Statistical vs. Computational Distance
—Lecture 6—

Christopher Brzuska

January 12, 2024

1 Statistical Tools: Markov and Chernoff Bound

In this lecture, we want to perform some statistical analysis of algorithms, and we will need several statistical tools in order to do this. For a randomized algorithm $\mathcal{A}$ which output a real number, we can define the *expectation* of $\mathcal{A}$ as

$$\sum_{z \in \mathrm{Supp}(\mathcal{A})} z \cdot \Pr[\mathcal{A} = z],$$

which the the average number which $\mathcal{A}$ outputs. For us, $\mathcal{A}$ will often be a randomized experiment or an algorithm which takes as input λ random bits, and in this case, we can write the expectation equivalently as

$$\mathbb{E}(\mathcal{A}) = \sum_{r \in \{0,1\}^\lambda} 2^{-\lambda} \mathcal{A}(r).$$

We have seen such probability statements already many times in the course in the case where the algorithm (adversary) $\mathcal{A}$ returns 0 or 1. Note that each r is chosen with probability $2^{-\lambda}$. We call an algorithm that maps random strings to a real number a *random variable*. We will now see two bounds that tell us whether a random variable is likely to be far from its expectation, and if so, how likely. To appreciate the quality of the different bounds, let us consider the following example:

> Example: We flip a coin which has a 0 on one side and a 1 on the other side, such that the probability of getting 0 is $\frac{1}{2}$. Now, if we flip a coin 1000 times, we get 500 zeroes in expectation. How likely is it that we get much more than this, say, that we get 750 zeroes?

My intuition says that this should be quite unlikely. Let's now look at some popular tail bounds and see what they say about our example. The first bound is the Markov bound that is valid for all random variables that are positive. In the lemma, we replace the explicit notation $\Pr_{r \leftarrow\$ \{0,1\}^n}[\mathcal{A}(r) \geq v]$ by the implicit notation $\Pr[\mathcal{A} \geq v]$, since this allows us to speak about arbitrary randomized algorithms/random variables without making explicit how long the random string is.

Lemma 1 (Markov Bound)**.** For all non-negative random variables $\mathcal{A}$ and all positive real numbers v, we have

$$\Pr[\mathcal{A} \geq v] \leq \frac{\mathbb{E}(\mathcal{A})}{v}$$

and

$$\Pr[\mathcal{A} \geq v \cdot \mathbb{E}(\mathcal{A})] \leq \frac{1}{v}.$$

The expectation of a random variable gives some information about a random variable, but not necessarily very much. For instance, the Markov bound only tells us some very weak relation between a random variable and its expectation. The Markov bound is not particularly good for repeated, independent experiments such as our example. However, Markov bound only tells us that this probability is lower than $\frac{2}{3}$ (which is not very informative). Therefore, we need better bounds such as the Chernoff bound (below) which tells us that the probability is really small. We now turn to repeated experiments with $0-1$ random variables Exp, i.e, random variables that either return 0 or 1 such as our security experiments.

Lemma 2 (Chernoff Bound)**.** For all $p \leq \frac{1}{2}$, for all $\mathsf{Exp}_1, \mathsf{Exp}_2, ..., \mathsf{Exp}_n$ independent 0-1 random variables so that for all i $\Pr[\mathsf{Exp}_i = 1] = p$, for all ϵ, $0 < \epsilon \leq p(1-p)$, we have

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} \mathsf{Exp}_i}{n} - p\right| > \epsilon\right] < 2 \cdot e^{-\epsilon^2 n}$$

and

$$\Pr\left[\sum_{i=1}^{n} \mathsf{Exp}_i > (p+\epsilon)n\right] < 2 \cdot e^{-\epsilon^2 n}$$

In our example, $p = \frac{1}{2}$, $n = 1000$ and $\epsilon = \frac{1}{4}$, which yields that the probability of obtaining more than 750 zeroes is very small.

Definition 1 (Expectation)**.** When $\mathcal{A}$ is a randomized algorithm or experiment which outputs a real number (e.g., 0, or 1), we define its expectation as

$$\mathbb{E}(\mathcal{A}) := \sum_{z \in \mathrm{Supp}(\mathcal{A})} z \cdot \Pr[\mathcal{A} = z].$$

Lemma 1 (Markov Bound)**.** For all non-negative random variables $\mathcal{A}$ and all positive real numbers v, we have

$$\Pr[\mathcal{A} \geq v] \leq \frac{\mathbb{E}(\mathcal{A})}{v}$$

and

$$\Pr[\mathcal{A} \geq v \cdot \mathbb{E}(\mathcal{A})] \leq \frac{1}{v}.$$

Lemma 2 (Chernoff Bound)**.** For all $p \leq \frac{1}{2}$, for all $\mathsf{Exp}_1, \mathsf{Exp}_2, ..., \mathsf{Exp}_n$ independent 0-1 random variables so that for all i $\Pr[\mathsf{Exp}_i = 1] = p$, for all ϵ, $0 < \epsilon \leq p(1-p)$, we have

$$\Pr\left[\left|\frac{\sum_{i=1}^{n} \mathsf{Exp}_i}{n} - p\right| > \epsilon\right] < 2 \cdot e^{-\epsilon^2 n}$$

and

$$\Pr\left[\sum_{i=1}^{n} \mathsf{Exp}_i > (p+\epsilon)n\right] < 2 \cdot e^{-\epsilon^2 n}$$