Lecturers: **Evangelos Kipouridis and Tomasz Kociumaka** **Winter 2025/26**
Tutors: **Anouk Duyster and Yonggang Jiang**

Randomized and Approximation Algorithms, Exercise Sheet 4

https://www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/winter-2025/26/randomized-and-approximation-algorithms

Total Points: **100** + 100 bonus points Due: **14:00** on Thursday, **November 20**, 2025

Exercise 1 **10** points

Recall the balls-into-bins setting, where m balls are placed independently and uniformly at random into n bins. Prove that, if $m = \Omega(n \log n)$, then the maximum bin load is at most

$$\frac{m}{n} + \mathcal{O}\left(\sqrt{\frac{m \log n}{n}}\right)$$

with probability at least $1 - n^{-c}$ for every constant $c = \Theta(1)$.

Exercise 2 **10** + **10** + **10** points + 15 bonus points

Recall the frequency estimation setting: a finite universe U of size n and an unknown subset $S \subseteq U$. We estimate each $|S|$ by sampling s elements $u_1, \ldots, u_s \in U$ uniformly at random and computing

$$E = \frac{n}{s}\left|\{i \in [1\,.\,.\,s] : u_i \in S\}\right|.$$

We proved that, for $\epsilon, \delta \in (0,1)$, a sufficiently large sample size $s = \mathcal{O}(\epsilon^{-2} \ln \delta^{-1})$ guarantees

$$\mathbb{P}[\,\big|\,|S| - E\,\big| \geq \epsilon n\,] \leq \delta.$$

a Show that, for $\epsilon, \delta \in (0,1)$,

$$\mathbb{P}[\,\big|\,|S| - E\,\big| \geq \epsilon |S|\,] \leq \delta$$

holds for sufficiently large $s = \mathcal{O}\left(\frac{n}{|S|}\epsilon^{-2} \ln \delta^{-1}\right)$. *Hint:* Use multiplicative Chernoff bounds.

b Prove that, for every $k \in [1\,.\,.\,N]$ and $\epsilon \in (0,1)$, a sufficiently large $s = \mathcal{O}\big(\frac{n}{k}\epsilon^{-2} \ln \delta^{-1}\big)$ suffices to distinguish between the cases $|S| \leq (1-\epsilon)k$ and $|S| > k$, with error probability at most δ.

c Suppose Alice has n strings $A_1, \ldots, A_n$, each of length m. For any given indices $i \in [1\,.\,.\,m]$ and $j \in [1\,.\,.\,n]$, she can reveal the character $A_j[i]$ to Bob. Bob wants to distinguish between the following two cases for some $k \in [1\,.\,.\,m]$:

- For all distinct indices j, j', the Hamming distance[1] between A_j and $A_{j'}$ is $\mathsf{hd}(A_j, A_{j'}) > k$.
- There exist distinct indices j, j' such that $\mathsf{hd}(A_j, A_{j'}) \leq (1-\epsilon)k$.

Show that Bob can succeed with probability at least $1 - \delta$ using $\mathcal{O}\big(\frac{nm}{k}\epsilon^{-2} \ln \frac{n}{\delta}\big)$ queries.

d (*) Suppose that $n = m$ and Alice's strings are substrings of a single longer string A of length $2n-1$, so that $A_j = A[j\,.\,.\,j+n-1]$. If Bob can directly query characters of A, how many queries are sufficient? *Hint:* Do not worry about the time complexity of your algorithm.

Exercise 3 **15** + **15** points + 5 bonus points

Random variables $X_1, \ldots, X_n$ are said to be *negatively correlated* if the following inequality is satisfied for every subset $S \subseteq [1\,.\,.\,n]$:

$$\mathbb{E}\left[\prod_{i \in S} X_i\right] \leq \prod_{i \in S} \mathbb{E}[X_i].$$

---

[1] $\mathsf{hd}(A_j, A_{j'}) = |\{i \in [1\,.\,.\,m] : A_j[i] \neq A_{j'}[i]\}|$.

a Suppose that we sample elements $s_1, \ldots, s_n$ uniformly at random **without replacement** from a universe U of size at least n (s_i is a uniformly random element of $U \setminus \{s_1, \ldots, s_{i-1}\}$). For a subset $V \subseteq U$, define indicator variables $X_i = \mathbb{1}_{s_i \in V}$. Prove that these variables are negatively correlated.

b Let $X_1, \ldots, X_n : \Omega \to \{0, 1\}$ be negatively correlated random variables, and let $X = \sum_{i=1}^{n} X_i$ with expected value $\mu = \mathbb{E}[X]$. Prove that the multiplicative Chernoff bound for the upper tail still holds; that is, for every $\delta > 0$:

$$\mathbb{P}[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{\mu}.$$

Hint: It suffices to show that $\mathbb{E}[e^{tX}] \leq e^{\mu(e^t - 1)}$ for every $t > 0$; note that $e^{tX_i} = 1 + X_i(e^t - 1)$.

c (*) Extend the previous statement to the case where $X_1, \ldots, X_n : \Omega \to [0, 1]$.

Exercise 4 — **10 + 15 + 5** points + 10 bonus points

Consider a casino that offers a game consisting of n rounds:

- You start with a budget of 1.
- In each round, you first choose a wager (any amount between 0 and your current budget). Then the casino tosses an unbiased coin. Depending on the outcome, your wager is either lost or doubled, and your budget is updated accordingly.
- Your final score is the budget after n rounds.

a Show that, regardless of your strategy, your expected final score is 1.

b Suppose a fortune-teller informs you that the number of rounds you will win (double your wager) will be at least $(1+\delta)\frac{n}{2}$ for some $0 < \delta < 1$ (without revealing which rounds you will win). Design a strategy that guarantees a final score of at least $\exp\big(\Omega(\delta^2 n)\big)$, assuming the fortune-teller is correct. *Hint:* Use the inequality $\ln(1+x) \geq x - x^2$ for $-0.5 \leq x \leq 0.5$; equivalently, $1 + x \geq e^{x - x^2}$ for $-0.5 \leq x \leq 0.5$.

c Deduce that the probability of observing at least $(1+\delta)\frac{n}{2}$ successful tosses is at most $\exp(-\Omega(\delta^2 n))$.

d (*) Improve the bound in part b to $\exp(\delta^2 n/2)$. *Hint:* Use the Taylor expansion of $\ln(1+x)$ for $-1 < x < 1$.

Exercise 5 — 15 + 5 + 10 + 40 bonus points

Recall the Floyd–Rivest median selection algorithm discussed in lecture. We showed that it performs $\frac{3}{2}n + \mathcal{O}(n^{3/4} \log n)$ comparisons, runs in $\mathcal{O}(n)$ time, and fails with probability $\mathcal{O}(n^{-1/4})$.

a (*) Prove that the failure probability is in fact $\exp(-\Omega(n^{1/4}))$. *Hint:* Derive a tighter bound on $\mathbb{P}[\,|\,Y_i - (is)/n\,| \geq c]$, where Y_i is the number of samples among the i smallest elements.

b (*) Choose different parameters for which the algorithm succeeds with high probability and makes at most $\frac{3}{2}n + \mathcal{O}(n^{2/3} \operatorname{polylog} n)$ comparisons.

c (*) Show that the analysis remains valid if samples are drawn *without replacement*. *Hint:* Tail bounds on Y_i are sufficient.

d (**) Design an algorithm that succeeds with high probability and makes at most $\frac{3}{2}n + \mathcal{O}(n^{1/2} \operatorname{polylog} n)$ comparisons. *Hint:* Permute the input array uniformly at random so that the sample can be chosen as a prefix. Imagine running multiple instances of the original algorithm with varying sample sizes s_t, starting small and gradually increasing up to n. View the tth instance as using the $(t-1)$th instance to filter out irrelevant samples.