CS 466/666: Algorithm Design and Analysis **University of Waterloo: Fall 2023**

Lecture 2

September 11, 2023

Instructor: Sepehr Assadi

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Topics of this Lecture

1 Probabilistic Background

We review basic probabilistic background that will be used in this course using a simple running example: Consider the probabilistic process of rolling two dices and observing the result.

- **Probability space**: The set of all possible outcomes of the probabilistic process.

  Here, all the 36 combinations of answers, i.e.,

$$(1,1), (1,2), (1,3), \cdots, (6,5), (6,6).$$

- **Event**: Any subset of the probability space.

  Here, one example of an event would be 'sum of the two dice is equal to 7'; another example is 'both dice rolled an even number'.

- **Probability distribution**: an assignment of $\Pr(e) \in [0,1]$ to every element e of the probability space such that $\sum_e \Pr(e) = 1$.

  Here, every one of the 36 elements of have the same probability 1/36, i.e.,

$$\Pr((1,1)) = \Pr((1,2)) = \Pr((1,3)) = \cdots = \Pr((6,6)) = \frac{1}{36}.$$

- **Probability**: for any event E, probability of E is $\Pr(E) = \sum_{e \in E} \Pr(e)$, i.e., the sum of the probabilities assigned by the probability distribution to the elements of this event.

  Here, for example,

$$\Pr(\text{sum of the two dice is 7}) = \sum_{e \in \{(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)\}} \Pr(e) = 6 \cdot \frac{1}{36} = \frac{1}{6};$$

$$\Pr(\text{both dice roll even}) = \sum_{e \in \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\}} \Pr(e) = 9 \cdot \frac{1}{36} = \frac{1}{4}.$$

- **Random variable**: Any function from the elements of the probability space to integers or reals. I.e., a random variable X is a function $X : \Omega \to \mathbb{R}$, where Ω is the probability space.

  Here, an example of a random variable X is the sum of the two dice, e.g., $X((1,3)) = 4$ and $X((2,5)) = 7$. Another example is a random variable Y that assigns one to the events that both dice roll even and is zero otherwise, e.g. $Y((2,2)) = 1$ and $Y((2,3)) = 0$ (this type of random variable that assigns 1 to a particular event and is zero otherwise is called an *indicator* random variable for that event).

- **Independence**: We say two events E_1 and E_2 are independent of each other, denoted by $E_1 \perp E_2$, whenever $\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2)$.

  For instance, the events 'the first dice rolls even' and 'the second dice rolls odd' are independent of each other, but the events 'the first dice rolls even' and 'both dice rolls even' are not independent.

  Similarly, two random variables X and Y are independent, denoted by $X \perp Y$, if for any values a and b, the events $X = a$ and $Y = b$ are independent. In other words, conditioning on any value of X (resp. Y), does not change the distribution of Y (resp. X).

- **Expected value**: The expected value of a random variable X, denoted by $\mathbb{E}[X]$, is the average of its value *weighted* according to the probability distribution, i.e., $\mathbb{E}[X] = \sum_e \Pr(e) \cdot X(e)$.

  Here, for the random variables X and Y defined two bullets above, we have,

$$\mathbb{E}[X] = \sum_e \Pr(e) \cdot X(e) = \sum_{i \in \{2,\ldots,12\}} \Pr(X = i) \cdot i = 7;$$

$$\mathbb{E}[Y] = \sum_e \Pr(e) \cdot Y(e) = \sum_e \Pr(Y(e) = 1) = \frac{1}{4}.$$

  A very important property of expected value is its *linearity*, often called **linearity of expectation**: for any two random variables X, Y,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

  For *independent* random variables, we further have that expectation is *multiplicative*, i.e., for any two random variables X, Y where $X \perp Y$,

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

- **Variance**: The variance of a random variable X, denoted by $\mathrm{Var}[X]$, is a measure of the 'distance' of an average value of X, from the expected value of X. More accurately,

$$\mathrm{Var}[X] := \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

  Variance of a random variable is a measure of its 'spread': the larger the variance, the more likely that the random variable takes a value 'far' from its expectation. Note that a simple calculation implies

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - 2 \cdot \mathbb{E}[X \cdot \mathbb{E}[X]] + (\mathbb{E}[X])^2 = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$

Note that unlike expectation, variance in general is *not* linear. However, for two *independent* random variables X, Y, it will be linear, i.e.,

$$\begin{aligned}
\mathrm{Var}\left[X+Y\right] &= \mathbb{E}\left[(X+Y)^2\right] - \left(\mathbb{E}\left[X+Y\right]\right)^2 \\
&= \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y\right]^2 + 2\cdot\mathbb{E}\left[X\cdot Y\right] - \left(\mathbb{E}\left[X\right]\right)^2 - \left(\mathbb{E}\left[Y\right]\right)^2 - 2\cdot\mathbb{E}\left[X\right]\cdot\mathbb{E}\left[Y\right] \\
&\hspace{20em}\text{(by expanding the terms)} \\
&= \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 + \mathbb{E}\left[Y^2\right] - \left(\mathbb{E}\left[Y\right]\right)^2 \qquad \text{(as } \mathbb{E}\left[X\cdot Y\right] = \mathbb{E}\left[X\right]\cdot\mathbb{E}\left[Y\right] \text{ when } X\perp Y\text{)} \\
&= \mathrm{Var}\left[X\right] + \mathrm{Var}\left[Y\right].
\end{aligned}$$

The above is an extremely short refresher of the probabilistic background. This cannot by any means replace a proper introduction to this amazing concept. You are strongly encouraged to take a look back at the materials from your previous courses on probability, or the further reading materials on the course page.

2 Concentration Inequalities

When working with random variables, perhaps the easiest way to "summarize" the variable is to focus on its expected value. However, expected value on its own can often be misleading: for instance, consider a random variable which is 0 with probability 1/2 and is 1 with the remaining value. Expected value of X is 1/2 but of course we do not 'expect' X to ever take the value of 1/2! This, and many other examples, suggest that summarizing a random variable just down to its expectation may lose too much information.

On the other extreme, a random variable can be uniquely identified by its probability distribution:

$$\mathbb{P}_X : k \to [0,1] \qquad \text{such that} \qquad \mathbb{P}_X(a) = \Pr(X = a).$$

Yet, the distribution of even very simple random variables can be quite cumbersome to work with. Consider, for instance, the simple example of throwing a fair coin 100 times and defining X to be the number of heads. Here, for every $a \in \{0, \ldots, 100\}$,

$$\Pr(X = a) = \binom{100}{a} \cdot 2^{-100},$$

which, even in this simple form, is rather tedious to work with.

Concentration inequalities are a saving grace between these two extremes: morally speaking (but not strictly speaking true), they allow us to extract (perhaps, the most) "important" information about the distribution of our random variables, without getting to compute the very precise distribution itself. More accurately, they allow us to bound the probability of *deviation* of a random variable from its expectation (as a function of its distance from the expectation).

We will study various concentration inequalities in the course of this term, as they arise quite frequently in the analysis of randomized algorithms (and way beyond). This lecture, includes two of the most basic and highly applicable ones.

2.1 Markov's Inequality

The simplest and most basic variant of concentration results is **Markov's inequality** or **Markov bound**:

Proposition 1 (**Markov Bound**). *For a* non-negative *random variable X and $t > 0$,*

$$\Pr\left(X \geqslant t \cdot \mathbb{E}\left[X\right]\right) \leqslant \frac{1}{t}.$$

Proof. Let $\mu := \mathbb{E}\left[X\right]$. We can use the law of total conditional probabilities to have:

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \mathbb{E}\left[X \mid X \geqslant t\cdot\mu\right]\cdot\Pr\left(X \geqslant t\cdot\mu\right) + \mathbb{E}\left[X \mid X < t\cdot\mu\right]\cdot\Pr\left(X < t\cdot\mu\right) \\
&\geqslant t\cdot\mu\cdot\Pr\left(X\geqslant t\cdot\mu\right) + 0.
\end{aligned}$$

(the first term since we conditioned on $X \geqslant t\cdot\mu$ and the second term since X is non-negative)

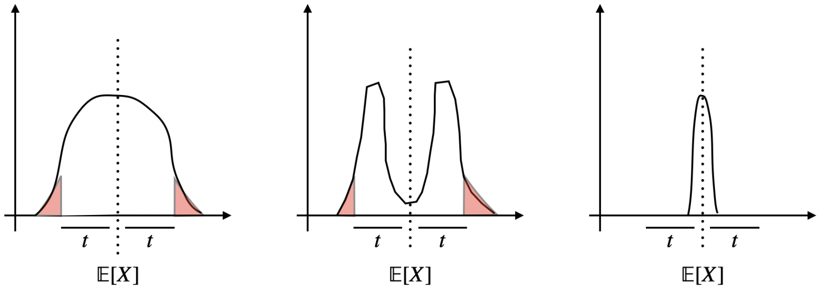

Figure 1: An illustration of three different random variables and their probability distributions. Here, all these variables have the same expected value. Moreover, the first two variables show roughly the same concentration for the particular choice of t; i.e., for the first two variables, the probability that their values are more than t away from the expectation is almost the same (the probability is the part shaded in red) even though the distributions are quite different; however, the third variable is much more concentrated.

Thus, $\Pr\left(X \geqslant t \cdot \mathbb{E}\left[X\right]\right) = \Pr\left(X \geqslant t \cdot \mu\right) \leqslant 1/t$, otherwise the RHS above will be larger than the LHS. □

Markov bound only bounds the *upper tail* of the distribution[1]: the probability that a random variable takes value t times larger than its expectation is at most $1/t$. This is a basic but extremely useful property. One can alternatively state the Markov bound as follows.

Corollary 2. *For a* non-negative *random variable X and $b > 0$,*

$$\Pr\left(X \geqslant b\right) \leqslant \frac{\mathbb{E}\left[X\right]}{b}.$$

Proof. The proof is by simply picking $t = b/\mathbb{E}\left[X\right]$ in Proposition 1. □

Note that a random variable X, which, with probability $\mathbb{E}\left[X\right]/b$ takes the value b and otherwise is 0 will be a tight example for Markov bound; i.e., one cannot expect to improve Markov bound in general.

Remark. Even though Markov bound may sound almost trivial (and it is indeed straightforward), it is the basis for proving all other concentration inequalities that we use in this course; moreover, Markov bound is used one way or another in analysis of almost every randomized algorithm.

2.2 Chebyshev's Inequality

We now consider our second concentration inequality: **Chebyshev's inequality**. Unlike Markov bound that only required a knowledge of the expected value of the random variable to bound its deviation probability, Chebyshev's inequality applies to the settings in which we could additionally bound the *variance* of the random variable as well.[2]

Recall that for a random variable X, **variance** of X is:

$$\mathrm{Var}\left[X\right] := \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2.$$

Chebyshev inequality allows us to bound deviation of a random variable based on its variance.

[1] Although one can use it to bound the lower tail in special cases as well, but the bounds there are generally very weak.

[2] As we showed earlier, Markov bound *can* be tight for certain random variables; thus, naturally whenever we need a stronger bound we should show that our variable satisfies additional guarantees that what is only required by Markov bound.

Proposition 3 (**Chebyshev's Inequality**)**.** *For any random variable X and $t > 0$,*

$$\Pr\left(|X - \mathbb{E}[X]| \geqslant t \cdot \mathbb{E}[X]\right) \leqslant \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2 \cdot t^2}.$$

Proof. Define a new random variable $Y := (X - \mathbb{E}[X])^2$. Clearly, Y is non-negative. Moreover, $|X - \mathbb{E}[X]| \geqslant t \cdot \mathbb{E}[X]$ if and only if $Y = (X - \mathbb{E}[X])^2 \geqslant t^2 \cdot \mathbb{E}[X]^2$. Hence,

$$\begin{aligned}\Pr\left(|X - \mathbb{E}[X]| \geqslant t \cdot \mathbb{E}[X]\right) = \Pr\left(Y \geqslant t^2 \cdot \mathbb{E}[X]^2\right) &\leqslant \frac{\mathbb{E}[Y]}{\mathbb{E}[X]^2 \cdot t^2} && \text{(by Markov bound of Corollary 2)}\\ &= \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2 \cdot t^2}. && \text{(as } \mathbb{E}[Y] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathrm{Var}[X] \text{ by definition)}\end{aligned}$$

□

A useful variant of Chebyshev's inequality is the following.

Corollary 4. *For a random variable X and $b > 0$,*

$$\Pr\left(|X - \mathbb{E}[X]| \geqslant b\right) \leqslant \frac{\mathrm{Var}[X]}{b^2}.$$

Proof. The proof is by simply picking $t = b/\mathbb{E}[X]$ in Proposition 3. □

3 Streaming Distinct Elements Problem

Our next step in this course is to see an example of Chebyshev's inequality in designing randomized algorithm in the *streaming* model of computation. A streaming algorithm processes its inputs in small chunks, one at a time, and thus does not need to store the entire input in one place. For motivation, consider a router in a network: the router needs to process a massive number of packets using a limited memory much smaller than what allows for storing all the packets it sees during its process.

3.1 The Streaming Model of Computation

We now define the streaming model formally as introduced by Alon, Matias, and Szegedy in [AMS96][3]. The input consists of n elements $e_1, e_2, \ldots, e_n$, where each e_i belongs to some universe $\mathcal{U}$ of m elements, that are received one at a time by the algorithm, sequentially. Every time a new element is received, the previous one is erased, so the algorithm only has access to the most recent element. The algorithm has a local memory available, separate from the input, which is (ideally) much smaller than the input (and so we cannot store the input entirely by the end of the stream).

The goal in this model is to design algorithms that use only a small amount of memory compared to the input size, typically (but not always) of size $\mathrm{poly}(\log n, \log m)$ bits.

Warm-Up: Before getting to our main problem, let us mention a standard warm-up puzzle:

- You are given $n - 1$ *distinct* numbers from $[n]$ in a stream in some arbitrary order. Find the missing element in $O(\log n)$ bits of space.

[3] The 2005 Gödel Prize—one of the highest award in Theoretical Computer Science—was awarded to Noga Alon, Yossi Matias, and Mario Szegedy for the introduction of this model. See the citation of the award here:

https://eatcs.org/index.php/component/content/article/503

- If you like to challenge yourself further, consider the same problem where you are given $n-k$ distinct numbers from $[n]$ and the goal is to find the k missing elements in $O(k^2 \cdot \log n)$ bits of space.
- If you really like to challenge yourself, solve the above problem in $O(k \log n)$ bits of space.

We will leave the answer to these questions as an exercise to the reader (*Note:* the first two questions are simple enough but the last one might be quite challenging without the "right" background—do not let that discourage you however!).

3.2 Distinct Elements Counting Problems

We now consider one of the first (and highly influential) problems considered in the streaming model, namely, the **distinct element (counting)** problem.

Problem 1. Given a stream of n elements from the universe $[m]$, output the number of *distinct* elements in the stream, denoted by DE.

For example, if $m = 5$, and the stream is $1, 2, 2, 1, 5, 4, 2, 2, 1$, then the answer is $\mathsf{DE} = 4$.

There are two naive solutions to this problem:

- Store the entire universe: Use a bitmap with m bits. Every time we see a new element, mark it. This requires $O(m)$ bits.
- Store the entire stream: Store a set of all the elements we receive. This requires $O(n \log m)$ bits.

These type of straightforward solutions are applicable to most streaming problems.

What about an algorithm using $\text{poly}(\log n, \log m)$ bits? While we will most likely not cover this topic in this course, one can show that this is not possible without randomization and approximation, by proving the following *lower bounds*:

- Every deterministic algorithm requires $\min\{\Omega(n), \Omega(m)\}$ bits, even if it is a, say, 1.1-approximation.
- Every exact randomized algorithm requires $\min\{\Omega(n), \Omega(m)\}$ bits.

Therefore, to find a sublinear space streaming algorithm we need to allow for both approximation and randomization. In addition, to make the problem a bit easier for us in this lecture, we will consider a relaxed version of the problem, sometimes called *threshold testing* (variant) of the problem, defined as follows:

- At the beginning of the stream, you are given a value $\widetilde{\mathsf{DE}} \in [m]$ as an 'estimated threshold' for the value of DE and a parameter $\varepsilon \in (0, 1)$; you need to, with probability at least $2/3$, output *Yes* if $\mathsf{DE} \geqslant \widetilde{\mathsf{DE}}$ and output *No* if $\mathsf{DE} < (1-\varepsilon) \cdot \widetilde{\mathsf{DE}}$; if the value of DE is between these two numbers, either answer is considered correct.

In the next lecture, we will give an algorithm that solves this problem using

$$\text{poly}(\log n, \log m, 1/\varepsilon)$$

bits of space, which is much more efficient than the naive approaches above.

References

[AMS96] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22-24, 1996*, pages 20–29, 1996. 5