Markov's inequality becomes particularly easy to memorize if one looks at the interpretation for finite random variables. For example, assume that we have n non-negative numbers $a_1, \ldots, a_n$, and we consider the random variable X that chooses a number uniformly at random. Then the expected value of X is just the arithmetic mean $\hat{a}$ of the numbers:

$$\mathbf{E}[X] = \sum_{i=1}^{n} \frac{1}{n} a_i = \frac{1}{n} \sum_{i=1}^{n} a_i = \hat{a}.$$

It is easy to prove that not too many numbers can deviate very much from $\hat{a}$. It is also intuitive: In a group of people with average income $\hat{a}$, not more than half of the people can earn $2\hat{a}$. Otherwise, the average would be higher. Similarly, assume that more than n/k of the a_i satisfy that $a_i \geq k\hat{a}$. Since the a_i are non-negative, the sum of all numbers is not smaller than the sum of these more than n/k numbers. Thus:

$$\sum_{i=1}^{n} a_i > \frac{n}{k} k\hat{a} = n \frac{1}{n} \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} a_i.$$

Clearly, $\sum_{i=1}^{n} a_i > \sum_{i=1}^{n} a_i$ is impossible, so our assumption was wrong. There can at most be n/k numbers a_i that satisfy $a_i \geq k\hat{a}$. This is exactly the statement of Markov's inequality for X:

$$\begin{aligned} & \frac{|\{a_i \mid a_i \geq k \cdot \hat{a}\}|}{n} = \mathbf{Pr}(X \geq k \cdot \mathbf{E}[X]) \leq \frac{1}{k} \\ \Leftrightarrow \quad & |\{a_i \mid a_i \geq k \cdot \hat{a}\}| \leq \frac{n}{k}. \end{aligned}$$

Observe that we used that the numbers are non-negative to argue that $\sum_{i=1}^{n} a_i > \frac{n}{k} k\hat{a}$. Indeed, if the numbers can be negative (like in the list of numbers $2, 2, 2, 2, -8$), then a large fraction of the numbers can be larger than the expected value ($4/5$ in the example, or $(n-1)/n$ if we have $n-1$ twos and $-2(n-1)$ as the nth number). They can even be a lot larger than the expected value (just make the nth number $-k$ for a large $k \in \mathbb{N}$).

Implications of Markov's inequality. We directly obtain the following consequences from Markov's inequality. Think of X as the running time of an algorithm or the approximation ratio of a randomized approximation algorithm. In both cases, $X \geq 0$ and $\mathbf{E}[X]$ exists. By Markov's inequality,

$$\mathbf{Pr}(X \geq (1+\varepsilon)\mathbf{E}[X]) \leq \frac{1}{1+\varepsilon}$$

holds for any constant $\varepsilon > 0$. We can use independent repetitions to decrease the probability even further. If X is the running time of an algorithm, stop the algorithm after $(1+\varepsilon)\mathbf{E}[X]$ steps. If the algorithm was not finished, try again, doing at most t repetitions. If X is the approximation ratio, simply repeat the algorithm t times and output the best solution. Let X_i be the running time / approximation ratio in the ith run, and set $t = \log_{1+\varepsilon} n \in \mathcal{O}(\log_2 n)$. Then

$$\mathbf{Pr}(\min_{i=1,\ldots,t} X_i \geq (1+\varepsilon)\mathbf{E}[X]) \leq \frac{1}{(1+\varepsilon)^t} = \frac{1}{n},$$

so the probability that we have at least one run with $X \leq (1+\varepsilon)\mathbf{E}[X]$ in $\Theta(\log n)$ tries is at least $1 - \frac{1}{n}$, and obtaining this statement only needed Markov's I nequality.

3.1 Variance and Chebyshev's inequality

Before we prove Chebyshev's inequality, we need to define the *variance* of a random variable X. It is the expected value of the squared deviation of X from its expected value and thus a very useful quantity for obtaining a tail bound.

Definition 3.2. *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space, let $X : \Omega \to \mathbb{R}$ be a discrete random variable. Assume that $\mathbf{E}[(X - \mathbf{E}[X])^2]$ exists. Then the* variance $\mathbf{Var}[X]$ *of X exists and is defined as*

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

Observe that $\mathbf{E}[X]$ is a constant. If it appears inside another expected value, we can use linearity. We use this fact to prove that the two equivalent definitions of the variance are indeed equal.

$$\begin{aligned}
\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] &= \mathbf{E}\left[X^2 - 2X\mathbf{E}[X] + (\mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2\right] - 2\mathbf{E}[X \cdot \mathbf{E}[X]] + \mathbf{E}\left[(\mathbf{E}[X])^2\right] \\
&= \mathbf{E}\left[X^2\right] - 2\mathbf{E}[X] \cdot \mathbf{E}[X] + (\mathbf{E}[X])^2 \\
&= \mathbf{E}\left[X^2\right] - 2(\mathbf{E}[X])^2 + (\mathbf{E}[X])^2 \\
&= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2.
\end{aligned}$$

Since the variance is inherently a squared measure, it is common to also define the *standard deviation* $\sigma(X)$ as $\sqrt{\mathbf{Var}[X]}$.

Example 3.3. *If X always attains the same value, e.g. $X = 7$, then $\mathbf{E}[X] = X$ and thus $\mathbf{E}[(X - \mathbf{E}[X])] = 0$, the random variable has zero variance. We have also seen an example for the other extreme case in Example 2.8 where we chose each number $i \in \mathbb{N}$ with probability c/i^4 for a constant $c > 0$ and let X be the chosen number. We saw that $\mathbf{E}[X]$ exists, but $\mathbf{E}[X^2]$ does not. Thus, $\mathbf{Var}[X]$ also does not exist.*

The rules for computing expected values in Theorem 2.7 can be used to prove the following rules for computing the variance of random variables.

Lemma 3.4. *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space and let $X, Y : \Omega \to \mathbb{R}$ be discrete random variables. Assume that $\mathbf{Var}[X]$ and $\mathbf{Var}[X]$ exist, let $a, b \in \mathbb{R}$ be two real values. Then the following statements hold.*

1. *$\mathbf{Var}[aX + b]$ exists and $\mathbf{Var}[aX + b] = a^2 \cdot \mathbf{Var}[X]$.*
2. *If X and Y are independent, then $\mathbf{Var}[X+Y]$ exists and $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$.*

Proof. Exercise. □

The second rule can be extended for sums of n random variables by induction.

Corollary 3.5. *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space and let $X_1, X_2, \ldots, X_n$ be independent random variables, assume that $\mathbf{Var}[X_i]$ exists for all $i \in \{1, \ldots, n\}$. Let $X := \sum_{i=1}^{n} X_i$. Then $\mathbf{Var}[X]$ exists and it holds that $\mathbf{Var}[X] = \sum_{i=1}^{n} \mathbf{Var}[X_i]$.*

We can use Corollary 3.5 to compute the variance of binomially distributed variables.

Example 3.6. *Let X be a Bernoulli variable with parameter p. We know that $\mathbf{E}[X^2]$ exists since Ω is finite in this case. We observe that $X^2 = X$ is always true, and thus, $\mathbf{E}[X^2] = p$. We use this and compute the variance of X:*

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X]) = p - p^2 = p(1-p).$$

By Corollary 3.5, we conclude that the variance of a binomially distributed variable $Y = Y_1 + \ldots + Y_n$ also exists and it is $\mathbf{Var}[Y] = np(1-p)$.

Now we can state and prove Chebyshev's inequality. The variance is the expected value of $X - \mathbf{E}[X]$, so it is an expected value itself. Chebychev's inequality arises from applying Markov's inequality to this expected value.

Theorem 3.7 (Chebyshev's inequality)**.** *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space, let $X : \Omega \to \mathbb{R}$ be a discrete random variable. Assume that $\mathbf{E}[X]$ and $\mathbf{Var}[X]$ exist. Let $a > 0$ be a constant. Then it holds that*

$$\mathbf{Pr}(|X - \mathbf{E}[X]\,| \geq a) \leq \frac{\mathbf{Var}[X]}{a^2}.$$

This also implies that

$$\mathbf{Pr}(|X - \mathbf{E}[X]\,| \geq t \cdot \mathbf{E}[X]) \leq \frac{\mathbf{Var}[X]}{t^2 \cdot (\mathbf{E}[X])^2}$$

holds for all $t > 1$.

Proof. First, we observe that

$$\mathbf{Pr}(|X - \mathbf{E}[X]\,| \geq a) = \mathbf{Pr}((X - \mathbf{E}[X])^2 \geq a^2).$$

Now, we apply Markov's inequality to the random variable $(X - \mathbf{E}[X])^2$. Notice that the expected value of this variable is $\mathbf{Var}[X]$, so it exists by the precondition of the theorem. Furthermore, observe that $(X - \mathbf{E}[X])^2$ only attains non-negative values. Thus, we can apply Markov's inequality and obtain

$$\mathbf{Pr}((X - \mathbf{E}[X])^2 \geq a^2) \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{a^2} = \frac{\mathbf{Var}[X]}{a^2},$$

the inequality we wanted to prove. For the second inequality, set $a = t \cdot \mathbf{E}[X]$. □

3.2 Chernoff/Rubin bounds

Chernoff bounds are a powerful type of concentration bounds. They are named after Herman Chernoff. However, Chernoff attributes them to Herman Rubin, for example in an article by himself [Che04] and in an interview [Bat96], thus they should rather be called *Rubin bounds*. Chernoff/Rubin bounds are obtained by applying Markov's inequality to the moment generating function of X. We skip the definition of moment generating functions and the derivation of Chernoff/Rubin bounds and just cite two versions. Both are proven in Section 4.2 of [MU05], see Theorem 4.4(2.+3.)/4.5(2.). Notice that the strongest bounds in [MU05] are Theorem 4.4(1.) and Theorem 4.5(1.), but we are satisfied with the slightly simplified variants.

Theorem 3.8. *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space and let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameters $p_1, \ldots, p_n$. Let $X = X_1 + \ldots + X_n$. Observe that $\mathbf{E}[X] = p_1 + \ldots, p_n$.*

1. *For every $0 < \delta \leq 1$, it holds that*
$$\mathbf{Pr}\Big(X \geq (1+\delta) \cdot \mathbf{E}[X]\Big) \leq e^{-\mathbf{E}[X] \cdot \delta^2/3}.$$
2. *For every $0 < \delta < 1$, it holds that*
$$\mathbf{Pr}\Big(X \leq (1-\delta) \cdot \mathbf{E}[X]\Big) \leq e^{-\mathbf{E}[X] \cdot \delta^2/2}.$$

The second version will be helpful for our second application lateron. It allows us to use upper bounds instead of the expected value, and it has an easy to memorize form.

Theorem 3.9. *Let $(\Omega, \mathbf{Pr})$ be a discrete probability space and let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameters $p_1, \ldots, p_n$. Let $X = X_1 + \ldots + X_n$. Then it holds for every $b \geq 6 \cdot \mathbf{E}[X] = 6(p_1 + \ldots + p_n)$ that*
$$\mathbf{Pr}(X \geq b) \leq 2^{-b}.$$

The following example compares the three types of concentration bounds that we have now seen.

Example 3.10. *We consider n independent tosses of a fair coin. We use the random variable X to model the number of heads that we see. This variable is binomially distributed with parameters n and $1/2$, and its expected value is $n/2$. With Markov's inequality, we observe that*
$$\mathbf{Pr}(X \geq \frac{3}{4}n) \leq \frac{\mathbf{E}[X]}{\frac{3}{4}n} = \frac{n/2}{\frac{3}{4}n} = \frac{2}{3}.$$
To apply Chebyshev's inequality, we observe that $\mathbf{Var}[X]$ exists and that it holds that $\mathbf{Var}[X] = n \cdot (1/2) \cdot (1 - (1/2)) = (n/4)$. Chebyshev's inequality thus gives the stronger bound
$$\mathbf{Pr}(X \geq \frac{3}{4}n) \leq \mathbf{Pr}\left(\left|X - \frac{1}{2}n\right| \geq \frac{1}{4}n\right) \leq \frac{\mathbf{Var}[X]}{(\frac{1}{4}n)^2} = \frac{n/4}{n^2/16} = \frac{4}{n}.$$

This bound is stronger. It tells us that the probability for $(3/4)n$ *heads in* n *coin tosses decreases polynomially in* n*. We observe that Theorem 3.8 can also be applied to binomially distributed random variables since it is a sum of Bernoulli variables. We apply the first inequality with* $\delta = 1/2$ *since* $(1 + (1/2)) \cdot (n/2) = (3/4)n$*. The result is that*

$$\mathbf{Pr}(X \geq \frac{3}{4}n) \leq e^{-(\mathbf{E}[X]\delta^2)/3} = e^{-\frac{1}{2}\cdot n\cdot\frac{1}{4}\cdot\frac{1}{3}} = e^{-\frac{n}{24}}.$$

That is the strongest of the three bounds, it says that the probability to see $(3/4)n$ *heads in* n *coin flips decreases exponentially in* n*.*

3.3 Applications

Now we discuss two different types of applications of concentration bounds. The first application is from statistics, we want to estimate the parameter of a distribution. That is a common task in statistics, we only look at the easy case of repeated Bernoulli experiments. The second application is from the area of parallel computing. We discuss oblivious routing strategies in a special type of graphs, namely hypercube graphs.

3.3.1 Estimating parameters

Assume that we are conducting a street survey to determine how many inhabitants of Bonn are fans of the Cologne soccer club 1. FC Köln. How many people do we need to ask to be sure that our estimation is close to the truth, assuming that the participants of our survey are chosen uniformly at random from the population of Bonn?

Observe that we want to estimate the parameter p of a Bernoulli random variable: When $100 \cdot p$ % of the population of Bonn are a fan, then the probability that a person chosen uniformly at random is a fan is p. Conducting our poll means that we repeat the Bernoulli experiment independently, giving rise to a binomially distributed variable X. We want to choose the number of repetitions n such that we can get a good estimation of p with high probability. Assume that we see $\tilde{p} \cdot n$ ones in n tries. The expected number of ones would be $p \cdot n$. The following definition formalizes the type of statement that we want to obtain.

Definition 3.11. *A* γ*-confidence interval for a parameter* p *is an interval* $[\tilde{p} - \delta, \tilde{p} + \delta]$ *such that*

$$\mathbf{Pr}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \geq \gamma.$$

For example, we might want to estimate the fraction p of the population within a 5% error, i.e. $\delta = 0.05$, and our estimation shall be correct with probability 99%, i.e. $\gamma = 0.99$. First, we rewrite the probability that p lies within $[\tilde{p} - \delta, \tilde{p} + \delta]$.

$$\begin{aligned}\mathbf{Pr}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) &= 1 - \mathbf{Pr}(p \notin [\tilde{p} - \delta, \tilde{p} + \delta]) \\ &= 1 - \mathbf{Pr}((p < \tilde{p} - \delta) \cup (p > \tilde{p} + \delta)) \\ &= 1 - \mathbf{Pr}(p < \tilde{p} - \delta) - \mathbf{Pr}(p > \tilde{p} + \delta)\end{aligned}$$

$$= 1 - \mathbf{Pr}(\tilde{p} > p + \delta) - \mathbf{Pr}(\tilde{p} < p - \delta)$$
$$= 1 - \mathbf{Pr}(\tilde{p}n > np + n\delta)) - \mathbf{Pr}(\tilde{p} < np - n\delta).$$

Second, we observe that np is the expected value of X, furthermore, $n\delta = np \cdot \delta/p = (\delta/p)\mathbf{E}[X]$. We can thus rewrite the failure probability into a term that we can apply Theorem 3.8 to, then we apply the theorem.

$$\begin{aligned}
&\mathbf{Pr}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \\
&= 1 - \mathbf{Pr}(\tilde{p}n > np + n\delta)) - \mathbf{Pr}(\tilde{p} < np - n\delta) \\
&= 1 - \mathbf{Pr}(X > \mathbf{E}[X] + \mathbf{E}[X] \cdot (\delta/p)) - \mathbf{Pr}(X < \mathbf{E}[X] - \mathbf{E}[X] \cdot (\delta/p)) \\
&= 1 - \mathbf{Pr}\Big(X > \mathbf{E}[X]\left(1 + \frac{\delta}{p}\right)\Big) - \mathbf{Pr}\Big(X < \mathbf{E}[X]\left(1 - \frac{\delta}{p}\right)\Big) \\
&> 1 - e^{-\mathbf{E}[X]\cdot\delta^2/(3p^2)} - e^{-\mathbf{E}[X]\cdot\delta^2/(2p^2)}.
\end{aligned}$$

In this form, the bound is not very useful because we do not know p. However, we know that $p \leq 1$, and replacing p by a larger value makes the above term smaller. We thus do this and then bring our bound in a short form.

$$\begin{aligned}
&\mathbf{Pr}(p \in [\tilde{p} - \delta, \tilde{p} + \delta]) \\
&> 1 - e^{-\mathbf{E}[X]\cdot\delta^2/(3p^2)} - e^{-\mathbf{E}[X]\cdot\delta^2/(2p^2)} \\
&= 1 - e^{-(np)\cdot\delta^2/(3p^2)} - e^{-(np)\cdot\delta^2/(2p^2)} \\
&= 1 - e^{-n\cdot\delta^2/(3p)} - e^{-n\cdot\delta^2/(2p)} \\
&\geq 1 - e^{-n\cdot\delta^2/3} - e^{-n\cdot\delta^2/2} \\
&\geq 1 - 2e^{-n\cdot\delta^2/3} \\
&\geq 1 - 2 \cdot \frac{1}{e^{n\cdot\delta^2/3}}.
\end{aligned}$$

We want that $1 - 2 \cdot \frac{1}{e^{n\cdot\delta^2/3}} \geq \gamma$. Rearranging this for n tells us that

$$n \geq \frac{3}{\delta^2} \cdot \ln\Big(\frac{2}{1-\gamma}\Big)$$

repetitions are sufficient to ensure that p lies in the interval $[\tilde{p}-\delta, \tilde{p}+\delta]$ with probability at least γ. For $\delta = 0.05$ and $\gamma = 0.01$, we get that

$$\frac{3}{\delta^2} \cdot \ln\Big(\frac{2}{1-\gamma}\Big) = \frac{3}{(0.05)^2} \cdot \ln\Big(\frac{2}{1-0.01}\Big) \leq 6358$$

repetitions are sufficient. Notice that this is a constant – it does not matter how large the population is to obtain the γ-confidence interval. However, our evaluation crucially depended on the assumption that we can choose the participants uniformly at random.