36-705: Intermediate Statistics **Fall 2018**

Lecture 4: September 4

Lecturer: Siva Balakrishnan

We will first continue our discussion of exponential concentration inequalities.

4.1 Levy's inequality

There is a similar concentration inequality that applies to functions of Gaussian random variables that are sufficiently smooth. In this case, the assumption is quite different. We assume that:

$$|f(X_1, \ldots, X_n) - f(Y_1, \ldots, Y_n)| \leq L\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2},$$

for all $X_1, \ldots, X_n, Y_1, \ldots, Y_n \in \mathbb{R}$.

For such functions we have that if $X_1, \ldots, X_n \sim N(0,1)$ then,

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq t) \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

4.2 χ^2 tail bounds

A χ^2 random variable with n degrees of freedom, denoted by $Y \sim \chi^2_n$, is a RV that is a sum of n i.i.d. standard Gaussian RVs, i.e. $Y = \sum_{i=1}^{n} X_i^2$ where each $X_i \sim N(0,1)$. Suppose that $Z_1, \ldots, Z_n \sim N(0,1)$, then the expected value $\mathbb{E}[Z_i^2] = 1$, and we have the χ^2 tail bound:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} Z_k^2 - 1\right| \geq t\right) \leq 2\exp(-nt^2/8) \quad \text{for all} \quad t \in (0,1).$$

You will derive this in your HW using the Chernoff method. Analogous to the class of sub-Gaussian RVs, χ^2 random variables belong to a class of what are known as *sub-exponential* random variables. The main note-worthy difference is that the Gaussian-type behaviour of the tail only holds for small values of the deviation t.

Detour: The union bound. This is also known as Boole's inequality. It says that if we have events $A_1, \ldots, A_n$ then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

In particular, if we consider a case when each event A_i is a failure of some type, then the above inequality says that the probability that even a single failure occurs is at most the sum of the probabilities of each failure.

Example: The Johnson-Lindenstrauss Lemma. One very nice application of χ^2 tail bounds is in the analysis of what are known as "random projections". Suppose we have a data set $X_1, \ldots, X_n \in \mathbb{R}^d$ where d is quite large. Storing such a dataset might be expensive and as a result we often resort to "sketching" or "random projection" where the goal is to create a map $F : \mathbb{R}^d \mapsto \mathbb{R}^m$, with $m \ll d$. We then instead store the mapped dataset $\{F(X_1), \ldots, F(X_n)\}$. The challenge is to design this map F in a way that preserves essential features of the original dataset. In particular, we would like that for every pair (X_i, X_j) we have that,

$$(1-\epsilon)\|X_i - X_j\|_2^2 \leq \|F(X_i) - F(X_j)\|_2^2 \leq (1+\epsilon)\|X_i - X_j\|_2^2,$$

i.e. the map preserves all the pair-wise distances up to a $(1 \pm \epsilon)$ factor. Of course, if m is large we might expect this is not too difficult.

The Johnson-Lindenstrauss lemma is quite stunning: it says that a simple randomized construction will produce such a map with probability at least $1 - \delta$ provided that,

$$m \geq \frac{16 \log(n/\delta)}{\epsilon^2}.$$

Notice that this is completely independent of the original dimension d and depends on logarithmically on the number of points n. This map can result in huge savings in storage cost while still essentially preserving all the pairwise distances.

The map itself is quite simple: we construct a matrix $Z \in \mathbb{R}^{m \times d}$, where each entry of Z is i.i.d $N(0,1)$. We then define the map as:

$$F(X_i) = \frac{ZX_i}{\sqrt{m}}.$$

Now let us fix a pair (X_j, X_k) and consider,

$$\begin{aligned}\frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} &= \left\| \frac{Z(X_j - X_k)}{\sqrt{m}\|X_j - X_k\|_2} \right\|_2^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} \underbrace{\langle Z_i, \frac{X_j - X_k}{\|X_j - X_k\|_2} \rangle^2}_{T_i}.\end{aligned}$$

Now, for some fixed numbers a_j the distribution of $\sum_{j=1}^d a_j Z_{ij}$ is Gaussian with mean 0 and variance $\sum_{j=1}^d a_j^2$. So each term T_i is an independent χ^2 random variable. Now applying the χ^2 tail bound, we obtain that,

$$\mathbb{P}\left(\left| \frac{\|F(X_j) - F(X_k)\|_2^2}{\|X_j - X_k\|_2^2} - 1 \right| \geq \epsilon \right) \leq 2\exp(-m\epsilon^2/8).$$

Thus for the fixed pair (X_i, X_j) the probability that our map fails to preserve the distance is exponentially small, i.e. is at most $2\exp(-m\epsilon^2/8)$. Now, to find the probability that our map fails to preserve *any* of our $\binom{n}{2}$ pairwise distances we simply apply the union bound to conclude that, the probability of any failure is at most:

$$\mathbb{P}(\text{failure}) \leq 2\binom{n}{2}\exp(-m\epsilon^2/8).$$

Now, it is straightforward to verify that if

$$m \geq \frac{16\log(n/\delta)}{\epsilon^2},$$

then this probability is at most δ as desired. An important point to note is that the *exponential concentration* is what leads to such a small value for m (i.e. it only needs to grow logarithmically with the sample size).

In the rest of this lecture we discuss the convergence of random variables. At a high-level, our first few lectures focused on non-asymptotic properties of averages i.e. the tail bounds we derived applied for any fixed sample size n. For the next few lectures we focus on asymptotic properties, i.e. we ask the question: what happens to the average of n i.i.d. random variables as $n \to \infty$.

Roughly, from a theoretical perspective the idea is that many expressions will considerably simplify in the asymptotic regime. Rather than have many different tail bounds, we will derive simple "universal results" that hold under extremely weak conditions.

From a slightly more practical perspective, asymptotic theory is often useful to obtain approximate confidence intervals (and p-values and other useful things) that although approximate are typically more useful. We will follow quite closely Section 5.5 of Casella and Berger.

4.3 Reminder: convergence of sequences

When we think of convergence of deterministic real numbers the corresponding notions are classical.

Formally, we say that a sequence of real numbers $a_1, a_2, \ldots$ converges to a fixed real number a if, for every positive number ϵ, there exists a natural number $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, $|a_n - a| < \epsilon$. We call a the limit of the sequence and write $\lim_{n\to\infty} a_n = a$.

Our focus today will in trying to develop analogues of this notion that apply to sequences of random variables. We will first give some definitions and then try to circle back to relate the definitions and discuss some examples.

Throughout, we will focus on the setting where we have a sequence of random variables $X_1, \ldots, X_n$ and another random variable X, and would like to define what is means for the sequence to converge to X. In each case, to simplify things you should also think about the case when X is deterministic, i.e. when $X = c$ with probability 1 (for some constant c).

Importantly, we will *not assume that the RVs $X_1, \ldots, X_n$ are independent.*

4.4 Almost sure convergence

We will not use almost sure convergence in this course so you should feel free to ignore this section. A natural analogue of the usual convergence would be to hope that,

$$\lim_{n\to\infty} X_n = X.$$

These are both however random variables so one has to at least specify on what event we are hoping for this statement to be true.

The correct analogue turns out to be to require:

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

There are measure theoretic subtleties to be aware of here. In particular, the sample space inside the probability statement here grows with n and it requires some machinery to be precise here.

There are other equivalent (this is somewhat difficult to see) ways to define almost sure convergence. Equivalently, we say that X_n converges almost surely to X if we let Ω be a set of probability mass 1, i.e. $\mathbb{P}(\Omega) = 1$, and for every $\omega \subseteq \Omega$, and for every $\epsilon > 0$, we have that there is some $n \geq N(\omega, \epsilon)$ such that:

$$|X_n(\omega) - X(\omega)| \leq \epsilon.$$

Roughly, the way to think about this type of convergence is to imagine that there is some set of exceptional events on which the random variables can disagree, but these exceptional events have probability 0 as $n \to \infty$. Barring, these exceptional events the sequence converges just like sequences of real numbers do. The exceptional events is where the "almost" in almost sure arises.

4.5 Convergence in probability

A sequence of random variables $X_1, \ldots, X_n$ converges in probability to a random variable X if for every $\epsilon > 0$ we have that,

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

To build intuition it is perhaps useful to consider the case when X is deterministic, i.e. $X = c$ with probability 1. Then convergence in probability is saying that as n gets large the distribution of X_n gets more peaked around the value c.

Again somewhat roughly, convergence in probability can be viewed as a statement about the convergence of probabilities, while almost sure convergence is a convergence of the values of a sequence of random variables.

We will not prove this statement but convergence in probability is implied by almost sure convergence. The notes contain a counterexample to the reverse implication but we most likely will not cover this in lecture.

Example: Weak Law of Large Numbers Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}(Y_i) = \sigma^2 < \infty$. Define, for $i \in \{1, \ldots, n\}$,

$$X_i = \frac{1}{i}\sum_{j=1}^{i} Y_j.$$

The WLLN says that the sequence $X_1, X_2, \ldots$ converges in probability to μ.

Proof: The proof is simply an application of Chebyshev's inequality. We note that by Chebyshev's inequality:

$$\mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

This in turn implies that,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) = 0,$$

as desired.

Notes:

1. Strictly speaking the WLLN is true even without the assumption of finite variance, as long as the first absolute moment is finite. This proof is a bit more difficult.
2. There is a statement that says that under similar assumptions the average converges almost surely to the expectation. This is known as the strong law of large numbers. This is actually quite a bit more difficult to prove.

Consistency: Convergence in probability will frequently recur in this course. Usually we will construct an estimator $\widehat{\theta}_n$ for some quantity θ^*. We will then say that the estimator is *consistent* if the sequence of RVs $\widehat{\theta}_n$ converges in probability to θ^*.

The WLLN/Chebyshev can already be used to prove some rudimentary consistency guarantees. For instance, if we consider the sample variance:

$$\widehat{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu}_n)^2,$$

then by Chebyshev's inequality we obtain,

$$\mathbb{P}(|\widehat{S}_n - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{Var}(\widehat{S}_n)}{\epsilon^2},$$

so a sufficient condition for consistency is that $\mathrm{Var}(\widehat{S}_n) \to 0$ as $n \to \infty$.

Convergence in probability does not imply almost sure convergence: This example is from Casella and Berger. Suppose we have a sample space $S = [0, 1]$, with the uniform distribution, we draw $s \sim U[0, 1]$ and define $X(s) = s$.

We define the sequence as:

$$X_1(s) = s + \mathbb{I}_{[0,1]}(s), \quad X_2(s) = s + \mathbb{I}_{[0,1/2]}(s), \quad X_3(s) = s + \mathbb{I}_{[1/2,1]}(s)$$
$$X_4(s) = s + \mathbb{I}_{[0,1/3]}(s), \quad X_5(s) = s + \mathbb{I}_{[1/3,2/3]}(s), \quad X_6(s) = s + \mathbb{I}_{[2/3,1]}(s).$$

Now one can check that this sequence converges in probability but not almost surely. Roughly, the "1 + s" spike becomes less frequent down the sequence (allowing convergence in probability) but the limit is not well defined. For any s, $X_n(s)$ alternates between 1 and $1 + s$.

4.6 Convergence in quadratic mean

An often useful way to show convergence in probability is to show something stronger known as convergence in quadratic mean. We say that a sequence converges to X in quadratic mean if:

$$\mathbb{E}(X_n - X)^2 \to 0,$$

as $n \to \infty$. We will return to this one when we discuss some examples.

4.7 Convergence in distribution

The other commonly encountered mode of convergence is convergence in distribution. We say that a sequence converges to X in distribution if:

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t),$$

for all points t where the CDF F_X is continuous. We will see why the exception matters in a little while but for now it is worth noting that convergence in distribution is the weakest form of convergence.

For instance, a sequence of i.i.d. $N(0,1)$ RVs converge in distribution to an independent $N(0,1)$ RV, even though the values of the random variables are not close in any meaningful sense (their distributions are however, identical). A famous example that we will spend a chunk of the next lecture on is the central limit theorem. The central limit theorem says that an average of i.i.d. random variables (appropriately normalized) converges in distribution to a $N(0,1)$ random variable.

The picture to keep in mind to understand the relationships is the following one:

$$\begin{array}{ccccc} & & \text{q.m.} & & \\ & & \downarrow & & \\ \text{a.s.} & \rightarrow & \text{prob} & \rightarrow & \text{distribution} \end{array}$$

We will re-visit this in the next lecture and perhaps try to prove some of the implications (or disprove some of the non-implications).

4.8 More Examples

Example 1: Suppose we consider a sequence $X_n = N(0, 1/n)$. Intuitively, it seems like this sequence converges to 0. Let us first consider what happens in distribution.

The CDF of the RV that is deterministically 0 is simply $F_X(x) = 0$, for $x < 0$ and $F_X(x) = 1$ for $x \geq 0$. Now, let us consider,

$$F_{X_n}(x) = \mathbb{P}(Z \leq \sqrt{n}x),$$

where $Z \sim N(0,1)$. If $x > 0$ this tends to 1, and if $x < 0$ this tends to 0. Interestingly, at $x = 0$, $F_{X_n}(x) = 1/2$, and does not converge to $F_X(0) = 1$. Remember, however, that we had an exception at points of discontinuity.

Example 2: Let us consider the same example and consider convergence in probability.

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \frac{\mathbb{E}[X_n^2]}{\epsilon^2} = \frac{1}{n\epsilon^2} \to 0,$$

so the sequence converges to 0 in probability.

Example 3: Let us consider another example from the Casella and Berger book. Suppose $X_1, \ldots \sim U[0,1]$. Let us define $X_{(n)} = \max_{1 \le i \le n} X_i$. Now, we verify two things:

1. $X_{(n)}$ converges in probability to 1. To see this observe that,

$$\begin{aligned}\mathbb{P}(|X_{(n)} - 1| \ge \epsilon) &= \mathbb{P}(X_{(n)} \le 1 - \epsilon) \\ &= \prod_{i=1}^{n} \mathbb{P}(X_i \le 1 - \epsilon) = (1-\epsilon)^n \\ &\to 0.\end{aligned}$$

2. The random variable $n(1 - X_{(n)})$ converges in distribution to an Exp(1) RV. To see this we compute:

$$\begin{aligned}F_{X_{(n)}}(t) &= \mathbb{P}(n(1 - X_{(n)}) \le t) = 1 - \mathbb{P}(X_{(n)} \le 1 - t/n) \\ &= 1 - (1 - t/n)^n \to 1 - \exp(-t) = F_X(t).\end{aligned}$$