

Lecture 5: Unbiased Estimators, Streaming

Lecturer: Shayan Oveis Gharan

January 23rd

Scribe: Ning Li

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 get a better estimate of μ . How many samples do we need to get a good estimate of μ ? We answer this question in the following section. Then, we use the solution to discuss a streaming algorithm for the second moment problem.

5.1 Unbiased Estimators

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

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

In this section we will see how many samples we need to approximate μ within $1 \pm \epsilon$ multiplicative factor.

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

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

Observe that $\mathbb{E}X = \mathbb{P}[X = 1] = s(A)/s(B)$, where $s(\cdot)$ denotes the surface area function. Since we can

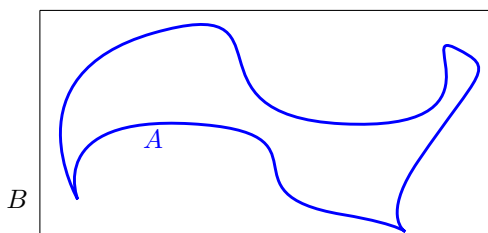


Figure 5.1: Estimating Area by Monte Carlo Method

exactly calculate $s(B)$, we can use $s(B)X$ is 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 $s(A)$ within a $1 \pm \epsilon$ factor.

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

$$t = \frac{\sigma^2(X)}{\mu^2} = \frac{s(A)/s(B)}{(s(A)/s(B))^2} = \frac{s(B)}{s(A)} \quad (5.2)$$

where we used that X is a Bernoulli random variable with prior $s(A)/s(B)$. So, we need $O(s(B) \log \frac{1}{\delta} / \epsilon^2 s(A))$ 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 .

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

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

$$= 1 - \frac{\sigma^2(X)}{\epsilon^2 k \mu^2} = 1 - \frac{t}{k\epsilon^2}. \quad (5.4)$$

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. In our particular case, we have no prior assumption on the distribution of X . We only assume that $\mathbb{E}X = \mu$. So, we will use a trick called “median trick”.

Fix, $k = O(t/\epsilon^2)$, such that

$$\mathbb{P}[(1 - \epsilon)\mu \leq Y \leq (1 + \epsilon)\mu] \geq 9/10. \quad (5.5)$$

This follows simply from (5.4). We modify previous algorithm by getting the median value from the ℓ independent estimates of Y . Call these samples, Y_1, \dots, Y_ℓ . 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

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

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

$$\mathbb{P}\left[\sum_{i=1}^{\ell} Z_i \leq \frac{\ell}{2}\right] \leq \mathbb{P}\left[\left|\sum_{i=1}^{\ell} Z_i - \mathbb{E}\left[\sum_{i=1}^{\ell} Z_i\right]\right| > \frac{\ell}{4}\right] \quad (5.6)$$

$$\leq e^{-\ell/8} \quad (5.7)$$

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

5.2 F_2 moment

We'll talk about algorithm for F_2 estimation, which is a classic streaming problem, detailed information can be referred in [AMS96]. The problem setup is as following:

Let $U = \{1, \dots, |U|\}$ be a large universe of numbers, and let X_1, \dots, X_n be a sequence of numbers in U . Let $f_i = \sum_{j=1}^n \mathbb{I}[X_j = i]$ be the number of times i appears in the sequence. We let

$$F_2 = \sum_{i=1}^{|U|} f_i^2$$

be the second moment of the vector $(f_1, \dots, f_{|U|})$.

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

Before discussing ideas to prove the above theorem, let us first discuss a natural idea for streaming problems. Since we are dealing with a "big data" problem, we may first down sample the input into a smaller length, then we calculate the second moment of the down sample and we use it to estimate the second moment of the original input. Consider the following set of two inputs.

$$\begin{array}{c} 1, 2, 3, 4, \dots, n \\ \underbrace{1, 1, \dots, 1}_{m \text{ times}}, \underbrace{2, 2, \dots, 2}_{m \text{ times}}, \dots, \underbrace{n/m, n/m, \dots, n/m}_{m \text{ times}} \end{array}$$

where $m = \Omega(\sqrt{n})$. Observe that any down samples of the first sequence gives completely distinct numbers, and any down sample of the second sequence of size $O(\sqrt{n})$ also gives completely distinct numbers with a constant probability. So, any streaming algorithm that is based on down sampling sees almost the same thing, i.e., completely distinct elements, in both cases. However, the second moment of the first sequence is n and the second moment of the second one is $O(n^{3/2})$, so we don't expect a streaming algorithm based on down sampling to size at most $O(\sqrt{n})$ obtain an estimate better than \sqrt{n} of the true second moment.

Our main plan for the proof is that we design an unbiased estimator for F_2 that uses $O(\log |U| + \log n)$ amount of memory and has a relative variance of $O(1)$. Then, by Theorem 5.2 we only need $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ independent samples of our unbiased estimator; so it is enough to run $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ independent copies of our algorithm on the input sequence and run the algorithm of Theorem 5.2 on the output of these independent copies.

Before designing algorithm for estimating F_2 , let's revisit the random walk example that we talked about few lectures ago. Denote random variable X_i

$$X_i = \begin{cases} +1, & \text{w.p. } 1/2 \\ -1, & \text{w.p. } 1/2 \end{cases}$$

Let $X = \sum_i X_i$. Using the Hoeffding bound, we have showed that for any constant,

$$\mathbb{P}[|X| \geq c\sqrt{n}] \leq e^{-\Omega(c^2)}.$$

We will show that $X \geq \Omega(\sqrt{n})$ with a constant probability, that is $\mathbb{P}[X \geq \Omega(\sqrt{n})] \geq \Omega(1)$. This conclusion follows from the central limit theorem (CLT). Because $\frac{1}{\sqrt{n}} \sum_i X_i \rightarrow_{n \rightarrow \infty} \mathcal{N}(0, 1)$. So,

$$\sum_i X_i \rightarrow_{n \rightarrow \infty} \mathcal{N}(0, \sqrt{n}).$$

Therefore, we expect $\sum_i X_i$ to be (almost) uniform in the interval $[-\sqrt{n}, \sqrt{n}]$. So, 0 the particle is almost uniformly distributed in the interval $[-\sqrt{n}, \sqrt{n}]$.

Here, we prove this fact without using CLT. Most importantly, our argument only uses low moments $\sum_i X_i$, so unlike CLT we don't need full independent. We show that $\mathbb{E}[X^2] \geq n$, which is enough to show $X \geq \Omega(\sqrt{n})$ with a constant probability:

$$\begin{aligned} \mathbb{E}[X^2] &= \mathbb{E} \left[\left(\sum_i X_i \right)^2 \right] \\ &= \mathbb{E} \left[\sum_{i,j} X_i X_j \right] \\ &= \sum_{i,j} \mathbb{E}[X_i X_j] \\ &= \sum_i \mathbb{E}[X_i^2] + \sum_{i \neq j} \mathbb{E}[X_i X_j] = \\ &= \sum_i 1 + \sum_{i \neq j} \mathbb{E}[X_i] \mathbb{E}[X_j] = n \end{aligned} \tag{5.8}$$

where we used the pairwise independence between X_i and X_j .

Back to estimating F_2 problem, we will apply a similar idea. Choose a pairwise independent hash function $h : U \rightarrow \{-1, 1\}$. Eventually we will see that we need to make h 4-wise independent. We start the algorithm letting $Y = 0$; when we read X_i , we'll update Y ,

$$Y \leftarrow Y + h(X_i).$$

Note that at the end of the algorithm $Y = \sum_{i=1}^n h(X_i)$. We claim that Y^2 is the unbiased estimator of F_2 .

Claim 5.4. Y^2 is an unbiased estimator of F_2 , i.e.,

$$\mathbb{E}[Y^2] = F_2 = \sum_{i=1}^{|U|} f_i^2.$$

Before proving the claim let us consider two special cases. First suppose $X_1 = 1, X_2 = 2, \dots, X_n = n$. Then, $\sum_i h(X_i)$ can be seen as a (pairwise independent) random walk of length n started from 0. So, by (5.8), $\mathbb{E}[Y^2] = n$ which is the same as F_2 . Note that in (5.8) we only use pairwise independence of X_i, X_j (for all i, j).

For the second example, suppose $X_1 = X_2 = \dots = X_n = 1$. Then, $f_2 = n^2$ and $Y = nh(1)$. So,

$$Y^2 = n^2 h(1)^2 = n^2 = f_2.$$

Now, we are ready to prove the claim.

Proof. First, observe that we can write $Y = \sum_i f_i h(i)$. Therefore,

$$\mathbb{E}[Y^2] = \mathbb{E}\left[\left(\sum_i f_i h(i)\right)^2\right] \quad (5.9)$$

$$= \mathbb{E}\left[\sum_{i,j} f_i f_j h(i) h(j)\right] \quad (5.10)$$

$$= \sum_{i,j} f_i f_j \mathbb{E}[h(i) h(j)] \quad (5.11)$$

$$= \sum_i f_i^2 \mathbb{E}[h(i)^2] + \sum_{i,j} f_i f_j \mathbb{E}[h(i)] \mathbb{E}[h(j)] \quad (5.12)$$

$$= \sum_i f_i^2 + 0 = F_2, \quad (5.13)$$

where the second to last equality follows by pairwise independent of h . Note that the expectations are over random choices for our pairwise independent hash function h . Therefore, Y^2 is the unbiased estimator of F_2 . \square

Now, by [Theorem 5.2](#), we upper bound the relative variance of Y^2 . First, we show that $\sigma^2(Y^2) \leq 2(F_2)^2$. Then, we show that the relative variance of Y^2 is $O(1)$. In the proof of the next claim we use that h is 4-wise independence.

Claim 5.5. $\sigma^2(Y^2) \leq 2(F_2)^2$.

Proof. Recall that $\sigma^2(Y^2) = \mathbb{E}Y^4 - (\mathbb{E}Y^2)^2$. So, first we upper bound $\mathbb{E}Y^4$.

$$\mathbb{E}Y^4 = \mathbb{E}\sum_i f_i h(i)^4 \quad (5.14)$$

$$= \mathbb{E}\sum_{i,j,k,l} f_i f_j f_k f_l h(i) h(j) h(k) h(l) \quad (5.15)$$

$$= \sum_{i,j,k,l} f_i f_j f_k f_l \mathbb{E}[h(i) h(j) h(k) h(l)] \quad (5.16)$$

$$= \sum_{i \neq j} f_i^2 f_j^2 \binom{4}{2} + \sum_i f_i^4 \quad (5.17)$$

Note that in the last equality we used 4-wise independence of h . In particular if any of i, j, k, l appear an odd number of times then $\mathbb{E}[h(i) h(j) h(k) h(l)]$ is zero. So, the only case that it is nonzero is if i appears 4

times or i, j each appear 2 times. Therefore,

$$\sigma^2(Y^2) = 6 \sum_{i \neq j} f_i^2 f_j^2 + \sum_i f_i^4 - \left(\sum_i f_i^2 \right)^2 \quad (5.18)$$

$$= 4 \sum_{i \neq j} f_i^2 f_j^2 \quad (5.19)$$

$$\leq 2 \left(\sum_i f_i^2 \right)^2 \quad (5.20)$$

$$= 2(F_2)^2 \quad (5.21)$$

as desired. \square

It follows from the above claim that the relative variance of Y^2 is

$$\frac{\sigma^2(Y^2)}{(F_2)^2} \leq \frac{2F_2^2}{F_2^2} = 2.$$

So, by [Theorem 5.2](#) we only need $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ independent samples of Y to give a $1 \pm \epsilon$ approximation of F_2 .

Now, let us discuss the final algorithm. We choose $\frac{20}{\epsilon^2} \log \frac{1}{\delta}$ independent hash functions; For $k = 20/\epsilon^2$, let h_1, \dots, h_k be k of those functions. Start with $Y_1 = Y_2 = \dots = Y_k = 0$, after reading X_i , let $Y_j = Y_j + h_j(X_i)$ for all $1 \leq j \leq k$. Then, $\frac{1}{k} \sum_j Y_j$ gives a $1 \pm \epsilon$ approximation of F_2 with probability $9/10$. To get $1 - \delta$ probability of success we just need to run the above idea on $\log \frac{1}{\delta}$ independent copies and return the median.

The total amount of memory that we use is as follows. We need to use $O(\log n)$ amount of memory to store each Y ; we need to use $O(\log |U|)$ amount of memory to store a 4-wise independent hash function. So, for each copy we need $O(\log n + \log |U|)$ amount of memory. Since we are using $O(\frac{1}{\epsilon^2} \log \frac{1}{\delta})$ independent copies we use $O(\frac{\log n + \log |U|}{\epsilon^2} \log \frac{1}{\delta})$ bits of memory as desired. This completes the proof of [Theorem 5.3](#). More details can be referred to [\[AMS96\]](#).

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 pp. [5-3](#), [5-6](#)).