Lecture 2

1 Concentration Inequalities and Limit Theorems

Theorem 1.1 (Chernoff-Hoeffding inequalities)**.** *Let $X_1, \ldots, X_n$ be mutually independent random variables such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$ for every $1 \leq i \leq n$, and let $X = \sum_{i=1}^n X_i$. Then, for every $t > 0$ it holds that*

$$\mathbb{P}(X \geq t) \leq e^{-t^2/(2n)},$$
$$\mathbb{P}(X \leq -t) \leq e^{-t^2/(2n)}.$$

In the proof of Theorem 1.1 we will make use of the following technical lemma.

Lemma 1.2. *For every real number $\lambda > 0$ it holds that*

$$\frac{e^\lambda + e^{-\lambda}}{2} \leq e^{\lambda^2/2}.$$

Proof. It follows from the Taylor series of the three functions e^λ, $e^{-\lambda}$, and $e^{\lambda^2/2}$ that

$$e^\lambda + e^{-\lambda} = \sum_{n=0}^\infty \frac{\lambda^n}{n!} + \sum_{n=0}^\infty \frac{(-1)^n \lambda^n}{n!} = 2\sum_{n=0}^\infty \frac{\lambda^{2n}}{(2n)!} \leq 2\sum_{n=0}^\infty \frac{\lambda^{2n}}{n!2^n} = 2\sum_{n=0}^\infty \frac{(\lambda^2/2)^n}{n!} = 2e^{\lambda^2/2},$$

where the inequality above follows from the fact that $(2n)! \geq n!2^n$ holds for every positive integer n (this is very easy to verify algebraically but also follows since $\frac{(2n)!}{n!2^n}$ is the number of ways to partition $2n$ people into n pairs and thus a non-negative integer; since, moreover, it is clearly positive, it must be at least 1). □

Proof of Theorem 1.1. We will prove that $\mathbb{P}(X \geq t) \leq e^{-t^2/(2n)}$; the proof of the complementary inequality is left as an exercise. For every real number $\lambda > 0$ and $1 \leq i \leq n$, it holds that

$$\mathbb{E}\left(e^{\lambda X_i}\right) = \frac{e^\lambda + e^{-\lambda}}{2} \leq e^{\lambda^2/2},$$

where the above inequality holds by Lemma 1.2. Since the random variables $X_1, \ldots, X_n$ are mutually independent by assumption, the random variables $e^{\lambda X_1}, \ldots, e^{\lambda X_n}$ are mutually independent as well. Hence

$$\mathbb{E}\left(e^{\lambda X}\right) = \mathbb{E}\left(e^{\lambda \sum_{i=1}^n X_i}\right) = \mathbb{E}\left(\prod_{i=1}^n e^{\lambda X_i}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{\lambda X_i}\right) \leq e^{\lambda^2 n/2}.$$

Therefore

$$\mathbb{P}\left(X \geq t\right) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda t}\right) \leq \frac{\mathbb{E}\left(e^{\lambda X}\right)}{e^{\lambda t}} \leq e^{\lambda^2 n/2 - \lambda t}, \tag{1}$$

where the first equality holds since $f(x) := e^{\lambda x}$ is an increasing function for every $\lambda > 0$ and the first inequality holds by Markov's inequality (note that $e^{\lambda X}$ is a non-negative random variable). Since our goal is to find the best possible upper bound on $\mathbb{P}\left(X \geq t\right)$ and since the inequality (1) holds for every $\lambda > 0$, we wish to minimize $g(\lambda) := e^{\lambda^2 n/2 - \lambda t}$ over $(0, \infty)$. Differentiating g yields

$$g'(\lambda) = (\lambda n - t)e^{\lambda^2 n/2 - \lambda t}$$

Comparing to zero then yields $\lambda = t/n$. It is easy to see that this point is indeed a minimum, as for every $\varepsilon > 0$ it holds that $g'(t/n - \varepsilon) < 0$ and $g'(t/n + \varepsilon) > 0$, that is, the function g is decreasing in $(0, t/n)$ and increasing in $(t/n, \infty)$. Substituting $\lambda = t/n$ in (1) yields

$$\mathbb{P}\left(X \geq t\right) \leq e^{t^2/(2n) - t^2/n} = e^{-t^2/(2n)}$$

as claimed. □

1.1 Limit Theorems

Theorem 1.3 (Weak Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of mutually independent identically distributed random variables with finite expectation μ. Then, for every $\varepsilon > 0$, it holds that*

$$\lim_{n \to \infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

We will prove only the special case of Theorem 1.3 in which the X_i's have a finite variance σ^2.

Proof of Theorem 1.3. By linearity of expectation

$$\mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)}{n} = \mu$$

holds for every positive integer n. Moreover, since the random variables $\{X_i\}_{i=1}^{\infty}$ are mutually independent

$$Var\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{Var(X_1) + \ldots + Var(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

holds for every positive integer n. Hence, for every $\varepsilon > 0$, it follows by Chebyshev's inequality that

$$\begin{aligned}\mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mathbb{E}\left(\frac{X_1 + \ldots + X_n}{n}\right)\right| \geq \varepsilon\right) \\ &\leq \frac{Var\left(\frac{X_1 + \ldots + X_n}{n}\right)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n}.\end{aligned}$$

We conclude that

$$0 \leq \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{\sigma^2}{\varepsilon^2 n} = 0$$

and thus

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0$$

as claimed. □

Theorem 1.4 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of mutually independent identically distributed random variables with finite expectation μ and finite variance. Then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \ldots + X_n}{n} = \mu\right) = 1.$$

At first glance the two laws of large numbers appear quite similar and it is not that obvious that, as their names suggest, Theorem 1.4 is stronger than Theorem 1.3. Both of them state that, under certain assumptions, some sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ converges to the random variable μ. The difference between these two theorems lies in the modes of convergence. Let us first define the two relevant modes of convergence.

Definition 1.5. *A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables is said to converge* in probability *to a random variable X as n tends to infinity, if for every $\varepsilon > 0$, it holds that*

$$\lim_{n\to\infty} \mathbb{P}\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}\right) = 0.$$

We denote this fact by $X_n \xrightarrow{p} X$.

Definition 1.6. *A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables is said to converge* almost surely *(or, with probability 1) to a random variable X as n tends to infinity, if*

$$\mathbb{P}\left(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1.$$

We denote this fact by $X_n \xrightarrow{a.s.} X$.

Looking back on Theorems 1.3 and 1.4, we see that the weak law of large numbers asserts that (under certain assumptions) $1/n \cdot \sum_{i=1}^{n} X_i \xrightarrow{p} \mu$ whereas the strong law of large numbers asserts that $1/n \cdot \sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$. It is known that almost sure convergence implies convergence in probability but the converse implication does not hold in general. The following example demonstrates the latter (we will not prove the former in this course).

Example 1: Let X denote the random variable 0 and let $\{X_n\}_{n=1}^{\infty}$ be a sequence of mutually independent random variables such that $\mathbb{P}(X_n = 1) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - 1/n$ hold for every positive integer n. Fix some $\varepsilon > 0$. Then

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) \leq \lim_{n\to\infty} \mathbb{P}(X_n = 1) = \lim_{n\to\infty} 1/n = 0.$$

That is, $X_n \xrightarrow{p} X$.

On the other hand, suppose for a contradiction that $X_n \xrightarrow{a.s.} X$. By Definition 1.6 this means that, with probability 1, for every $\varepsilon > 0$, there exists an integer m such that $|X_n - X| \leq \varepsilon$ holds for every integer $n \geq m$. It then follows by the distributions of the X_n's that, with probability 1, there exists an integer m such that $X_n = 0$ holds for every integer $n \geq m$. However, since the X_n's are mutually independent, for every positive integer m, the probability that $X_n = 0$ for every $n \geq m$ is

$$\prod_{n=m}^{\infty} (1 - 1/n) \leq \prod_{n=m}^{\infty} e^{-1/n} = e^{-\sum_{n=m}^{\infty} 1/n} = 0.$$

This contradiction implies that $X_n \overset{a.s.}{\not\to} X$.

We end this lecture with several examples of the importance (at least theoretical) of Theorems 1.3 and 1.4.

Example 2: Consider flipping a fair coin n times, all coin flips being mutually independent. In the previous lecture we used several quantitative tools to devise upper bounds on the probability that the number of coin flips whose outcome is heads is far from $n/2$. Theorems 1.3 and 1.4 are qualitative results, in particular, they cannot be used to devise such upper bounds for any given n. Instead, they assert that if we flip the coin enough times, then with very high probability, the number of coin flips whose outcome is heads will be very close to $n/2$. Formally, it follows from the weak law of large numbers that for every real numbers $\varepsilon > 0$ and $\delta > 0$, there exists an integer n_0 such that for every $n \geq n_0$, if we flip the coin n times, then with probability at least $1 - \delta$ the number of coin flips whose outcome is heads will be at least $(1/2 - \varepsilon)n$ and at most $(1/2 + \varepsilon)n$. This demonstrates the power of probability theory in describing natural occurrences. It also demonstrates a counter-intuitive but ubiquitous phenomenon in mathematics: while the outcome of one coin flip is a mystery, the outcome of many coin flips is (in some sense) predictable. The more coin flips we have, the more certain we are that the number of coin flips whose outcome is heads will be very close to half the total number of coin flips.

Example 3: Imagine that we would like to devise a way to approximate π up to some predetermined precision. Consider the following experiment. Let $X_1, Y_1, X_2, Y_2, \ldots$ be mutually independent random variables such that X_i and Y_i are distributed uniformly over $[-1/2, 1/2]$ for every positive integer i (we did not yet cover random variables which are distributed over uncountable sets, but hopefully the uniform distribution is simple enough so that we understand, at least intuitively, the behaviour of the X_i's and Y_i's). For every positive integer i let Z_i be the indicator random variable for the event "$X_i^2 + Y_i^2 \leq 1/4$". Observe that for every positive integer i, the ordered pair (X_i, Y_i) is a point in the square $[-1/2, 1/2] \times [-1/2, 1/2]$, chosen uniformly at random, and $Z_i = 1$ if and only if the point (X_i, Y_i) is in the circle of radius $1/2$ which is centered at $(0, 0)$. Clearly, the area of this circle

is $\pi/4$ and the area of the square $[-1/2, 1/2] \times [-1/2, 1/2]$ is 1. Hence, for every positive integer i, it holds that

$$\mathbb{E}(Z_i) = \mathbb{P}(Z_i = 1) = \pi/4.$$

Note also that, for every positive integer i, it holds that

$$Var(Z_i) = \mathbb{E}(Z_i^2) - (\mathbb{E}(Z_i))^2 = \pi/4 - (\pi/4)^2,$$

in particular, $Var(Z_i)$ is finite. It thus follows from the strong law of large numbers that

$$\frac{Z_1 + \ldots + Z_n}{n} \xrightarrow{a.s.} \frac{\pi}{4}.$$

That is, for every $\varepsilon > 0$, there exists an integer n_0 such that for every $n \geq n_0$, if we sample n points of the square $[-1/2, 1/2] \times [-1/2, 1/2]$, count how many of them are in the circle of radius $1/2$ which is centered at $(0, 0)$, and multiply the result by $4/n$, then with probability 1, the resulting number is at least $\pi - \varepsilon$ and at most $\pi + \varepsilon$.