36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 6: September 12

Lecturer: Siva Balakrishnan

6.1 Review and Outline

Last class we saw:

- Markov's inequality
- Chebyshev's inequality
- Exponential concentration (Mill's inequality)

This class we will state another famous exponential concentration inequality, prove the weak law of large numbers, and then talk about O_p notation and then talk about convergence of random variables.

6.2 Hoeffding's inequality

The main drawback was that Mill's inequality only applies to Gaussian random variables. Another commonly useful exponential concentration inequality applies to bounded random variables. This is called Hoeffding's inequality.

Hoeffding's inequality: Suppose that $X_1, \ldots, X_n$ are independent and that, $a_i \leq X_i \leq b_i$, and $\mathbb{E}[X_i] = 0$. Then for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

We will not prove this one, but Wasserman's book has a proof if you are curious.

Hoeffding's inequality looks a bit different from the other inequalities we have seen yesterday, but let us rearrange it a bit. Equivalently,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\sqrt{\frac{\sum_{i=1}^{n}(b_i - a_i)^2}{n^2}}\right) \leq 2\exp\left(-2t^2\right).$$

This is more like the earlier inequalities, but notice that we don't really have the standard deviation any more. That said, if $a_i \leq X_i \leq b_i$ then $\mathrm{Var}(X_i) \leq (b_i - a_i)^2$.

Exercise: Prove the above fact.

So that:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \leq \frac{\sum_{i=1}^{n}(b_i - a_i)^2}{n^2},$$

and this will allow us to interpret Hoeffding's inequality in a more familiar way. Roughly, it says that the probability that the sample average is more than t standard deviations from its expectation is at most $\exp(-2t^2)$.

Let us now use Hoeffding's inequality in our case study example of coin tosses. There each random variable is between -1 and 1 so we have that by Hoeffding's inequality:

$$\mathbb{P}\left(|Y| \geq \frac{2t}{\sqrt{n}}\right) \leq 2\exp(-2t^2).$$

Observe once again this inequality is similar to Chebyshev's inequality on the left hand side, i.e. the deviation is on the order of $1/\sqrt{n}$ but the right hand side is much smaller $\exp(-t^2)$ instead of $1/t^2$.

6.3 Sample Sizes and Exponential Concentration

Lets just do some basic calculations to get used to concentration inequalities. I am sampling $X_1, \ldots, X_n$ which are each i.i.d. 0 or $+1$ with some (unknown) probabilities. In order to estimate the probability of 1 (i.e. the expected value of X), I use the estimate:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The variance of $\hat{\mu}$ is

$$\mathrm{Var}(\hat{\mu}) = \frac{\mu(1-\mu)}{n} \leq \frac{1}{4n}.$$

Suppose I want to be 95% sure that my sample average is within 0.01 of the true average, how many samples do I need?

1. If I only had Chebyshev's inequality: Recall,

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t/(2\sqrt{n})) \leq \frac{1}{t^2},$$

so that we need:

$$\frac{t}{2\sqrt{n}} \leq 0.01$$
$$\frac{1}{t^2} \leq 0.05.$$

In particular, we would choose $t = \sqrt{1/0.05} \approx 4.47$, and conclude that I need $(\frac{t}{0.02})^2 \approx 50000$ samples.

2. If I instead had Hoeffding's inequality: Recall that by Hoeffding's inequality we would have:

$$\mathbb{P}(|\hat{\mu} - \mu| \geq t/\sqrt{n}) \leq 2\exp(-t^2/2),$$

so that we need:

$$\frac{t}{\sqrt{n}} \leq 0.01$$
$$2\exp(-t^2/2) \leq 0.05.$$

In this case, we would use:

$$t = \sqrt{-\frac{\ln 0.025}{2}} \approx 1.36,$$

and need

$$n \geq 18500,$$

samples. Roughly a 3-fold reduction. Of course the difference can be even more stark if we change the requirements.

There is a slightly different take away here, which is that one can see that small imprecisions (using an upper bound on the variance instead of the exact variance or having only approximately tight concentration inequalities) can have huge impacts on our sample size requirements. This is sometimes troubling from a practical perspective. It is this fact that often motivates large sample approximations - where we assume $n \to \infty$ and then derive very precise results about the distribution of various estimators. In essence, finite-sample tail bounds can be loose but are always correct, large-sample theory can be much tighter but is only approximately correct for finite sample-sizes.

6.3.1 Confidence Intervals

We will cover this in much more detail later on but here is a quick preview. Hoeffding's inequality gives us a simple way to create a confidence interval for a binomial parameter p. Fix $\alpha > 0$ and let,

$$t = \sqrt{\frac{1}{2n}\log(2/\alpha)}.$$

By Hoeffding's inequality,

$$\mathbb{P}(|\hat{\mu} - p| \geq t) \leq 2\exp(-2nt^2) = \alpha.$$

Let $C = (\hat{\mu} - t, \hat{\mu} + t)$. Then,

$$\mathbb{P}(p \notin C) \leq \alpha.$$

Hence, the random interval C traps the true parameter value p with probability at least $1 - \alpha$; we call C a $1 - \alpha$ confidence interval.

6.4 The Weak Law of Large Numbers

The weak law of large numbers essentially assures us that the average of independent and identically distributed random variables "converges" to the expectation.

We will assume that $X_1, \ldots, X_n$ are i.i.d with $\text{Var}(X) < \infty$. Then the weak law of large numbers says that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| \geq \epsilon\right) \to 0,$$

as $n \to \infty$.

This type of convergence is quite common and has a name: it is called convergence in probability. So the weak LLN tells us that the sample average converges to the population average in probability. This is called the weak law because convergence in probability is often referred to as weak convergence. We will discuss convergence more systematically soon.

Proof: The proof is a simple consequence of Chebyshev's inequality. We have that for any positive ϵ, via Chebyshev's inequality:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| \geq \epsilon\right) &\leq \frac{\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)}{\epsilon^2} \\ &= \frac{\text{Var}(X)}{n\epsilon^2} \to 0,\end{aligned}$$

as $n \to \infty$.

6.5 Introduction to Stochastic Convergence

The weak LLN is an example where we tried to reason about the limiting behaviour of a sequence of random variables:

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In statistics we commonly estimate parameters using data, and in this case our estimate forms a sequence of random variables (as we collect more data). We would like to reason about the limiting behaviour of our estimates (do they "converge" to the truth, what is their distribution around the truth and so on). This is what we refer to as **large sample theory.**

Suppose we have a sequence of random variables $X_1, \ldots, X_n$, and another random variable X. Let F_n denote the CDF of X_n, and let F be the CDF of X.

The two most basic forms of stochastic convergence are:

1. Convergence in Probability: We say the sequence converges to X if for every $\epsilon > 0$,

   $$\mathbb{P}(|X_n - X| \geq \epsilon) \to 0,$$

   as $n \to \infty$. An important example is the weak law.

2. Convergence in Distribution: The sequence converges in distribution to X if

   $$\lim_{n\to\infty} F_n(t) = F(t),$$

   at all points t where F is continuous. An important example of this type of convergence is the central limit theorem. We will see this in much more detail but roughly the central limit theorem says that the average of i.i.d. RVs, rescaled appropriately converges in distribution to a standard normal distribution, i.e. that

   $$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \to N(0, 1).$$

   If you think about it, this is a really stunning result. You do not assume anything about the RVs except their first two moments need to exist, and you conclude that the average approaches a Gaussian distribution.