ETH Zürich
Institute of Theoretical Computer Science
Prof. Rasmus Kyng
Prof. Johannes Lengler
Prof. Vera Traub

FS 2025

Algorithmen und Wahrscheinlichkeit
Theoretical Exercise 4

Submit by Moodle () until 16:00 on 17.04.2025.

Exercise 1 – *AMS sketch*

We consider a sequence of elements $x_1, \dots x_m \in [N]$ arriving one by one. Let $F_i := |\{k : x_k = i\}|$ be the number of occurences of an element $i \in [N]$. We would like to approximate $\sum_{i\in[N]} F_i^2$ without storing the entire sequence x_i, or the sequence of counts F_i. Specifically, assume that we have access to independent random $\varepsilon_i \in \{\pm 1\}$ for $i \in [N]$. For convenience, let us define $F^* := \sum_i F_i^2$.

(a) Let $X = \sum_{i\in[N]} \varepsilon_i F_i$ where F_i are (deterministic) numbers, and $\varepsilon_i \in \{\pm 1\}$ are independent random variables with $\Pr[\varepsilon_i = 1] = 1/2$. Show that $\mathbb{E}X^2 = F^*$.

(b) Show that $\mathbb{E}X^4 \le 3(F^*)^2$. Conclude that $\mathrm{Var}(X^2) \le 3\mathbb{E}[X^2]^2$. (Hint $\mathcal{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell]$ is non-zero, only in three possible scenarios: when $i = j$ and $k = \ell$, or when $i = k$ and $j = \ell$, or when $i = \ell$ and $j = k$.)

(c) Let $X_1, \dots X_k$ for $k = C/\varepsilon^2$ for some constant C be k independent copies of X. (That is, each $X_t := \sum_i \varepsilon_{t,i} X_i$ where all $\varepsilon_{t,i}$ are independent random variables.) Show that

$$\Pr[|\frac{1}{k}\sum X_i^2 - F^*| > \varepsilon F^*] \le 2/3.$$

We can now use the following algorithm to estimate $\sum F_i^2$ on a sequence of elements $x_1, \dots x_m$. We store values of $X_1, X_2, \dots X_k$ in memory (i.e. only $O(1/\varepsilon^2)$ numbers). Whenever an element x_i appears on a stream, we update all X_t for $t \in [k]$ by $X_t \leftarrow X_t + \varepsilon_{t,x_i}$. At the end as an estimate of F^* we give $\hat{F} := \frac{1}{k}\sum F_i^2$. By the exercise (c), with probability $2/3$, the estimate satisies $(1-\varepsilon)F^* \le \hat{F} \le (1+\varepsilon)F^*$.

Solution 1

(a) Using linearity of expectation we have

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[\sum_{i,j\in[N]} \varepsilon_i\varepsilon_j F_i F_j] \\ &= \sum_{i,j\in[N]} F_i F_j \mathbb{E}[\varepsilon_i\varepsilon_j]\end{aligned}$$

Now since for $i \neq j$ variables ε_i and ε_j are independent we have $\mathbb{E}[\varepsilon_i\varepsilon_j] = \mathbb{E}[\varepsilon_j]\mathbb{E}[\varepsilon_j] = 0$. When $i = j$ we have $\mathbb{E}[\varepsilon_i\varepsilon_j] = \mathbb{E}[\varepsilon_i^2] = 1$, since $\varepsilon_i \in \{\pm 1\}$, and ε_i^2 is deterministic 1. This gives

$$\mathbb{E}[X^2] = \sum_{i\in[N]} F_i^2 = F^*,$$

as desired.

(b) Again, using linearity of expectation, we have

$$\mathbb{E}[X^4] = \sum_{i,j,k,\ell\in[N]} F_i F_j F_k F_\ell \mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell].$$

Note that if among the $\{i,j,k,\ell\}$ any index appears odd number of times, then $\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell]=0$ (since random variables ε_i are independent, and $\mathbb{E}\varepsilon_i = \mathbb{E}\varepsilon_i^3 = 0$). On the other hand, if among the indices $\{i,j,k,\ell\}$ all appear even number of times, we have $\mathbb{E}[\varepsilon_i\varepsilon_j\varepsilon_k\varepsilon_\ell]=1$ (since $\varepsilon_i^{2k} = (\varepsilon_i^2)^k = 1^k = 1$ is deterministic). This gives

$$\mathbb{E}[X^4] = \sum_{i\in[N]} F_i^4 + \sum_{i\neq j\in[N]} 6F_i^2F_j^2 \leq 3\sum_{i,j\in[N]} F_i^2F_j^2 = 3(\sum_i F_i^2)^2.$$

(c) Since each X_i satisfies $\mathcal{E}[X_i^2] = F^*$, and $\mathcal{E}[X_i^4] \leq 3(F^*)^2$, we also have $\mathrm{Var}(X_i^2) = \mathbb{E}X_i^4 - \mathbb{E}[X_i^2]^2 \leq 2(F^*)^2$. Now, by linearity of expectation, if $\hat{X} := \frac{1}{k}\sum_{i\leq k} X_i^2$, we have $\mathbb{E}\hat{X} = F^*$, and since all X_i are indepndentent we have

$$\mathrm{Var}(\hat{X}) = \frac{1}{k^2}\mathrm{Var}(\sum_{i\leq k} X_i^2) = \frac{1}{k}\mathrm{Var}(X_1^2) \leq 2(F^*)^2/k.$$

Now, by Chebyshev inequality

$$\Pr(|\hat{X} - \mathbb{E}[\hat{X}]| > \varepsilon F^*) \leq \frac{\mathrm{Var}[\hat{X}]}{\varepsilon^2(F^*)^2]} \leq \frac{2}{\varepsilon^2 k}.$$

When $k > C/\varepsilon^2$, this probability is bounded by $\frac{2}{C}$. Taking C to be sufficiently large constant we can make this probability as small as desirable.