CSE 521: Design and Analysis of Algorithms I	Fall 2023

Lecture 5: Unbiased Estimators, Streaming

Lecturer: Shayan Oveis Gharan — *10/12/2023*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

In previous lectures, we showed that by Chebyshev's inequality, any random variable has chance at least $1 - \frac{1}{k^2}$ of taking a value in interval $[\mu - k\sigma, \mu + k\sigma]$, where μ, σ are the mean and standard deviation, respectively. If we take t independent samples (sometimes pairwise independent is also enough), then the variance of the sample average is σ^2/t. Hence by increasing t, we can a better estimate of μ. How many samples do we need to get a good estimate of μ? In particular, to get ϵ-additive approximation of μ with probability $1 - \delta$ it is enough to use $O(1/\delta\epsilon^2)$ many independent samples.

An ϵ-additive approximation is not desirable in many applications because the range of the ϵ may be independent of the magnitude of μ. For example, if μ is in the interval $[0.001, 0.002]$, a 0.1-additive approximation to μ has no information. Instead, a multiplicative approximation scales proportional to the magnitude of μ. In the next section we will see how many samples we need to obtain a $1 \pm \epsilon$ approximation to μ.

5.1 Unbiased Estimators

We say a random variable X is an unbiased estimator of μ if

$$\mathbb{E}[X] = \mu.$$

It turns out the the number of samples is proportional to the relative variance of X.

Definition 5.1 (Relative Variance)**.** *Say X is an unbiased estimator of μ, then, the relative variance of X is defined as*

$$\frac{\sigma^2(X)}{\mu^2}, \tag{5.1}$$

where by $\sigma^2(X) = \mathbb{E}[X]^2 - (\mathbb{E}[X])^2$ is the variance of X. We typically use t to denote the relative variance.

The following theorem is the main result of this section.

Theorem 5.2. *Given $\epsilon, \delta > 0$, and an unbiased estimator of μ, X. We can approximate μ within $1 \pm \epsilon$ multiplicative factor using only $O(\frac{t}{\epsilon^2} \log \frac{1}{\delta})$ independent samples of X with probability $1 - \delta$.*

Before going into the details of the proof let us discuss a motivating example.

Dart throwing method of estimating areas. Suppose we want to estimate the area of a closed curve on the plane (see curve A of section 5.1). We can use the well-known Monte Carlo method. The idea is to draw a rectangle B that includes A. Then, we randomly sample a point in B. Let

$$X = \begin{cases} 1, & \text{if the point is belong to } B \\ 0, & \text{otherwise} \end{cases}$$

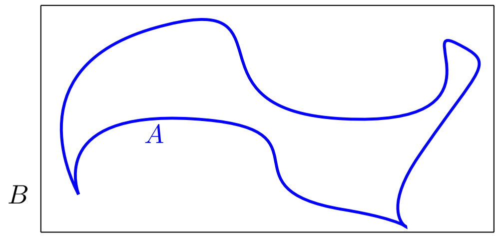

Figure 5.1: Estimating Area by Monte Carlo Method

Observe that $\mathbb{E}X = \mathbb{P}[X = 1] = s(A)/s(B)$, where $s(.)$ denotes the surface area function. Since we can exactly calculate $s(B)$, we can use $Y = s(B)X$ as an unbiased estimator of $s(A)$. Now, we can use Theorem 5.2 to find the number of independent samples of X that we need to estimate Y within a $1 \pm \epsilon$ factor.

All we need to know is that relative variance of X; we have

$$t = \frac{\sigma^2(Y)}{\mu^2} = \frac{s(B)^2\sigma^2(X)}{s(A)^2} \leq \frac{s(B)^2\mathbb{E}\left[X^2\right]}{s(A)^2} = \frac{s(A)\cdot s(B)}{s(A)^2} = \frac{s(B)}{s(A)} \tag{5.2}$$

where we used that X is a Bernoulli random variable with prior $s(A)/s(B)$. So, we need $O(\frac{s(B)}{s(A)} \cdot \frac{\log\frac{1}{\delta}}{\epsilon^2})$ independent samples of X to find an $1 \pm \epsilon$ approximation of $s(A)$ with probability $1 - \delta$. Note that we need to sample X t many times (in expectation) to just get 1 point that belongs to A.

Remark. The above Monte Carlo method is a fairly general method to estimate a quantity of interest. Suppose we have an object X that we want to measure. The idea is to find a bigger object Y that contains X such that (i) We can measure Y and (ii) We can generate samples from Y. Then, we can use the above method to approximately measure X. As we discussed the number of samples that we need is proportional to the measure of Y with respect to X, up to a $\log(1/\delta)/\epsilon^2$ factor.

Proof of Theorem 5.2. First, we give an algorithm that estimates μ within $1\pm\epsilon$ factor with probability 9/10 using only $O(t/\epsilon^2)$ samples. Then, we show how we can boost the success probability to $1 - \delta$ using the "median trick".

Let $X_1, \ldots, X_k$ be k independently chosen samples of X. Since X is an unbiased estimator, for all i, $\mathbb{E}[X_i] = \mu$. Let $Y = \frac{1}{k}(X_1 + \cdots + X_k)$ be the average of X_i's; by linear property of expectation $\mathbb{E}[Y] = \mu$. So, by Chebyshev's inequality, we have

$$\begin{aligned} \mathbb{P}[(1-\epsilon)\mu \leq Y \leq (1+\epsilon)\mu] &= \mathbb{P}[|Y - \mu| \leq \epsilon\mu] \\ &\geq 1 - \frac{\sigma^2(Y)}{\epsilon^2\mu^2} \qquad (5.3) \\ &= 1 - \frac{\sigma^2(X)}{\epsilon^2 k\mu^2} = 1 - \frac{t}{k\epsilon^2}. \qquad (5.4) \end{aligned}$$

Hence, by taking $k = O(\frac{t}{\epsilon^2})$, samples, we can get a $1 \pm \epsilon$ approximation of μ with probability 9/10.

To obtain $\log\frac{1}{\delta}$ probability of success we need to use Chernoff type of bounds. However, these bounds usually need some specific assumption on the distribution of the random variables that we average out, e.g., that the third or fourth moments are bounded. In our particular case, we have no prior assumption on the distribution of X. We only have a handle on the expectation and variance of X because we know the relative variance.

The idea is to use a trick called "median trick". Fix, $k = O(t/\epsilon^2)$, such that

$$\mathbb{P}\left[(1-\epsilon)\mu \le Y \le (1+\epsilon)\mu\right] \ge 9/10. \tag{5.5}$$

This follows simply from (5.4). We output the median value from the ℓ independent samples of Y. Call these samples, $Y_1, \ldots, Y_\ell$. Observe that the median of Y_i's will be in the interval $[(1-\epsilon)\mu, (1+\epsilon)\mu]$ if at least half of Y_i's are in this interval $[(1-\epsilon)\mu, (1+\epsilon)\mu]$.

We show that the probability that half of the Y_i's are outside this interval is very small. Define

$$Z_i := \mathbb{I}\left[Y_i \in [(1-\epsilon)\mu, (1+\epsilon)\mu]\right]$$

be the random variable indicating that Y_i is in $[(1-\epsilon)\mu, (1+\epsilon)\mu]$. Note that by (5.5) for each i, $\mathbb{P}[Z_i] \ge 9/10$. By linearity property of expectation, we have $\mathbb{E}\left[\sum_i Z_i\right] = \sum_i \mathbb{E} Z_i \ge \frac{9\ell}{10}$. By Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^{\ell} Z_i \le \frac{\ell}{2}\right] &\le \mathbb{P}\left[\left|\sum_{i=1}^{\ell} Z_i - \mathbb{E}\left[\sum_{i=1}^{\ell} Z_i\right]\right| > \frac{\ell}{4}\right] && (5.6)\\ &\le e^{-\ell/8} && (5.7)\end{aligned}$$

where in the first inequality we used that $\mathbb{E}\left[\sum_i Z_i\right] \ge 9\ell/10$. Choosing ℓ such that $e^{-\ell/8} \le \delta$, i.e., $\ell = O(\log 1/\delta)$, we only need $O(t \log \frac{1}{\delta}/\epsilon^2)$ samples of X to obtain a $1 \pm \epsilon$ approximation of μ with probability at least $1-\delta$. □

5.2 Introduction to Streaming Algorithms

As an application of hashing and the unbiased estimator, we are going to discuss streaming algorithms. Streaming algorithms has become a hot topic in computer science nowadays because of the massive amount of data that we have to process. Typically, we do not have enough space to store the entire data. Instead, we process the data in a streaming fashion, and sketch the information we want from the data by a few passes.

We will talk about algorithms for F_0 and F_2 estimation. Those are classic results appeared in the first paper of streaming algorithms [AMS96]. The problem is as follows.

Let $\mathcal{U} = \{1, ..., |\mathcal{U}|\}$ be a large universe of numbers, and let $X_1, ..., X_n$ be a sequence of numbers in $\mathcal{U}$. Let $f_i = \sum_{j=1}^{n} \mathbb{I}[X_j = i]$ be the number of times i appears in the sequence. For $0 \le k \le \infty$, we let $F_k = \sum_{i=1}^{|\mathcal{U}|} f_i^k$, where we define $0^0 = 0$. The interesting values of k for us are

- When $k = 0$, F_0 counts the number of distinct elements in the sequence.
- When $k = 2$, F_2 is the second moment of the vector $(f_1, ..., f_{|\mathcal{U}|})$.
- When $k = \infty$, F_∞ corresponds to the number of times the most frequent number shows up in the sequence.

The following theorem is proven in [AMS96]

Theorem 5.3. *There is a streaming algorithm that for any sequence $x_1, \ldots, x_n$ of the universe $\{1, 2, \ldots, |U|\}$ gives a $(1-\epsilon)$ approximation of F_0 and F_2 using $O(\frac{\log|U| + \log n}{\epsilon^2} \cdot \log \frac{1}{\delta})$ space with probability $1-\delta$.*

Here, we only give an algorithm for F_2 and we leave the algorithm for F_0 as a homework exercise.

We remark that allowing randomness and approximated solution is crucial. There is no hope to use a deterministic or exact algorithm to achieve logarithmic amount of space. Please see Tim Roughgarden's Lecture notes for more details.

References

[AMS96] N. Alon, Y. Matias, and M. Szegedy. "The space complexity of approximating the frequency moments". In: *STOC*. ACM. 1996, pp. 20–29 (cit. on p. 5-3).