Chapter 26

Frequency Estimation over a Stream

> "See? Genuine-sounding indignation. I programmed that myself. It's the first thing you need in a university environment: the ability to take offense at any slight, real or imagined."
>
> Robert Sawyer, Factoring Humanity

By Sariel Har-Peled, March 19, 2024[①]

26.1. The art of estimation

26.1.1. The problem

Assume we would like to estimate well some quantity $\rho > 0$ - specifically, for a fixed parameter $\varepsilon \in (0,1)$, we would like to compute a quantity ρ' such that $\rho' \in [(1-\varepsilon)\rho, (1+\varepsilon)\rho]$ with good probability. To this end, assume we have access to a distribution $\mathcal{D}$, such that if we sample X according to this distribution (i.e., $X \sim \mathcal{D}$), we have that $\mathbb{E}[X] = \rho$. We can use X to estimate our desired quantity, but this might not provide the desired estimation.

Example: Estimating p for a coin. Assume we have a coin that is head with probability p. A natural way to estimate p is to flip the coin once and return 1 if it is head, and zero otherwise. Let X be the result of the coin flip, and observe that $\mathbb{E}[X] = p$. But this is not very useful estimator.

26.1.2. Averaging estimator: Success with constant probability

26.1.2.1. The challenge

The basic problem is that $X \sim \mathcal{D}$ might be much bigger than ρ. Or more specifically, its variance might be huge, where $\mathcal{D}$ is a distribution we have access to. Let

$$\rho = \mathbb{E}[\mathcal{D}] \qquad \text{and} \qquad \nu = \mathbb{V}[\mathcal{D}]\,.$$

We would to generate a variable Z, such that

$$\mathbb{E}[Z] = \rho \qquad \text{and} \qquad \mathbb{V}[Z] \le (\varepsilon^2/4)\rho^2. \tag{26.1}$$

This would imply by Chebychev's inequality that

$$\mathbb{P}[|Z-\rho| \ge \varepsilon\rho] = \mathbb{P}\Big[|Z-\mathbb{E}[Z]| \ge 2\sqrt{(\varepsilon^2/4)\rho^2}\Big] \le \mathbb{P}\Big[|Z-\mathbb{E}[Z]| \ge 2\sqrt{\mathbb{V}[Z]}\Big] \le \frac{1}{4}.$$

[①]This work is licensed under the Creative Commons Attribution-Noncommercial 3.0 License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

26.1.2.2. Taming of the variance

The basic idea is to take $\alpha = \left\lceil \frac{\nu}{\rho} \right\rceil$ independent variables $X_1, \ldots, X_\alpha \sim \mathcal{D}$, and let $Y = \sum_i X_i/\alpha$. We have by linearity of expectation that

$$\mathbb{E}[Y] = \sum_i \mathbb{E}[X_i]/\alpha = \mathbb{E}[X] = \rho.$$

Using the independence of $X_1, \ldots, X_\alpha$, we have

$$\mathbb{V}[Y] = \mathbb{V}\left[\sum_i X_i/\alpha\right] = \frac{1}{\alpha^2}\mathbb{V}\left[\sum_i X_i\right] = \frac{1}{\alpha^2}\sum_i \mathbb{V}[X_i] = \frac{1}{\alpha^2}\alpha\nu = \frac{\nu}{\alpha}.$$

Guided by Eq. (26.1), we want this quantity to be smaller than $\leq (\varepsilon^2/4)\rho^2$. Thus,

$$\frac{\nu}{\alpha} \leq (\varepsilon^2/4)\rho^2 \qquad \Longleftarrow \qquad \alpha \geq \left\lceil \frac{4}{\varepsilon^2} \cdot \frac{\nu}{\rho^2} \right\rceil = \left\lceil \frac{4\,\mathbb{V}[X]}{\varepsilon^2\,\mathbb{E}[X]^2} \right\rceil.$$

We thus summarize the result.

Lemma 26.1.1. *Let $\mathcal{D}$ be a non-negative distribution with $\rho = \mathbb{E}[\mathcal{D}]$ and $\nu = \mathbb{V}[\mathcal{D}]$, and let $\varepsilon \in (0,1)$ be a parameter. For $\alpha \geq \left\lceil \frac{4\,\mathbb{V}[\mathcal{D}]}{\varepsilon^2(\mathbb{E}[\mathcal{D}])^2} \right\rceil$, consider sampling variables $X_1, \ldots, X_\alpha \sim \mathcal{D}$, and let $Z = \sum_{i=1}^{\alpha} X_i/\alpha$. Then Z is a "good" estimator for ρ. Formally, we have*

$$\mathbb{P}\Big[(1-\varepsilon)\rho \leq Z \leq (1+\varepsilon)\rho\Big] \geq \frac{3}{4}.$$

26.1.3. Median estimator: Success with high probability

We would like to get a better estimator, where the probability of success is high probability. Formally, we would have parameter φ, and we would like the estimator to succeed with probability $\geq 1-\varphi$. A natural approach is to try and use Chernoff to bound the probability of failure for the averaging estimator. This would work in some cases, but is limited to the case when Z lies in a small bounded range. This would not work in general if sampling from $\mathcal{D}$ might return a huge value with tiny probability. Instead, we are going to boost the averaging estimator. Assume, we generating

$$\beta = O\Big(\log \frac{1}{\varphi}\Big)$$

instances of the averaging estimators: $Z_1, \ldots, Z_\beta$ of Lemma 26.1.1. The ***median estimator*** returns the median value of the Zs as the desired estimate.

Analysis. Let $\mathcal{E}_i$ be the event that $Z_i \in [(1-\varepsilon)\rho, (1+\varepsilon)\rho]$. Let G_i be an indicator variable for $\mathcal{E}_i$. By Lemma 26.1.1, $\mathbb{P}[\mathcal{E}_i] = \mathbb{P}[G_i = 1] \geq 3/4$. The median estimator fails if $\sum_{i=1}^{\beta} G_i < \beta/2$. Using Chernoff inequality, we get that this happens with probability $\leq \varphi$. We thus get the following.

Theorem 26.1.2. *Let $\mathcal{D}$ be a non-negative distribution with $\mu = \mathbb{E}[\mathcal{D}]$ and $\nu = \mathbb{V}[\mathcal{D}]$, and let $\varepsilon, \varphi \in (0,1)$ be parameters. For some absolute constant $c > 0$, let $M \geq 24\left\lceil \frac{4\nu}{\varepsilon^2\mu^2} \right\rceil \ln \frac{1}{\varphi}$, and consider sampling variables $X_1, \ldots, X_M \sim \mathcal{D}$. One can compute, in, $O(M)$ time, a quantity Z from the sampled variables, such that*

$$\mathbb{P}\Big[(1-\varepsilon)\mu \leq Z \leq (1+\varepsilon)\mu\Big] \geq 1-\varphi.$$

Proof: Let $m = \lceil 4\nu/(\varepsilon^2\mu^2) \rceil$ and $M = \left\lceil 24 \ln \frac{1}{\varphi} \right\rceil$. Build M averaging estimators, each one using m samples. That is let Z_i be the average of m samples $s_{i,1}, \ldots, s_{i,m}$ from $\mathcal{D}$, for $i = 1, \ldots, M$. Formally,

$$Z_i = \frac{1}{m} \sum_{j=1}^{m} s_{i,j} \qquad \text{for} \qquad i = 1, \ldots, M.$$

The estimate returned is the value median$(Z_1, \ldots, Z_M)$.

By Lemma 26.1.1 each one of the averaging estimator is in the "good" range with probability $\geq 3/4$. As such, let X_i, for $i = 1, \ldots, M$, be an indicator variable, that is 1 if the ith averaging estimator is in the range $[(1-\varepsilon)\mu, (1+\varepsilon)\mu]$. Let $Y = \sum_{i=1}^{M} X_i$. We have that $\mathbb{E}[Y] \geq (3/4)M$. As such, by Lemma **??**, we have

$$\mathbb{P}[\text{bad output}] = \mathbb{P}[Y < (1/2)M] \leq \mathbb{P}[Y < (1 - 1/3)\,\mathbb{E}[Y]] \leq \exp\left(-\frac{(1/3)^2}{2}\,\mathbb{E}[Y]\right).$$

The later quantity is bounded by $\exp\left(-\frac{1}{18}\frac{3}{4}M\right) = \exp(-M/24) = \exp\left(-\left\lceil 24 \ln \varphi^{-1} \right\rceil /24\right) \leq \varphi$. ■

26.2. Frequency estimation over a stream for the kth moment

Let $\mathcal{S} = (s_1, \ldots, s_m)$ be a stream (i.e., sequence) of m elements from $N = \{1, \ldots, n\}$. Let f_i be the number of times the number i appears in $\mathcal{S}$. For $k \geq 0$, let

$$F_k = \sum_{i=1}^{n} f_i^k$$

be the ***kth frequency moment*** of $\mathcal{S}$. The quantity, $F_1 = m$ is the length of the stream $\mathcal{S}$. Similarly, F_0 is the number of distinct elements (where we use the convention that $0^0 = 0$ and any other quantity to the power 0 is 1). It is natural to define $F_\infty = \max_i f_i$.

Here, we are interested in approximating up to a factor of $1 \pm \varepsilon$ the quantity F_k, for $k \geq 1$ using small space, and reading the stream $\mathcal{S}$ only once.

26.2.1. An estimator for the kth moment

26.2.1.1. Basic estimator

One can pick a representative element from a stream uniformly at random by using reservoir sampling. That is, sample the ith element s_i to be the representative with probability $1/i$. Once sampled, the algorithm counts how many times it see the representative value later on in the stream (the counter is initialized to 1, to account for the chosen representative itself). In particular, if s_p is the chosen representative in the end of the stream (i.e., the algorithm might change the representative several times), then the counter value is

$$r = \left| \left\{ j \;\middle|\; j \geq p \text{ and } s_j = s_p \right\} \right|.$$

The output of the algorithm is the quantity

$$X = m\,(r^k - (r-1)^k),$$

where m is the number of elements seen in the stream. Let V be the random variable that is the value of the representative in the end of the sequence.

26.2.1.2. Analysis

Lemma 26.2.1. *We have* $\mathbb{E}[X] = F_k$.

Proof: Observe that since we choose the representative uniformly at random, we have

$$\mathbb{E}[X \mid V = i] = \sum_{j=1}^{f_i} \frac{1}{f_i} m(j^k - (j-1)^k) = \frac{m}{f_i} \sum_{j=1}^{f_i} (j^k - (j-1)^k) = \frac{m}{f_i} f_i^k.$$

As such, we have $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid V]] = \sum_{i:f_i \neq 0} \frac{f_i}{m} \frac{m}{f_i} f_i^k = \sum_i f_i^k = F_k$. ■

Remark 26.2.2. In the above, we estimated the function $g(x) = x^k$, over the frequency numbers $f_1, \ldots, f_k$, but the above argumentation, on the expectation of X, would work for any function $g(x)$ such that $g(0) = 0$, and $g(x) \geq 0$, for all $x \geq 0$.

Lemma 26.2.3. *For $k > 1$, we have* $\sum_{i=1}^{n} \left(i^k - (i-1)^k\right)^2 \leq kn^{2k-1}$.

Proof: Observe that for $x \geq 1$, we have that $x^k - (x-1)^k \leq kx^{k-1}$. As such, we have

$$\sum_{i=1}^{n} \left(i^k - (i-1)^k\right)^2 \leq \sum_{i=1}^{n} ki^{k-1}\left(i^k - (i-1)^k\right) \leq kn^{k-1} \sum_{i=1}^{n} \left(i^k - (i-1)^k\right) = kn^{k-1} n^k = kn^{2k-1}.$$ ■

Lemma 26.2.4. *We have* $\mathbb{E}\left[X^2\right] \leq kmF_{2k-1}$.

Proof: By Lemma 26.2.3, we have

$$\mathbb{E}\left[X^2 \;\middle|\; V = i\right] = \sum_{j=1}^{f_i} \frac{1}{f_i} m^2 (j^k - (j-1)^k)^2 \leq \frac{m^2}{f_i} k f_i^{2k-1} = m^2 k f_i^{2k-2},$$

and thus $\mathbb{E}[X^2] = \mathbb{E}[\mathbb{E}[X^2 \mid V]] = \sum_{i:f_i \neq 0} \frac{f_i}{m} \cdot m^2 k f_i^{2k-2} = mkF_{2k-1}$. ■

Lemma 26.2.5. *For any non-negative numbers $f_1, \ldots, f_n$, and $k \geq 1$, we have*

$$\sum_{i=1}^{n} f_i \leq n^{(k-1)/k} \left(\sum_{i=1}^{n} f_i^k\right)^{1/k}.$$

Proof: This is immediate from Hölder inequality, but here is a self contained proof. The above is equivalent to proving that $\sum_i f_i/n \leq \left(\sum_{i=1}^{n} f_i^k/n\right)^{1/k}$. Raising both sides to the power k, we need to show that $(\sum_i f_i/n)^k \leq \sum_{i=1}^{n} f_i^k/n$. Setting $g(x) = x^k$, we have $g(\sum_i f_i/n) \leq \sum_{i=1}^{n} g(f_i)/n$. The last inequality holds by the convexity of the function $g(x)$ (indeed, $g'(x) = kx^{k-1}$ and $g''(x) = k(k-1)x^{k-2} \geq 0$, for $x \geq 0$). ■

Lemma 26.2.6. *For any n numbers $f_1, \ldots, f_n \geq 0$, we have* $\left(\sum_i f_i\right)\left(\sum_i f_i^{2k-1}\right) \leq n^{1-1/k}\left(\sum_i f_i^k\right)^2$.

Proof: Let $M = \max_i f_i$ and $m = \sum_i f_i$. We have

$$\sum_i f_i^{2k-1} \le M^{k-1} \sum_i f_i^k \le M^{k(k-1)/k} \sum_i f_i^k \le \Big(\sum_i f_i^k\Big)^{(k-1)/k} \sum_i f_i^k \le \Big(\sum_i f_i^k\Big)^{(2k-1)/k}.$$

By Lemma 26.2.5, we have $\sum_{i=1}^n f_i \le n^{(k-1)/k}\Big(\sum_i f_i^k\Big)^{1/k}$. Multiplying the above two inequality implies the claim. ■

Lemma 26.2.7. *We have* $\mathbb{V}[X] \le kn^{1-1/k}F_k^2$.

Proof: Since $m = \sum_i f_i$, Lemma 26.2.4 and Lemma 26.2.6 together implies that

$$\mathbb{V}[X] = \mathbb{E}\Big[X^2\Big] - (\mathbb{E}[X])^2 \le \mathbb{E}\Big[X^2\Big] \overbrace{\le}^{\text{L26.2.4}} kmF_{2k-2} = k\Big(\sum_i f_i\Big)\Big(\sum_i f_i^{2k-1}\Big) \overbrace{\le}^{\text{L26.2.6}} kn^{1-1/k}F_k^2.$$ ■

26.2.2. An improved estimator: Plugin

We have an estimator for F_k using constant space $O(1)$. Specifically, $\mu = \mathbb{E}[X] = F_k$ see Lemma 26.2.1, and $\nu = \mathbb{V}[X] \le kn^{1-1/k}F_k^2$. Let

$$M = 24\Big\lceil \frac{4\nu}{\varepsilon^2\mu^2} \Big\rceil \ln \frac{1}{\varphi}$$

We compute M estimators as the above (in parallel on the stream), and combine them as specified by Theorem 26.1.2, to get a new estimate Z. We have that

$$\mathbb{P}\Big[(1-\varepsilon)\mu \le Z \le (1+\varepsilon)\mu\Big] \ge 1 - \varphi.$$

Thus, the amount of space this streaming algorithm is using is proportional to M, and we have

$$M = O\left(\frac{\nu}{\varepsilon^2\mu^2} \ln \frac{1}{\varphi}\right) = O\left(\frac{kn^{1-1/k}F_k^2}{\varepsilon^2 F_k^2} \ln \frac{1}{\varphi}\right) = O\left(\frac{kn^{1-1/k}}{\varepsilon^2} \ln \frac{1}{\varphi}\right).$$

We thus proved the following.

In the following, we consider a computer ***word*** to be sufficiently large to contain $\lg n$ or $\lg m$ bits. This readily implies the following.

Theorem 26.2.8. *Let $S = (s_1, \ldots, s_n)$ be a stream of numbers from the set $\{1, \ldots, n\}$. Let $k \ge 1$ be a parameter. Given $\varepsilon, \varphi \in (0,1)$, one can build a data-structure using $O(kn^{1-1/k}\varepsilon^{-2}\log\varphi^{-1})$ words, such that one can $(1\pm\varepsilon)$-approximate the kth moment of the elements in the stream; that is, the algorithm outs a number Z, such that $(1-\varepsilon)F_k \le Z \le (1+\varepsilon)F_k$, where $F_k = \sum_{i=1}^n f_i^k$, and f_i is the number of times i appears in the stream S. The algorithm succeeds with probability $\ge 1-\varphi$.*

26.3. Better estimation for F_2

26.3.1. Pseudo-random k-wide independent sequence of signed bits

In the following, assume that we sample $O(\log n)$ bits, such that given an index i, one can compute (quickly!) a random signed bit $b(i) \in \{-1, +1\}$. We require that the resulting bits $b(1), b(2), \ldots, b(n)$ are 4-wise independent.

To this end, pick a prime p, that is, say bigger than n^{10}. This can easily be done by sampling a number in the range $[n^{10}, n^{11}]$, and checking if it is prime (which can done in polynomial time).

Once we have such a prime, we generate a random polynomial $g(i) = \sum_{i=0}^{5} c_i x^i \bmod p$, by choosing $c_0, \ldots, c_5$ from $\mathbb{Z}_p = \{0, \ldots, p-1\}$. We had seen that $g(0), g(1), \ldots, g(n)$ are uniformly distributed in $\mathbb{Z}_p$, and they are, say, 6-wise independent (see Theorem 26.5.4).

We define

$$b(i) = \begin{cases} 0 & g(i) = p-1 \\ +1 & g(i) \text{ is odd} \\ -1 & g(i) \text{ is even.} \end{cases}$$

Clearly, the sequence $b(1), \ldots, b(n)$ are 6-wise independent. There is a chance that one of these bits might be zero, but the probability for that is at most n/p, which is so small, that we just assume it does not happen. There are known constructions that do not have this issue at all (one of the bits is zero), but they are more complicated.

Lemma 26.3.1. *Given a parameter $\varphi \in (0,1)$, in polynomial time in $O(\log(n/\varphi))$, one can construct a function $b(\cdot)$, requiring $O(\log(n/\varphi))$ bits of storage (or $O(1)$ words), such that $b(1), \ldots, b(n) \in \{-1, +1\}$ with equal probability, an they are 6-wise independent. Furthermore, given i, one can compute $b(i)$ in $O(1)$ time.*

The probability of this sequence to fail having the desired properties is smaller than φ.

Proof: We repeat the above construction, but picking a prime p in the range, say, $n^{10}/\varphi \ldots n^{11}/\varphi$. ■

26.3.2. Estimator construction for F_2

26.3.2.1. The basic estimator

As before we have the stream $\mathcal{S} = s_1 \ldots, s_m$ of numbers from the set $1, \ldots, n$. We compute the 6-wise independent sequence of random bits of Lemma 26.3.1, and in the following we assume this sequence is good (i.e., has only -1 and $+1$ in it). We compute the quantity

$$T = \sum_{i=1}^{m} b(i) f_i = \sum_{j=1}^{m} b(s_j),$$

which can be computed on the fly using $O(1)$ words of memory, and $O(1)$ time per time in the stream.

The algorithm returns $X = T^2$ as the desired estimate.

Analysis.

Lemma 26.3.2. *We have $\mathbb{E}[X] = \sum_i f_i^2 = F_2$ and $\mathbb{V}[X] \le 2F_2^2$.*

Proof: We have that $\mathbb{E}[X] = \mathbb{E}\Big[\Big(\sum_{i=1}^{n} b(i) f_i\Big)^2\Big]$, and as such

$$\mathbb{E}[X] = \mathbb{E}\Big[\sum_{i=1}^{n} (b(i))^2 f_i^2 + 2\sum_{i<j} b(i)b(j) f_i f_j\Big] = \sum_{i=1}^{m} f_i^2 + 2\sum_{i<j} f_i f_j \,\mathbb{E}[b(i)b(j)] = \sum_{i=1}^{m} f_i^2 = F_2,$$

since $\mathbb{E}[b(i)] = 0$, $\mathbb{E}\Big[b(i)^2\Big] = 1$, and $\mathbb{E}[b(i)b(j)] = \mathbb{E}[b(i)]\,\mathbb{E}[b(j)] = 0$ (assuming the sequence $b(1), \ldots, b(n)$ has not failed), by the 6-wise Independence of the sequence of signed bits.

We next compute $\mathbb{E}\Big[X^2\Big]$. To this end, let $N = \{1, \ldots, n\}$, and $\Gamma = N \times N \times N \times N$. We split this set into several sets, as follows:

(i) $\Gamma_0 = \left\{(i,i,i,i) \in N^4\right\}$: All quadruples that are all the same value.

(ii) Γ_1: Set of all quadruples (i,j,k,l) where there is at least one value that appears exactly once.

(iii) Γ_2: Set of all (i,j,k,ℓ) with only two distinct values, each appearing exactly twice.

Clearly, we have $N^4 = \Gamma_0 \cup \Gamma_1 \cup \Gamma_2$.

For a tuple $(i,i,i,i) \in \Gamma_0$, we have $\mathbb{E}[b(i)b(i)b(i)b(i)] = \mathbb{E}\left[b(i)^4\right] = 1$.

For a tuple $(i,j,k,\ell) \in \Gamma_1$ with i being the unique value, we have that

$$\mathbb{E}[b(i)b(j)b(k)b(\ell)] = \mathbb{E}[b(i)]\,\mathbb{E}[b(j)b(k)b(\ell)] = 0\,\mathbb{E}[b(j)b(k)b(\ell)] = 0,$$

using that the signed bits are 4-wise independent.

For a tuple $(i,i,j,j) \in \Gamma_2$, we have $\mathbb{E}[b(i)b(i)b(j)b(j)] = \mathbb{E}\left[b(i)^2 b(j)^2\right] = \mathbb{E}\left[b(i)^2\right]\mathbb{E}\left[b(j)^2\right] = 1$, and the same argumentation applies to any tuple of Γ_2. Observe that for any $i < j$, there are $\binom{4}{2} = 6$ different tuples in Γ_2 that are made out of i and j. As such, we have

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \mathbb{E}\left[\Big(\sum_{i=1}^{n} b(i) f_i\Big)^4\right] = \mathbb{E}\Big[\sum_{(i,j,k,\ell)\in\Gamma} b(i)b(j)b(k)b(\ell) f_i f_j f_k f_\ell\Big] \\
&= \sum_{(i,i,i,i)\in\Gamma_0} \mathbb{E}\left[b(i)^4\right] f_i^4 + \sum_{(i,j,k,\ell)\in\Gamma_1} f_i f_j f_k f_\ell\, \mathbb{E}[b(i)b(j)b(k)b(\ell)] + 6\sum_{i<j} \mathbb{E}\left[b(i)^2 b(j)^2\right] f_i^2 f_j^2 \\
&= \sum_{i=1}^{n} f_i^4 + 6\sum_{i<j} f_i^2 f_j^2.
\end{aligned}$$

As such, we have

$$\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \sum_{i=1}^{n} f_i^4 + 6\sum_{i<j} f_i^2 f_j^2 - \Big(\sum_{i=1}^{m} f_i^2\Big)^2 = 4\sum_{i<j} f_i^2 f_j^2 \le 2F_2^2. \qquad \blacksquare$$

26.3.3. Improving the estimator

We repeat the same scheme as above. Let $\varphi, \varepsilon \in (0,1)$ be parameters. In the following, let

$$\alpha = 16/\varepsilon^2 \qquad \text{and} \qquad \beta = 4\ln\frac{1}{\varphi}.$$

Let $X_{i,j}$ be a basic estimator for F_2, using the estimator of Section 26.3.2.1, for $i = 1,\ldots,\beta$ and $j = 1,\ldots,\alpha$. Let $Y_i = \sum_{j=1}^{\alpha} X_{i,j}/\alpha$, for $i = 1,\ldots,\beta$. Let Z be the median of $Y_1,\ldots,Y_\beta$, and the algorithm returns Z as the estimator.

Theorem 26.3.3. *Given a stream $\mathcal{S} = s_1,\ldots,s_m$ of numbers from $\{1,\ldots,n\}$, and parameters $\varepsilon,\varphi \in (0,1)$, one can compute an estimate Z for $F_2(\mathcal{S})$, such that $\mathbb{P}[|Z - F_2| > \varepsilon F_2] \le \varphi$. This algorithm requires $O(\varepsilon^{-2}\log\varphi^{-1})$ space (in words), and this is also the time to handle a new element in the stream.*

Proof: The scheme is described above. As before, using Chebychev's inequality, we have that

$$\mathbb{P}[|Y_i - F_2| > \varepsilon F_2] = \mathbb{P}\left[|Y_i - F_2| > \frac{\varepsilon F_2}{\sqrt{\mathbb{V}[Y_i]}}\sqrt{\mathbb{V}[Y_i]}\right] \le \frac{\mathbb{V}[Y_i]}{\varepsilon^2 F_2^2} = \frac{\mathbb{V}[X]/\alpha}{\varepsilon^2 F_2^2} \le \frac{2F_2^2}{\alpha\varepsilon^2 F_2^2} = \frac{1}{8},$$

by Lemma 26.3.2. Let U be the number of estimators in $Y_1,\ldots,Y_\beta$ that are outside the acceptable range. Arguing as in Lemma **??**, we have

$$\mathbb{P}[Z \text{ is bad}] \le \mathbb{P}[U \ge \beta/2] = \mathbb{P}[U \ge (1+3)\beta/8] \le \exp(-(\beta/8)3^2/4) \le \exp\left(-\ln\frac{1}{\varphi}\right) = \varphi,$$

by Chernoff inequality (Lemma 26.5.2), and $\blacksquare$

26.4. Bibliographical notes

The beautiful results of this chapter are from a paper from Alon *et al.* [AMS99].

26.5. From previous lectures

Theorem 26.5.1 (Chebyshev's inequality). *Let X be a real random variable, with $\mu_X = \mathbb{E}[X]$, and $\sigma_X = \sqrt{\mathbb{V}[X]}$. Then, for any $t > 0$, we have $\mathbb{P}[|X - \mu_X| \geq t\sigma_X] \leq 1/t^2$.*

Lemma 26.5.2. *Let $X_1, \ldots, X_n$ be n independent Bernoulli trials, where $\mathbb{P}[X_i = 1] = p_i$, and $\mathbb{P}[X_i = 0] = 1 - p_i$, for $i = 1, \ldots, n$. Let $X = \sum_{i=1}^{b} X_i$, and $\mu = \mathbb{E}[X] = \sum_i p_i$. For $\delta \in (0, 4)$, we have*

$$\mathbb{P}[X > (1 + \delta)\mu] < \exp\Big(-\mu\delta^2/4\Big),$$

Lemma 26.5.3. *Let $X_1, \ldots, X_n \in \{0, 1\}$ be n independent random variables, with $p_i = \mathbb{P}[X_i = 1]$, for all i. For $X = \sum_{i=1}^{n} X_i$, and $\mu = \mathbb{E}[X] = \sum_i p_i$, we have that $\mathbb{P}[X < (1 - \delta)\mu] < \exp\Big(-\mu\delta^2/2\Big)$.*

Theorem 26.5.4. *let p be a prime number, and pick independently and uniformly k values $b_0.b_1, \ldots, b_{k-1} \in \mathbb{Z}_p$, and let $g(x) = \sum_{i=0}^{k-1} b_i x^i \bmod p$. Then the random variables*

$$Y_0 = g(0), \ldots, Y_{p-1} = g(p - 1).$$

are uniformly distributed in $\mathbb{Z}_p$ and are k-wise independent.

References

[AMS99] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1): 137–147, 1999.

[MR95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 1995.