_i$.
  if x is not in any $SC_j, j < i$ **then**
    $c \leftarrow c + 1$
  end if
end for

As $|\Omega'|$ is known, we need only to approximate the ratio of S to $|\Omega'|$ in order to estimate the $|\Omega|$ and thus #F. Because any assignment can satisfy at most l clauses, this ratio must be at least $1/l$.

$$\mathbb{E}(y_i) = Pr[y_i = 1] = \frac{\#F}{\Omega'} \geq \frac{1}{l}$$

5 Estimating Number of Independent Sets

Definition 5.1. *We let w be the output of a sampling algorithm for a finite sample space Ω. If for an subset $|S|$ of Ω, $\left|Pr(w \in S), \frac{|S|}{|\Omega|}\right| \leq \epsilon$, the sampling algorithm generates an ϵ-uniform sample. If it additionally, runs in time polynomial in $ln\epsilon^{-}1$ and the size of input x, it is a fully polynomial almost uniform sampler (FPAUS).*

We consider the example of estimating the number of independent sets in a graph $G = (V, E)$. A uniform sampling takes as input our graph G and a parameter ϵ, outputting an ϵ-uniform sample in time polynomial with respect to $\epsilon^{-}1$. We demonstrate that given an FPAUS for sampling independent sets, we can produce an FPRAS for estimating the counting their number.

We assume our edgeset E contains m edges and impose on them some arbitrary ordering $e_1, ..., e_m$. By E_i, we denote the set of the first i edges in E and we define $G_i = (V, E_i)$. $\Omega(G_i)$ is the set of independent sets in the graph G_i.

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

By r_i, we express the ratio between the number of independent sets in consecutive graphs.

$$r_i = \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|}, i = 1, ..., m$$

Substituting into equation 1, gives

$$\Omega(G) = 2^n \prod_{i=1}^{m} r_i$$

.

Therefore good estimates of the ratios of $|\Omega(G_i)|$ to $|\Omega(G_{i-1})|$ for all i combined with the knowledge that $|\Omega(G_0)| = 2^n$ can be used to approximate $\Omega(G_m)$. We label our estimated ratios $\tilde{r}_i$. And thus our estimate of ΩG is expressed

$$2^n \prod_{i=1}^{m} \tilde{r}_i$$

.

To capture an (ϵ, δ) approximation of the$|\Omega(G)|$, we must bound the expression

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

.

Lemma 5.2. *Suppose that for all* $i, 1 \leq i \leq m, \tilde{r}_i$ *is an* $(\epsilon/2m, \delta/m)$*-approximation for* r_i*. Then*

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

Proof. For all i such that $1 \leq i \leq m$,

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

Applying the union bound, the probability that for any i, $|\tilde{r} - r_i| > (\epsilon/2m)r_i$ is at most δ. Combining these bounds yields:

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

□

Thus, given some method for producing an $(\epsilon/2m, \delta/m)$-approximation for each r_i, we can produce an (ϵ, δ) approximation of the number of independent sets in G. Using the Monte Carlo method and an FPAUS for sampling independent sets, we can produce such an approximation of r_i. The algorithm below uses an FPAUS and the Monte Carlo method to produce such an approximation of r_i.

Algorithm 4 Estimating r_i

Input: Graphs G_{i-1} and G_i
Output: $\tilde{r}_i$, an approximation for r_i
$X \leftarrow 0$
for $M = 1 : 1296m^2\epsilon^{-2}ln(2m/\delta)$ **do**
    Generate an $(\epsilon/6m)$-uniform sample from $\Omega(G_{i-1})$
    if The sample is in the independent set of G_i **then**
        $X \leftarrow X + 1$
    end if
end for
return $\tilde{r}_i \leftarrow X/M$

5.1 Correctness of Algorithm 4

Lemma 5.3. *When $m \geq 1$ and $0 \leq \epsilon \leq 1$, algorithm 4 produces and $(\epsilon/2m, \delta/m)$-approximation for r_i.*

Proof. First, we lower bound r_i by 1/2. This is clear because the introduction of an edge introduces one pair of vertices (u, v) that can no longer coexist in an independent set. By considering that any independent set containing both u and v has a corresponding independent set containing only u and not v. By simply removing v from consideration we eliminate at most half of the independent sets, i.e., $r_i \geq 1/2$.

By the definition of an FPAUS and because our samples are produced by a $(\epsilon/6m)$-uniform sampler,

$$\left| Pr(X_k = 1) - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} \right| \leq \frac{\epsilon}{6m},$$

Where each X_k is true if the kth sample from $\Omega(G_{i-1})$ is contained in $\Omega(G_i)$. Because each X_k is a Bernoulli random variable,

$$\left| \mathbb{E}(X_k) - \frac{|\Omega(G_i)|}{|\Omega(G_{i-1})|} \right| \leq \frac{\epsilon}{6m}.$$

Therefore,

$$\mathbb{E}(\tilde{r}_i) - r_i \leq \frac{\epsilon}{6m}.$$

Because with enough samples, $\tilde{r}_i$ is close to $\mathbb{E}(\tilde{r}_i)$ and $\mathbb{E}(\tilde{r}_i)$ is close to r_i, and because r_i is lower bounded by $1/2$, we can show that

$$\mathbb{E}(\tilde{r}_i) \geq r_i - \frac{\epsilon}{6m} \geq \frac{1}{2} - \frac{\epsilon}{6m} \geq \frac{1}{3}$$

We now apply the following theorem: [1] Given independent and identically distributed indicator random variables $X_1, ..., X_m$, and $m \geq (3ln(2/\delta))/\epsilon^2\mathbb{E}(X_i))$

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

Therefore if

$$M \geq \frac{3ln(2m/\delta)}{(\epsilon/12m)^2(1/3)} = 1296m^2\epsilon^{-2}ln\frac{2m}{\delta},$$

then

$$Pr\left(\left|\frac{\tilde{r}_i}{\mathbb{E}(\tilde{r}_i)} - 1\right| \geq \frac{\epsilon}{12m}\right) =\leq \frac{\delta}{m},$$

i.e., with probability $1 - \frac{\delta}{m}$,

$$1 - \frac{\epsilon}{12m} \leq \frac{\tilde{r}_i}{\mathbb{E}(\tilde{r}_i)} \leq 1 + \frac{\epsilon}{12m} \tag{2}$$

Because $r_i \geq 1/2$,

$$1 - \frac{\epsilon}{3m} \leq \frac{\mathbb{E}(\tilde{r}_i)}{r_i} \leq 1 + \frac{\epsilon}{3m}. \tag{3}$$

Combining equation 2 and equation 3, with probability $1 - \delta/m$,

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

This is the $(\epsilon/2m, \delta/m)$ approximation of r_i that we require. Combined with Lemma 5.3, this gives us an (ϵ, δ) approximation of the number of independent set in time polynomial to $|x|$, $\frac{1}{\epsilon}$, and $ln\frac{1}{\delta}$. □

References

[1] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis.* Cambridge University Press, 2005.