LECTURE NOTES 7

1 Stochastic convergence

Stochastic convergence refers to convergence of sequences of random variables. Recall from the last notes that we are concerned with two types of convergence in probability and convergence in distribution. Convergence in probability implies convergence in distribution but the reverse is not, in general, true.

Let $Y_n \sim N(0, n^{-1})$ for $n = 1, 2, \ldots$. Then $Y_n \xrightarrow{P} 0$. We also have that $Y_n \rightsquigarrow Z$ where Z is degenerate at 0, that is $P(Z = 0) = 1$. Also, $\sqrt{n}Y_n \rightsquigarrow N(0, 1)$. In fact, a stronger statement is true: $\sqrt{n}Y_n \stackrel{d}{=} N(0, 1)$ for all n.

Now suppose that $Y_n \sim N(n, 1)$. Then Y_n does not converge to anything.

We say that a sequence Y_n converges to Y in quadratic mean if:

$$\mathbb{E}(Y_n - Y)^2 \to 0,$$

as $n \to \infty$. This is once again a convergence of values of a sequence of random variables. In fact, convergence in quadratic mean $\implies$ convergence in probability since by Chebyshev's inequality we know that:

$$\mathbb{P}(|Y_n - Y| \geq \epsilon) \leq \frac{\mathbb{E}(Y_n - Y)^2}{\epsilon^2} \to 0,$$

as $n \to \infty$. Usually, we are concerned with convergence in quadratic mean to a constant c. This means that $\mathbb{E}(Y_n - c)^2 \to 0$. This implies that $Y_n \xrightarrow{P} c$.

We say that a sequence Y_n converges to Y in ℓ_1 if:

$$\mathbb{E}|Y_n - Y| \to 0,$$

as $n \to \infty$. Convergence in quadratic mean $\implies$ convergence in ℓ_1. To prove this we can just use the Cauchy-Schwarz inequality:

$$\mathbb{E}|Y_n - Y| \leq \sqrt{\mathbb{E}(Y_n - Y)^2} \to 0,$$

as $n \to \infty$.

We say that Y_n converges almost surely to c if

$$\mathbb{P}(\lim_{n\to\infty} Y_n = c) = 1.$$

This is stronger than convergence in probability.

2 The Central Limit Theorem (CLT)

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and

$$Z_n = \frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma}.$$

Note that $\mathbb{E}[Z_n] = 0$ and $\mathrm{Var}[Z_n] = 1$.

Theorem 1 *Z_n converges in distribution to a standard Gaussian. That is, $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$.*

We can use the CLT to approximate probability calcuations. For example:

$$\begin{aligned} P(a \leq \overline{X} \leq b) &= P\left(\frac{\sqrt{n}(a-\mu)}{\sigma} \leq Z_n \leq \frac{\sqrt{n}(b-\mu)}{\sigma}\right) \\ &\approx P\left(\frac{\sqrt{n}(a-\mu)}{\sigma} \leq Z \leq \frac{\sqrt{n}(b-\mu)}{\sigma}\right) \\ &= \Phi\left(\frac{\sqrt{n}(b-\mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a-\mu)}{\sigma}\right) \end{aligned}$$

where Φ is the cdf of a standard Normal.

Theorem 2 *Let*

$$T_n = \frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{s}$$

where $s^2 = n^{-1}\sum_i (X_i - \overline{X}_n)^2$. Then $T_n \rightsquigarrow N(0,1)$.

(The theorem is also true if we define $s^2 = (n-1)^{-1}\sum_i (X_i - \overline{X}_n)^2$. We'll see later why we might want to do this.) It then follows that

$$P(a \leq \overline{X} \leq b) \approx \Phi\left(\frac{\sqrt{n}(b-\mu)}{s}\right) - \Phi\left(\frac{\sqrt{n}(a-\mu)}{s}\right).$$

If we define the distance between the CDF of the average, and the CDF of a Gaussian appropriately, we can ask how far the two CDFs are for a $f$inite sample size n. The answer is $C/\sqrt{n}$ for some constant C. These results are typically called Berry-Esseen bounds. They assure us that the convergence to normality can happen quite quickly in some important cases.

If we average a collection of independent random vectors then they will converge in distribution to a multivariate Gaussian.

Given that Y_n converges in distribution to a Gaussian, one can ask about functions of Y_n. Under some regularity conditions these also converge to a Gaussian, and the delta method tells us how to compute the mean and variance of the new Gaussian. In detail, If $Y_n \rightsquigarrow N(\mu, \sigma^2)$ and r is a smooth function, then $r(Y_n) \rightsquigarrow N(r(\mu), (r'(\mu))^2\sigma^2)$.

3 O_P and o_P

In statistics and machine learning, we make use of o_P and O_P notation.

Recall first, that $a_n = o(1)$ means that $a_n \to 0$ as $n \to \infty$. $a_n = o(b_n)$ means that $a_n/b_n = o(1)$. $a_n = O(1)$ means that a_n is eventually bounded, that is, for all large n, $|a_n| \leq C$ for some $C > 0$. $a_n = O(b_n)$ means that $a_n/b_n = O(1)$.

We write $a_n \sim b_n$ if both a_n/b_n and b_n/a_n are eventually bounded. In computer science this is written as $a_n = \Theta(b_n)$ but we prefer using $a_n \sim b_n$ since, in statistics, Θ often denotes something else.

Now we move on to the probabilistic versions. Say that $Y_n = o_P(1)$ if $Y_n \xrightarrow{P} 0$. Say that $Y_n = o_P(a_n)$ if, $Y_n/a_n = o_P(1)$.

Say that $Y_n = O_P(1)$ if, for every $\epsilon > 0$, there is a $C > 0$ such that

$$\mathbb{P}(|Y_n| > C) \leq \epsilon.$$

Say that $Y_n = O_P(a_n)$ if $Y_n/a_n = O_P(1)$.

Let's use Hoeffding's inequality to show that sample proportions are $O_P(1/\sqrt{n})$ within the the true mean. Let $Y_1, \ldots, Y_n$ be coin flips i.e. $Y_i \in \{0, 1\}$. Let $p = \mathbb{P}(Y_i = 1)$. Let

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

We will show that: $\widehat{p}_n - p = o_P(1)$ and $\widehat{p}_n - p = O_P(1/\sqrt{n})$.

We have that

$$\mathbb{P}(|\widehat{p}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2} \to 0$$

and so $\widehat{p}_n - p = o_P(1)$. Also,

$$\begin{aligned} \mathbb{P}(\sqrt{n}|\widehat{p}_n - p| > C) &= \mathbb{P}\left(|\widehat{p}_n - p| > \frac{C}{\sqrt{n}}\right) \\ &\leq 2e^{-2C^2} < \delta \end{aligned}$$

if we pick C large enough. Hence, $\sqrt{n}(\widehat{p}_n - p) = O_P(1)$ and so

$$\widehat{p}_n - p = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Make sure you can prove the following:

$$\begin{aligned}
O_P(1)o_P(1) &= o_P(1) \\
O_P(1)O_P(1) &= O_P(1) \\
o_P(1) + O_P(1) &= O_P(1) \\
O_P(a_n)o_P(b_n) &= o_P(a_n b_n) \\
O_P(a_n)O_P(b_n) &= O_P(a_n b_n)
\end{aligned}$$