ECSE 2500
Lec 23
April 24

Topic: Law of large number
Central limit theory

Recall in the lecture:

In the limit, for n iid. RVs,

$$\lim_{n\to\infty} Var(M_n) = 0,$$

$$\lim_{n\to\infty} E[M_n] = \mu.$$

□ Weak Law of large number

Q: How fast do we get these values?

A natural way to estimate the speed is through the Chebyshev inequality:

$$P(|M_n - E[M_n]| \geq \epsilon) \leq \frac{1}{\epsilon^2} Var(M_n)$$

$$\Leftrightarrow \quad P(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

or $$P(|M_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

Message: This lets us bound how many observations we need to ensure that our sample mean M_n is within ϵ distance of the actual mean μ.

Say $Var(X_i) = 5$ and we want to be 99% sure that our estimate is within 0.1 of the true mean.

Then we want

$$1 - \frac{\sigma^2}{n\epsilon^2} \geq 0.99$$

$$\Rightarrow \frac{5}{n \cdot (0.1)^2} \leq 0.01 \Rightarrow n \geq \frac{5}{(0.01)^2} = 50000$$

We need at least 50000 samples so that the sample mean M_n is close to the actual mean by 0.1.

Important)

Week law of large number:

If X_i are i.i.d., $E[X_i] = \mu < \infty$, $\epsilon > 0$,

$$\boxed{\lim_{n \to \infty} P(|M_n - \mu| < \epsilon) = 1}$$

Plain language: If we are given ϵ, we can take n large enough so that M_n is that close to the mean with high probability.

$\boxed{\text{Implication:}}$ If we have a biased coin with $P(\text{head}) = p$, then we can this as a Bernoulli RV with $P(0) = 1-p$, $P(1) = p$. If we can compute M_n, then WLLN says

$$\lim_{n\to\infty} P\left(\left|\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{M_n} - p\right| < \epsilon\right) = 1$$

Averge # of success gets very close to p as n gets very large.

$\boxed{\text{Limitation of WLLN:}}$ The chebyshev inequality we use is very conservative, so the # of suggested samples/trials may be an overestimate.

Therefore, to overcome this limitation, there is also a more powerful version of WLLN, which is

(SLLN): Strong Law of large number

If X_i is a sequence of iid RVs with

$$E[X_i] = \mu < \infty, \quad Var(X_i) = \sigma^2 < \infty$$

then $$P\left(\lim_{n\to\infty} M_n = \mu\right) = 1.$$

The difference between WLLN and SLLN is a bit subtle. It basically says that any sequence of sample means must eventually converge to the actual mean.

Central limit theorem

The WLLN and SLLN talk about how the sample mean converges to the actual mean.

The central limit theorem is stronger: It talks about how the PDF/CDF of the sample mean M_n converges. The result is that no matter the incoming distribution of X_i, the limiting distribution of M_n is (in certain sense)

Gaussian !

Recall $S_n = \sum_{i=1}^{n} X_i$; $E[S_n] = n\mu$, $Var(S_n) = n\sigma^2$

$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$; $E[M_n] = \mu$, $Var(M_n) = \frac{\sigma^2}{n}$

If we define a new RV Z_n :

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

$$E[Z_n] = \frac{1}{\sigma\sqrt{n}}\left(E[S_n] - n\mu\right) = 0$$

$$Var(Z_n) = \frac{1}{\sigma^2 n} Var(S_n) = 1$$

At each value n, Z_n is a zero-mean, unit variance random variable.

$$\boxed{\text{The Central limit theorem states that } \lim_{n\to\infty} P(Z_n \le z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{x^2}{2}}\,dx}$$

$$\Rightarrow \lim_{n\to\infty} F_{Z_n}(z) = \phi(z) = 1 - Q(z)$$

CDF of standard Gaussian RV (mean 0, variance 1)

That is, the CDF of Z_n converges to the Gaussian CDF for any value z, as $n \to \infty$.

To rigorously prove CLT, we need more advanced tools in probability, but you can easily use numerical software to verify this result.

$\boxed{\text{Implication of CLT}}$ - When n is big enough, we can use CLT to get a good approximation of the probabilities of events by creating a corresponding Gaussian RV and looking at the Q table.

More specifically, CLT says that

$$P(Z_n \le z) \simeq \Phi(z)$$

$$\Rightarrow \quad P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le z\right) \simeq \Phi(z)$$

$$\Rightarrow \quad P\left(\frac{nM_n - n\mu}{\sigma\sqrt{n}} \le z\right) \simeq \Phi(z)$$

$$\Rightarrow \quad P\left(|M_n - \mu| \le \frac{z\sigma}{\sqrt{n}}\right) \simeq \Phi(z)$$

$$\simeq 1 - 2Q(z)$$

$\boxed{\text{Example}}$ Suppose $X_1, X_2, \cdots X_n$ are iid RV with Bernoulli distribution, i.e.,

$$P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$$

We have $E[X_i] = \frac{1}{2}$, $Var(X_i) = \frac{1}{4} = \sigma^2$

Q: How close is M_n to $\frac{1}{2}$?

Say $n = 10000$ and $z = 1$

$$P\left(|M_n - \mu| \le \frac{z\sigma}{\sqrt{n}}\right) \simeq 1 - 2\Phi(z)$$

$$\Rightarrow P\left(\left|M_n - \frac{1}{2}\right| \le \frac{1}{200}\right) \simeq 1 - 2\Phi(1)$$

check Q-table

$= 0.682$

Q: For what value of n, we will have

$$P\left(\left|M_n - \frac{1}{2}\right| \le 0.005\right) \ge 0.9 \, ?$$

$\Rightarrow 1 - 2\Phi(z) \ge 0.9$, $\Phi(z) \le 0.05$

Check Q-table, $z \ge 1.65$

So it means that

$$\frac{z\sigma}{\sqrt{n}} = 0.005$$

$$\Rightarrow \frac{1.65 \times \frac{1}{2}}{\sqrt{n}} = 0.005 \Rightarrow \sqrt{n} = 165$$

$$\Rightarrow n = 27225$$

Note: When n is big enough that the CDF of Z_n approximates Gaussian, then the CLT will give better results than standard concentration inequalities (e.g., Markov, Chebyshev or Chernoff), but when n is small, the Gaussian approximation in CLT may not be accurate!

$\boxed{\text{Exam 3}}$ May 4 noon - 1:30PM Exam 3

Two pages ~~two-sided~~ notes

6~7 problems, ~2 prob from HWs

2 ~ material before Exam 1, 2 ~ related Exam 2, 2-3 after Exam 2.