Advanced Probability: WEEK 7

1 Concentration inequalities

The weak law of large numbers states that \( S_n \) converges in probability to \( E(X_1) \), when \( S_n = X_1 + \ldots + X_n \) and the \( X \)'s are i.i.d. random variables. This is exactly saying that for every fixed \( t > 0 \),

\[
P(\{|S_n/n - E(X_1)| > t\}) \to n \to \infty 0.
\]

However, this law does not say anything about the speed of convergence to 0 of this probability. The answer to this question is provided by concentration inequalities.

1.1 Hoeffding’s inequality

**Theorem 1.1.** Let \((X_n, n \geq 1)\) be a sequence of i.i.d. and integrable random variables, defined on a common probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and such that \(|X_1(\omega) - E(X_1)| \leq 1\), for all \( \omega \in \Omega \). Let also \( S_n = X_1 + \ldots + X_n \). Then

\[
P(\{|S_n/n - E(X_1)| > t\}) \leq 2 \exp(-nt^2/2), \quad \forall t > 0, n \geq 1.
\]

Before proving this theorem, let us make a few observations.

**Remarks.** - From this result, we easily recover the strong law of large numbers. Indeed, for all \( t > 0 \), we have

\[
\sum_{n \geq 1} \mathbb{P}(\{|S_n/n - E(X_1)| > t\}) \leq 2 \sum_{n \geq 1} \exp(-nt^2/2) < \infty,
\]

which implies by the Borel-Cantelli lemma that

\[
P(\{|S_n/n - E(X_1)| > t \text{ infinitely often}\}) = 0.
\]

This therefore says that \( S_n/n \to_{n \to \infty} E(X_1) \) almost surely.

- Note the universal character of this result (as already observed for the central limit theorem): the upper bound on the probability does not depend on the distribution of the \( X \)'s (except for the fact that these are bounded random variables by assumption).

- Note also the following: replacing \( t \) by \( u\sqrt{n} \) in the above statement, we obtain

\[
P(\{|S_n/n - E(X_1)| > u\sqrt{n}\}) \leq 2 \exp(-u^2/2).
\]

Recalling that the cdf of a \( \mathcal{N}(0,1) \) random variable behaves as

\[
F(u) \sim 1 - \exp\left(-\frac{u^2}{2}\right), \quad \text{when } u \text{ is large},
\]

we observe here another analogy with the central limit theorem.
The following lemma is the key to the proof of Theorem 1.1.

**Lemma 1.2.** Let $Z$ be a random variable such that $|Z(\omega)| \leq 1$ for all $\omega \in \Omega$ and $\mathbb{E}(Z) = 0$. Then

$$\mathbb{E}(e^{sz}) \leq \exp\left(\frac{s^2}{2}\right), \quad \forall s \in \mathbb{R}.$$ 

**Proof.** First observe that the mapping $z \mapsto e^{sz}$ is convex for any $s \in \mathbb{R}$, so for any $z \in [-1, 1]$, we have

$$e^{sz} \leq e^s \left(\frac{1+z}{2}\right) + e^{-s} \left(\frac{1-z}{2}\right) = \frac{e^s + e^{-s} - s}{2} = \cosh(s) + z \sinh(s).$$

Therefore, as $|Z(\omega)| \leq 1$ for all $\omega \in \Omega$ and $\mathbb{E}(Z) = 0$, we obtain

$$\mathbb{E}(e^{sz}) \leq \cosh(s) + \mathbb{E}(Z) \sinh(s) = \cosh(s).$$

The rest of the proof is calculus. Let $f(s) = \log \cosh(s)$; then

$$f'(s) = \frac{\sinh(s)}{\cosh(s)} = \tanh(s) \quad \text{and} \quad f''(s) = 1 - (\tanh(s))^2 \leq 1.$$

So for $s \geq 0$, $f'(s) = f'(0) + \int_0^s f''(t) \, dt \leq 0 + \int_0^s \, dt = s$. Similarly,

$$f(s) = f(0) + \int_0^s f'(t) \, dt \leq 0 + \int_0^s t \, dt = \frac{s^2}{2}.$$

This implies that $\cosh(s) \leq \exp\left(\frac{s^2}{2}\right)$. The same reasoning can be applied to the case $s \leq 0$, leading to the same conclusion. This proves the lemma. \(\square\)

**Proof of Theorem 1.1.** Let us compute

$$P \left( \left\{ \frac{S_n}{n} - \mathbb{E}(X_1) \right\} > t \right) = P \left( \{|S_n - n \mathbb{E}(X_1)| > nt\} \right) = P \left( \{|S_n - \mathbb{E}(S_n)| > nt\} \right) + P \left( \{|S_n - \mathbb{E}(S_n)| < -nt\} \right).$$

Let us focus on the first term, as the second can be handled exactly in the same way. By Chebychev’s inequality (using $\varphi(x) = e^{sx}$ with $s \geq 0$), we obtain

$$P \left( \{|S_n - \mathbb{E}(S_n)| > nt\} \right) \leq \frac{\mathbb{E}(e^{s(S_n - \mathbb{E}(S_n))})}{e^{nts}} = e^{-nts} \mathbb{E} \left( \prod_{j=1}^n e^{s(X_j - \mathbb{E}(X_j))} \right).$$

By the assumptions made, the random variable $Z = X_1 - \mathbb{E}(X_1)$ satisfies the assumptions of Lemma 1.2, so $\mathbb{E}(e^{s(X_1 - \mathbb{E}(X_1))}) \leq e^{s^2/2}$. This implies finally that

$$P \left( \{|S_n - \mathbb{E}(S_n)| > nt\} \right) \leq e^{-nts + ns^2/2} = e^{nts^2-st}.$$

As $s \geq 0$ is a free parameter, we choose it so as to minimize the term inside the parentheses, namely we choose $s^* = t$. This gives $P \left( \{|S_n - \mathbb{E}(S_n)| > nt\} \right) \leq e^{-nt^2/2}$. As already mentioned, a similar reasoning gives the same upper bound on the second term in (1), and this concludes the proof. \(\square\)

**Generalization.** (This is actually Hoeffding’s original statement!)

Let $(X_n, n \geq 1)$ be a sequence of independent and integrable random variables (so not necessarily i.i.d.) such that $X_n(\omega) \in [a_n, b_n]$ for all $n \geq 1$ and $\omega \in \Omega$. Let also $S_n = X_1 + \ldots + X_n$. Then

$$P \left( \{|S_n - \mathbb{E}(S_n)| > nt\} \right) \leq 2 \exp \left( -\frac{2n^2t^2}{\sum_{j=1}^n (b_j - a_j)^2} \right), \quad \forall t > 0, n \geq 1.$$ 

The proof is strictly speaking the same as above, but note that in this general case, $\frac{S_n}{n}$ need not converge to a limit as $n \to \infty$.  

2
1.2 Large deviations principle

Large deviations estimates lead to a refinement of Hoeffding’s inequality. Rather than stating the result from the beginning, let us discover it together!

Let \((X_n, n \geq 1)\) be a sequence of i.i.d. random variables defined on a common probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and such that \(\mathbb{E}(e^{sX_1}) < \infty\) for all \(|s| < s_0\), for some \(s_0 \in \mathbb{R}\). Let also \(t > \mathbb{E}(X_1)\) and \(S_n = X_1 + \ldots + X_n\). Using then again Chebychev’s inequality (with \(\varphi(x) = e^{sx}\) and \(s \geq 0\)), we obtain

\[
\mathbb{P}(|S_n| > nt) \leq \frac{\mathbb{E}(e^{sS_n})}{e^{n}\mathbb{E}(e^{sX_1})^n} = e^{-nst} \exp(n \log \mathbb{E}(e^{sX_1})) = \exp(-n(st - \log \mathbb{E}(e^{sX_1})))
\]

Optimizing this upper bound over \(s \geq 0\), we obtain

\[
\mathbb{P}(|S_n| > nt) \leq \exp(-n \max_{s \geq 0} (st - \log \mathbb{E}(e^{sX_1}))), \quad \forall t > \mathbb{E}(X_1).
\]

Let us make a slightly technical observation at this point. First, the function \(st - \log \mathbb{E}(e^{sX_1})\) takes the value 0 in \(s = 0\), so the above maximum is greater than or equal to 0. Second, for all \(s < 0\), we obtain, using Jensen’s inequality:

\[
st - \log \mathbb{E}(e^{sX_1}) \leq s(t - \mathbb{E}(X_1)) < 0,
\]

as \(s < 0\) and \(t - \mathbb{E}(X_1) > 0\) by assumption. In the above inequality, we may therefore replace the maximum over \(s \geq 0\) by the maximum over all \(s \in \mathbb{R}\), leading to:

\[
\mathbb{P}(|S_n| > nt) \leq \exp(-n \max_{s \in \mathbb{R}} (st - \log \mathbb{E}(e^{sX_1}))), \quad \forall t > \mathbb{E}(X_1).
\]

Let us now define the function \(\Lambda(s) = \log \mathbb{E}(e^{sX_1})\), for \(s \in \mathbb{R}\). This function might take the value \(+\infty\) for some values of \(s\) above \(s_0\), but this is not a problem here.

Let us also define what is called the Legendre transform of \(\Lambda\): \(\Lambda^*(t) = \max_{s \in \mathbb{R}} (st - \Lambda(s))\). It is a non-negative and convex function, which of course depends on the distribution of \(X_1\). By the above inequality, we have:

\[
\mathbb{P}(|S_n| > nt) \leq \exp(-n \Lambda^*(t)), \quad \forall t > \mathbb{E}(X_1).
\]

This is our first large deviations estimate, which is more precise than Hoeffding’s inequality. This is normal, as we take into account here the specificity of the distribution; we are not after a universal upper bound. Note also that the only inequality in the above derivation comes from the use of Chebychev’s inequality at the beginning. All the rest are equalities. Moreover, we optimize our choice over a large set of functions \((\varphi(x) = e^{sx})\) while using Chebychev’s inequality, so this upper bound is hopefully tight.

Likewise, for \(t < \mathbb{E}(X_1)\), we obtain for \(s \geq 0\):

\[
\mathbb{P}(|S_n| < nt) = \mathbb{P}(-S_n > nt) \leq \mathbb{E}(e^{-sS_n})e^{-nst} \mathbb{E}(e^{-sX_1})^n = e^{nst} \exp(n \log \mathbb{E}(e^{-sX_1})) = \exp(-n(st - \log \mathbb{E}(e^{-sX_1}))).
\]

Optimizing over \(s \geq 0\), we further obtain

\[
\mathbb{P}(|S_n| < nt) \leq \exp(-n \max_{s \geq 0} (st - \log \mathbb{E}(e^{-sX_1}))), \quad \forall t < \mathbb{E}(X_1).
\]

and for similar reasons as before, the maximum can be turned into a maximum over \(\mathbb{R}\), so that

\[
\mathbb{P}(|S_n| < nt) \leq \exp(-n \max_{s \in \mathbb{R}} (st - \log \mathbb{E}(e^{sX_1}))) = \exp(-n \Lambda^*(t)), \quad \forall t < \mathbb{E}(X_1).
\]

\footnote{One can show that this condition is equivalent to saying that there exists \(c > 0\) such that \(\mathbb{P}(|X_1| > x) \leq \exp(-cx)\) as \(x \to \infty\).}
What do these two equations (2) and (3) actually mean?

One can check that \( \Lambda^*(t) = 0 \) if and only if \( t = \mathbb{E}(X_1) \), so we see that in both cases (\( t > \mathbb{E}(X_1) \) and \( t < \mathbb{E}(X_1) \)), the upper bound on the probability is decreasing exponentially in \( n \), as it was the case with Hoeffding’s inequality. What changes here is the multiplicative factor \( \Lambda^*(t) \) which differs from (and is generally larger than) \( t^2/2 \), as we will see in the examples below.

**Generalization.** Before that, let us mention the generalization of the above results, also known as Cramér’s theorem. Let \( A \) be a “nice” subset of \( \mathbb{R} \) (think e.g. of an interval). Then

\[
\Pr \left( \left\{ \frac{S_n}{n} \in A \right\} \right) \simeq \exp \left( -n \inf_{t \in A} \Lambda^*(t) \right).
\]

Note therefore that the above probability is decreasing exponentially in \( n \) if and only if \( \mathbb{E}(X_1) \notin A \).

In the particular cases where \( A \) is either the interval \( ]-\infty, t[ \) with \( t < \mathbb{E}(X_1) \), or the interval \( [t, +\infty[ \) with \( t > \mathbb{E}(X_1) \), one recovers the above equations (2) and (3). Indeed, one can check that the infimum of \( \Lambda^* \) on \( ]-\infty, t[ \) is achieved in \( t \) when \( t < \mathbb{E}(X_1) \), and likewise for the interval on the positive axis.

Finally, let us mention that Cramér’s theorem not only provides an upper bound on the probability, but also a corresponding lower bound which is matching the upper bound in some asymptotic sense. This is therefore a quite remarkable and complete result.

**Examples.** - Let \( X_1 \sim \mathcal{N}(0, 1) \). Note that \( X_1 \) is an unbounded random variable, so Hoeffding’s inequality does not apply here. Let us compute

\[
\Lambda(s) = \log \mathbb{E} \left( e^{sX_1} \right) = \log \left( \int_{\mathbb{R}} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \right) = \log \left( e^{s^2/2} \right) = \frac{s^2}{2},
\]

and

\[
\Lambda^*(t) = \max_{s \in \mathbb{R}} \left( st - \frac{s^2}{2} \right) = \frac{t^2}{2}, \quad \text{attained in } s^* = t.
\]

Also, \( \mathbb{E}(X_1) = 0 \), so \( \Pr(\{S_n \geq nt\}) \leq \exp \left( -\frac{nt^2}{2} \right) \), for all \( t > 0 \). Surprisingly perhaps, this gives exactly the same upper bound as the one derived by Hoeffding (even though the random variables \( X \)'s are unbounded here).

- Let \( X_1 \) be such that \( \Pr(\{X_1 = 1\}) = \Pr(\{X_1 = 0\}) = \frac{1}{2} \). In this case,

\[
\Lambda(s) = \log \left( \frac{e^{s} + 1}{2} \right) \quad \text{and} \quad \Lambda^*(t) = \max_{s \in \mathbb{R}} \left( st - \log \left( \frac{e^{s} + 1}{2} \right) \right).
\]

Looking for the value where the maximum is attained, we obtain \( s^* = \log \left( \frac{1}{t-1} \right) \) and correspondingly, after some computations:

\[
\Lambda^*(t) = t \log t + (1 - t) \log(1 - t) + 2.
\]

Also \( \mathbb{E}(X_1) = \frac{1}{2} \), so \( \Pr(\{S_n \geq nt\}) \leq \exp \left( -n \Lambda^*(t) \right) \), for all \( t > \frac{1}{2} \). Let us compare this result with Hoeffding’s inequality, which reads in this case (watch out that here, \( |X_1(\omega) - \mathbb{E}(X_1)| \leq \frac{1}{2} \) for all \( \omega \in \Omega \)):

\[
\Pr \left( \left\{ S_n - \frac{n}{2} > nu \right\} \right) \leq \exp \left( -2nu^2 \right), \quad \forall u > 0.
\]

Replacing \( u \) by \( t - \frac{1}{2} \), we obtain

\[
\Pr(\{S_n > nt\}) \leq \exp \left( -2n \left( t - \frac{1}{2} \right)^2 \right), \quad \forall t > \frac{1}{2}.
\]

It can be observed that for \( t > \frac{1}{2} \), the above function \( \Lambda^*(t) \) dominates the function \( 2 \left( t - \frac{1}{2} \right)^2 \) obtained via Hoeffding’s inequality.