Chapter 4

Chebychev, Sampling and Selection

By Sariel Har-Peled, April 26, 2022[1]

> During a native rebellion in German East Africa, the Imperial Ministry in Berlin issued the following order to its representatives on the ground: "The natives are to be instructed that on pain of harsh penalties, every rebellion must be announced, in writing, six weeks before it breaks out."
>
> Dead Funny: Humor in Hitler's Germany, Rudolph Herzog

4.1. Chebyshev's inequality

4.1.1. Example: A better inequality via moments

Let $X_i \in \{-1, +1\}$ with probability half for each value, for $i = 1, \ldots, n$ (all picked independently). Let $Y = \sum_i X_i$. We have that

$$\mathbb{E}\big[Y\big] = \mathbb{E}\Big[\sum_i X_i\Big] = \sum_i \mathbb{E}\big[X_i\big] = n \cdot 0 = 0.$$

A more interesting quantity is

$$\mathbb{E}\big[Y^2\big] = \mathbb{E}\Big[\Big(\sum_i X_i\Big)^2\Big] = \mathbb{E}\Big[\sum_i X_i^2 + 2\sum_{i<j} X_i X_j\Big] = \sum_i \mathbb{E}\big[X_i^2\big] + 2\,\mathbb{E}\Big[\sum_{i<j} X_i X_j\Big] = n + 2\sum_{i<j} \mathbb{E}\big[X_i X_j\big]$$
$$= n + 2\sum_{i<j} \mathbb{E}\big[X_i\big]\,\mathbb{E}\big[X_j\big] = n.$$

Lemma 4.1.1. *Let $X_i \in \{-1, +1\}$ with probability half for each value, for $i = 1, \ldots, n$ (all picked independently). We have that $\mathbb{P}\big[\left|\sum_i X_i\right| > t\sqrt{n}\big] < 1/t^2$.*

Proof: Let $Y = \sum_i X_i$ and $Z = Y^2$. We have

$$\mathbb{P}\Big[\Big|\sum_i X_i\Big| > t\sqrt{n}\Big] = \mathbb{P}\Big[\Big(\sum_i X_i\Big)^2 > t^2 n\Big] = \mathbb{P}\Big[Y^2 > t^2\,\mathbb{E}\big[Y^2\big]\Big] = \mathbb{P}\big[Z > t^2\,\mathbb{E}[Z]\big] \leq 1/t^2,$$

by Markov's inequality. ∎

4.1.2. Chebychev's inequality

As a reminder, the variance of a random variable X is $\mathbb{V}[X] = \mathbb{E}\big[(X - \mu_X)^2\big] = \mathbb{E}\big[X^2\big] - \mu_X^2$.

Theorem 4.1.2 (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 $\mathbb{P}\big[|X - \mu_X| \geq t\sigma_X\big] \leq 1/t^2$.*

Proof: Set $Y = (X - \mu_X)^2$, and observe that

$$\sigma_X^2 = \mathbb{V}[X] = \mathbb{E}[Y] = \mathbb{E}\big[(X - \mu_X)^2\big] = \mathbb{E}\big[X^2\big] - \mu_X^2.$$

As such, we have that

$$\mathbb{P}\big[|X - \mu_X| \geq t\sigma_X\big] = \mathbb{P}\big[\,(X - \mu_X)^2 \geq t^2\sigma_X^2\big] = \mathbb{P}\big[Y \geq t^2\,\mathbb{E}[Y]\big] \leq \frac{1}{t^2},$$

by Markov's inequality. ■

4.2. Estimation via sampling

One of the big advantages of randomized algorithms, is that they sample the world; that is, learn how the input looks like without reading all the input. For example, consider the following problem: We are given a set of U of n objects $u_1, \ldots, u_n$. and we want to compute the number of elements of U that have some property. Assume, that one can check if this property holds, in constant time, for a single object, and let $\psi(u)$ be the function that returns 1 if the property holds for the element u. and zero otherwise. Now, let Γ be the number of objects in U that have this property. We want to reliably estimate Γ without computing the property for all the elements of U.

A natural approach, would be to pick a random sample R of m objects, $r_1, \ldots, r_m$ from U (with replacement), and compute $Y = \sum_{i=1}^{m} \psi(r_1)$. The estimate for Γ is $\beta = (n/m)Y$. It is natural to ask how far is β from the true value Γ.

Lemma 4.2.1. *Let U be a set of n elements, with Γ of them having a certain property ψ. Let R be a uniform random sample from U (with repetition) of size m, and let Y be the number of elements in R that have the property ψ, and let $Z = (n/m)Y$ be the estimate for Γ. Then, for any $t \geq 1$, we have that*

$$\mathbb{P}\Big[\Gamma - t\frac{n}{2\sqrt{m}} \leq Z \leq \Gamma + t\frac{n}{2\sqrt{m}}\Big] \geq 1 - \frac{1}{t^2}.$$

Similarly, we have that $\mathbb{P}\big[\mathbb{E}[Y] - t\sqrt{m}/2 \leq Y \leq \mathbb{E}[Y] + t\sqrt{m}/2\big] \geq 1 - 1/t^2$.

Proof: Let $Y_i = \psi(r_i)$ be an indicator variable that is 1 if the ith sample r_i has the property ψ, for $i = 1, \ldots, m$. Consider the random variable $Y = \sum_i Y_i$, and the estimate $Z = (n/m)Y$ for Γ. Observe that

$$\mathbb{E}[Z] = \mathbb{E}[(n/m)Y] = \frac{n}{m}\,\mathbb{E}[Y] = \frac{n}{m}\,\mathbb{E}\Big[\sum_i Y_i\Big] = \frac{n}{m}\sum_{i=1}^{m}\mathbb{E}[Y_i] = \frac{n}{m}\sum_{i=1}^{m}\frac{\Gamma}{n} = \frac{n}{m}\cdot m \cdot \frac{\Gamma}{n} = \Gamma.$$

The variable Y is a binomial distribution with probability $p = \Gamma/n$, and m samples; that is, $Y \sim \mathrm{Bin}(m, p)$. We saw in the previous lecture that, $\mathbb{E}[Y] = mp$, $\mathbb{V}[Y] = mp(1 - p)$, and its standard deviation is

$$\sigma_Y = \sqrt{mp(1 - p)} \leq \sqrt{m}/2,$$

as $\sqrt{p(1 - p)}$ is maximized for $p = 1/2$.

By Chebychev's inequality, we have that $\mathbb{P}\big[|Y - \mathbb{E}[Y]| \geq t\sigma_Y\big] \leq 1/t^2$. Since $(n/m)\,\mathbb{E}[Y] = \mathbb{E}[Z] = \Gamma$, this implies that

$$\begin{aligned}\frac{1}{t^2} &\geq \mathbb{P}\big[|Y - \mathbb{E}[Y]| \geq t\sigma_Y\big] \geq \mathbb{P}\Big[\Big|\frac{n}{m}Y - \frac{n}{m}\,\mathbb{E}[Y]\Big| \geq \frac{n}{m}t\sigma_Y\Big] = \mathbb{P}\Big[|Z - \Gamma| \geq \frac{n}{m}t\sigma_Y\Big]\\ &\geq \mathbb{P}\Big[|Z - \Gamma| \geq \frac{n}{m}t \cdot \frac{\sqrt{m}}{2}\Big] = \mathbb{P}\Big[|Z - \Gamma| \geq t\frac{n}{2\sqrt{m}}\Big].\end{aligned}$$

■

4.3. Randomized selection – Using sampling to learn the world

4.3.1. Inverse estimation

We are given a set $U = \{u_1, \ldots, u_n\}$ of n distinct numbers. Let $U_{\langle i \rangle}$ denote the ith smallest number in U – that is $U_{\langle i \rangle}$ is the number of ***rank*** i in U.

Lemma 4.3.1. *Given a set U of n numbers, a number k, and parameters $t \geq 1$ and $m \geq 1$, one can compute, in $O(m \log m)$ time, two numbers $r_-, r_+ \in U$, such that:*

(A) The number of rank k in U is in the interval $\mathcal{I} = [r_-, r_+]$.

(B) There are at most $8tn/\sqrt{m}$ numbers of U in $\mathcal{I}$.

The above two properties hold with probability $\geq 1 - 3/t^2$.

(Namely, as t increases, the interval $\mathcal{I}$ becomes bigger, and the probability it contains the desired element increases.)

Proof: **(A)** Compute a random sample R of U of size m in $O(m)$ time (assuming the input numbers are given in an array, say). Next sort the numbers of R in $O(m \log m)$ time. Let

$$\ell_- = \left\lfloor m\frac{k}{n} - t\sqrt{m}/2 \right\rfloor - 1 \qquad \text{and} \qquad \ell_+ = \left\lceil m\frac{k}{n} + t\sqrt{m}/2 \right\rceil + 1.$$

Set $r_- = \mathsf{R}[\ell_-]$ and $r_+ = \mathsf{R}[\ell_+]$.

Let Y be the number of elements in the sample R that are $\leq U_{\langle k \rangle}$. By Lemma 4.2.1, we have $\mathbb{P}\left[\mathbb{E}[Y] - t\sqrt{m}/2 \leq Y \leq \mathbb{E}[Y] + t\sqrt{m}/2\right] \geq 1 - 1/t^2$. In particular, if this happens, then $r_- \leq U_{\langle k \rangle} \leq r_+$.

(B) Let $g = k - t\frac{n}{\sqrt{m}} - 3\frac{n}{m}$, and let g_{R} be the number of elements in R that are smaller than $U_{\langle g \rangle}$. Arguing as above, we have that $\mathbb{P}\left[g_{\mathsf{R}} \leq \frac{g}{n}m + t\sqrt{m}/2\right] \geq 1 - 1/t^2$. Now

$$\frac{g}{n}m + t\sqrt{m}/2 = \frac{m}{n}\left(k - t\frac{n}{\sqrt{m}} - 3\frac{n}{m}\right) + t\sqrt{m}/2 = k\frac{m}{n} - t\sqrt{m} - 3 + t\sqrt{m}/2 = k\frac{m}{n} - t\sqrt{m}/2 - 3 < \ell_-.$$

This implies that the g smallest numbers in U are outside the interval $[r_-, r_+]$ with probability $\geq 1 - 1/t^2$.

Next, let $h = k + t\frac{n}{\sqrt{m}} + 3\frac{n}{m}$. A similar argument, shows that all the $n - h$ largest numbers in U are too large to be in $[r_-, r_+]$. This implies that

$$|[r_-, r_+] \cap U| \leq h - g + 1 = 6\frac{n}{m} + 2t\frac{n}{\sqrt{m}} \leq 8\frac{tn}{\sqrt{m}}. \qquad \blacksquare$$

4.3.1.1. Inverse estimation – intuition

Here we are trying to give some intuition to the proof of the previous lemma. Feel free to skip this part if you feel you already understand what is going on.

Given k, we are interested in estimating $s_k = U_{\langle k \rangle}$ quickly. So, let us take a sample R of size m. Let $\mathsf{R}_{\leq s_k}$ be the set of all the numbers in R that are $\leq s_k$. For $Y = \left|\mathsf{R}_{\leq s_k}\right|$, we have that $\mu = \mathbb{E}[Y] = m\frac{k}{n}$. Furthermore, for any $t \geq 1$, Lemma 4.2.1 implies that $\mathbb{P}\left[\mu - t\sqrt{m}/2 \leq Y \leq \mu + t\sqrt{m}/2\right] \geq 1 - 1/t^2$. In particular, with probability $\geq 1 - 1/t^2$ the number $r_- = \mathsf{R}_{\langle \ell_- \rangle}$, for $\ell_- = \left\lfloor \mu - t\sqrt{m}/2 \right\rfloor - 1$, is smaller than s_k, and similarly, the number $r_+ = \mathsf{R}_{\langle \ell_+ \rangle}$ of rank $\ell_+ = \left\lceil \mu + t\sqrt{m}/2 \right\rceil + 1$ in R is larger than s_k.

One can conceptually think about the interval $\mathcal{I}(k) = [r_-, r_+]$ as a *confidence interval* – we know that $s_k \in \mathcal{I}(k)$ with probability $\ge 1 - 1/t^2$. But how heavy is this interval? Namely, how many elements are there in $\mathcal{I}(k) \cap U$?

To this end, consider the interval of ranks, *in the sample*, that might contain the kth element. By the above, this is $\mathcal{I}(k,t) = k\frac{m}{n} + \left[-t\sqrt{m}/2 - 1, t\sqrt{m}/2 + 1\right]$. In particular, consider the maximum $\nu \le k$, such that $\mathcal{I}(\nu, t)$ and $\mathcal{I}(k,t)$ are disjoint. We have the condition that $\nu\frac{m}{n} + t\sqrt{m}/2 + 1 \le k\frac{m}{n} - t\sqrt{m}/2 - 1 \implies \nu \le k - t\frac{n}{\sqrt{m}} - 2\frac{n}{m}$. Let $g = k - t\frac{n}{\sqrt{m}} - 2\frac{n}{m}$ and $h = k + t\frac{n}{\sqrt{m}} + 2\frac{n}{m}$. We have that $\mathcal{I}(g,t)$, $\mathcal{I}(k,t)$ and $\mathcal{I}(h,t)$ are all disjoint with probability $\ge 1 - 3/t^2$.

To this end, let $g = k - \left\lceil 2\left(t\frac{n}{2\sqrt{m}}\right)\right\rceil$ and $h = k + \left\lceil 2\left(t\frac{n}{2\sqrt{m}}\right)\right\rceil$. It is easy to verify (using the same argumentation as above) that with probability at least $1 - 3/t^2$, the three confidence $\mathcal{I}(g), \mathcal{I}(k)$ and $\mathcal{I}(h)$ do not intersect. As such, we have $\left|\mathcal{I}(k) \cap U\right| \le h - g \le 4\left(t\frac{n}{2\sqrt{m}}\right)$.

4.3.2. Randomized selection

4.3.2.1. The algorithm

Given an array S of n numbers, and the rank k. The algorithm needs to compute $S_{\langle k \rangle}$. To this end, set $t = \lceil n^{1/8} \rceil$, and $m = \lceil n^{3/4} \rceil$.

Using the algorithm of Lemma 4.3.1, in $O(m \log m)$ time, we get two numbers r_- and r_+, such that $S_{\langle k \rangle} \in [r_i, r_+]$, and

$$\underbrace{|S \cap (r_i, r_+)|}_{S_m} = O\left(tn/\sqrt{m}\right) = O\left(n^{1/8} n / m^{3/8}\right) = O(n^{3/4}).$$

To this end, we break S into three sets:

(i) $S_< = \{s \in S \mid s \le r_-\}$,

(ii) $S_m = \{s \in S \mid r_- < s < r_+\}$,

(iii) $S_> = \{s \in S \mid r_+ \le s\}$.

This three way partition can be done using $2n$ comparisons and in linear time. We now can readily compute the rank of r_- in S (it is $|S_<|$) and the rank of r_+ in S (it is $|S_<| + |S_m| + 1$). If $r_{-\langle S \rangle} > k$ or $r_{+\langle S \rangle} < k$ then the algorithm failed. The other possibility for failure is that S_m is too large. That is, larger than $8tn/\sqrt{m} = O(n^{3/4})$. If any of these failures happen, then we rerun this algorithm from scratch.

Otherwise, the algorithm need to compute the element of rank $k - |S_<|$ in the set S_m, and this can be done in $O(|S_m| \log |S_m|) = O(n^{3/4} \log n)$ time by using sorting.

4.3.2.2. Analysis

The correctness is easy – the algorithm clearly returns the desired element. As for running time, observe that by Lemma 4.3.1, by probability $\ge 1 - 1/n^{1/4}$, we succeeded in the first try, and then the running time is $O(n + (m \log m)) = O(n)$. More generally, the probability that the algorithm failed in the first α tries to get a good interval $[r_-, r_+]$ is at most $1/n^{\alpha/4}$.

One can slightly improve the number of comparisons performed by the algorithm using the following modifications.

Lemma 4.3.2. *Given the numbers r_-, r_+, one can compute the sets $S_<, S_m, S_>$ using in expectation (only!) $1.5n + O(n^{3/4})$ comparisons.*

Proof: We need to compute the sets $S_<, S_m, S_>$. Namely, we need to compare all the numbers of S to r_- and r_+. Since only $O(n^{3/4})$ numbers fall in S_m, almost all of the numbers are in either $S_<$ or $S_>$. If a number of is in $S_<$ (resp. $S_>$), then comparing it r_- (resp. r_+) is enough to verify that this is indeed the case. Otherwise, perform the other comparison and put the element in its proper set (in this case we had to perform two comparisons to handle the element).

So let us guess, by a coin flip, for each element of S whether they are in $S_<$ or $S_>$. If we are right, then the algorithm would require only one comparison to put them into the right set. Otherwise, it would need two comparisons. Let X_s be the random variable that is the number of comparisons used by this algorithm for an element $s \in S$. We have that if $s \in S_< \cup S_>$ then $\mathbb{E}[X_s] = 1(1/2) + 2(1/2) = 3/2$. If $s \in S_m$ then both comparisons will be performed, and thus $\mathbb{E}[X_s] = 2$ in this case.

Thus, the expected numbers of comparisons for all the elements of S, by linearity of expectations, is $\frac{3}{2}(n - |S_m|) + 2|S_m| = (3/2)n + |S_m|/2$. ■

Theorem 4.3.3. *Given an array S with n numbers and a rank k, one can compute the element of rank k in S in expected linear time. Formally, the resulting algorithm performs in expectation $1.5n + O(n^{3/4} \log n)$ comparisons.*

Proof: Let X be the random variable that is the number of iteration till the interval is good. We have that X is a geometric variable with probability of success $\geq 1 - 1/n^{1/4}$. As such, the expected number of rounds till success is $\leq 1/p \leq 1 + 2/n^{1/4}$. As such, the expected number of comparisons performed by the algorithm is $\mathbb{E}\Big[X \cdot \Big(1.5n + O(n^{3/4} \log n)\Big)\Big] = 1.5n + O(n^{3/4} \log n)$. ■