EE 126 Probability and Random Processes
February 13, 2018 Convergence

1 Reading

5.2. The Weak Law of Large Numbers

The weak law of large numbers asserts that the sample mean of a large number of i.i.d. random variables is very close to the true mean, with high probability. If $X_1, ..., X_n$ is a sequence of i.i.d. random variables with mean μ and variance σ^2, then the sample mean is

$$M_n = \frac{X_1 + ... + X_n}{n}$$

the expectation of the sample mean is

$$\mathbb{E}[M_n] = \frac{\mathbb{E}[X_1] + ... + \mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu$$

and the variance of the sample mean is

$$var(M_n) = \frac{var(X_1 + ... + X_n)}{n^2} = \frac{var(X_1) + ... + var(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Applying the Chebyshev inequality, we have

$$P(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \quad \text{for any } \epsilon > 0$$

As n increases, the Chebyshev bound goes to 0! As a consequence, we obtain the formal WLLN:

The Weak Law of Large Numbers
Let $X_1, ..., X_n$ be i.i.d. random variables with mean μ. For every $\epsilon > 0$, we have

$$P(|M_n - \mu| \geq \epsilon) = P\left(\left|\frac{X_1 + ... + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \quad \text{as } n \to \infty$$

The WLLN suggests that for large n, the bulk of the distribution of M_n is concentrated near μ.
Essentially, the sample mean should converge to the true mean.

5.3. Convergence in Probability

Convergence of a Deterministic Sequence
Let $a_1, ..., a_n$ be a sequence of real numbers, and let a be another real number. We say that the sequence a_n converges to a, or $\lim_{n\to\infty} a_n = a$, if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon \quad \text{for all } n \geq n_0$$

Convergence in Probability
Let $Y_1, ..., Y_n$ be a sequence of random variables (not necessarily independent), and let a be a real number. We say that the sequence Y_n **converges to** a **in probability** if, for every $\epsilon > 0$, we have

$$\lim_{n\to\infty} P(|Y_n - a| \geq \epsilon) = 0$$

In other words, for every $\epsilon > 0$ and every $\delta > 0$, there exists some n_0 such that

$$P(|Y_n - a| \geq \epsilon) \leq \delta \quad \text{for all } n \geq n_0$$

If we call ϵ the "accuracy" level and δ the "confidence" level, then Y_n can be equal to a within any level of accuracy and confidence – provided n is large enough.

Even if a sequence Y_n converges to a, $\mathbb{E}[Y_n]$ might not converge to a!

5.4. The Central Limit Theorem

Let $X_1, ..., X_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. We define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

We can see that

$$\mathbb{E}[Z_n] = \frac{\mathbb{E}[X_1 + ... + X_n] - n\mu}{\sigma\sqrt{n}} = 0$$

and

$$var(Z_n) = \frac{var(X_1 + ... + X_n)}{\sigma^2 n} = \frac{var(X_1) + ... + var(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1$$

which brings us to the central limit theorem:

The Central Limit Theorem
Let $X_1, ..., X_n$ be a sequence of i.i.d. random variables with common mean μ and variance σ^2, and define

$$Z_n = \frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}$$

Then the CDF of Z_n converges to the standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

in the sense that

$$\lim_{n\to\infty} P(Z_n \leq z) = \Phi(z) \quad \text{for every } z$$

In other words, the sum of a large number of independent random variables is approximately normal. The central limit theorem allows us to calculate probabilities related to Z_n as if Z_n were normal. Note that this is equivalent to treating S_n as a normal random variable with mean $n\mu$ and variance $n\sigma^2$.

Normal Approximation Based on the Central Limit Theorem
Let $S_n = X_1 + ... + X_n$, where the X_i are i.i.d. random variables with mean μ and variance σ^2. If n is large, the probability $P(S_n \leq c)$ can be approximated by treating S_n as if it were normal, as per the procedure

1. Calculate the mean $n\mu$ and the variance $n\sigma^2$ of S_n.
2. Calculate the normalized value $z = (c - n\mu)/\sqrt{n\sigma^2}$.
3. Use the approximation $P(S_n \leq c) \approx \Phi(z)$, where $\Phi(z)$ is available from standard normal CDF tables.

5.4.1. De Moivre-Laplace Approximation to the Binomial

Recall: a binomial random variable S_n with parameters n and p can be viewed as the sum of n independent Bernoulli variables $X_1, ..., X_n$ with common parameter p:

$$S_n = X_1 + ... + X_n$$

where $\mu = \mathbb{E}[X_i] = p$ and $\sigma = \sqrt{var(X_i)} = \sqrt{p(1-p)}$.

We can use the CLT to provide an approximation for $P(\{k \leq S_n \leq l\})$, where k and l are given integers. We express this event as a standardized random variable:

$$k \leq S_n \leq l \iff \frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{l-np}{\sqrt{np(1-p)}}$$

By the central limit theorem, $(S_n - np)/\sqrt{np(1-p)}$ has (approximately) a standard normal distribution. Hence

$$\begin{aligned} P(k \leq S_n \leq l) &= P\left(\frac{k-np}{\sqrt{np(1-p)}} \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{l-np}{\sqrt{np(1-p)}}\right) \\ &\approx \Phi\left(\frac{l-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Essentially we're treating S_n as a normal random variable with mean np and variance $np(1-p)$. Notably, a more accurate approximation can be obtained if we replace k with $k - 1/2$ and l with $l + 1/2$. This gives us the following:

De Moivre-Laplace Approximation to the Binomial
If S_n is a binomial random variable with parameters n and p, n is large, and k, l are nonnegative integers, then

$$P(k \leq S_n \leq l) \approx \Phi\left(\frac{l + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

The quality of the above approximation is best when p is close to 0.5, and worse when p is close to 0 or 1. (When p is close to 0 or 1, a larger value of n will be needed to maintain the same accuracy.)

5.5. The Strong Law of Large Numbers

The SLLN also deals with the convergence of the sample mean to the true mean.

The Strong Law of Large Numbers
Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with mean μ. Then the sequence of sample means $M_n = (X_1 + ... + X_n)/n$ converges to μ **with probability 1**, i.e.

$$P\left(\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} = \mu\right) = 1$$

By the strong law: for any given $\epsilon > 0$ the probability that the difference $|M_n - \mu|$ will exceed ϵ an infinite number of times is equal to 0. With the weak law, it is possible that M_n deviates significantly from μ every once in a while.

The strong law asserts that M_n converges to μ with probability 1. The weak law asserts that M_n converges to μ in probability.

Convergence with Probability 1
Let $Y_1, Y_2, ...$ be a sequence of random variables (not necessarily independent). We say that Y_n converges to c with probability 1 (or **almost surely**) if

$$P\left(\lim_{n\to\infty} Y_n = c\right) = 1$$

Convergence with probability 1 implies convergence in probability, but the converse is not necessarily true.

W2.3. Laws of Large Numbers for i.i.d. Random Variables

If X_n and X are random variables defined on a common probability space, X_n **converges in probability** to X if for all $\epsilon > 0$ we have

$$P(|X_n - X| \geq \epsilon) \to 0 \quad \text{as } n \to \infty$$

The notation for "converges in probability" is $X_n \xrightarrow{p} X$.

Meanwhile, X_n converges **almost surely** to X as $n \to \infty$ if

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

The notation for "converges almost surely" is $X_n \to X$.

"Almost sure" convergence implies "convergence in probability," so the SLLN is stronger than the WLLN. By these laws of large numbers, the sample mean values $Y_n = (X_0 + ... + X_{n-1})/n$ converge to the expected value with probability 1.

Modes of Convergence

When working with sequences of real numbers, the definition of convergence is easy. A sequence of real numbers $(a_n)_{n\in\mathbb{N}}$ is said to converge to a limit, i.e. $a_n \xrightarrow[n\to\infty]{} a$, if for every $\epsilon > 0$ there exists a positive integer N such that the sequence after N is always within ϵ of the supposed limit a.

On the other hand, the notion of convergence becomes more subtle as we discuss functions (e.g. random variables) instead of numbers. We will examine the modes of convergence for a fixed probability space Ω, sequence of random variables $(X_n)_{n\in\mathbb{N}}$, and other random variable X.

Almost Sure Convergence

The sequence $(X_n)_{n\in\mathbb{N}}$ converges almost surely (/with probability one) to X if the set of outcomes $\omega \in \Omega$ for which $X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)$ forms an event of probability one.

The **SLLN** states that the sample average of i.i.d. random variables converges a.s. to the expected value of their common distribution.

For example, in SGD from ML we want to know whether iterates converge a.s. to the true minimizer of the function.

Convergence in Probability

The sequence $(X_n)_{n\in\mathbb{N}}$ converges in probability to X if, for every $\epsilon > 0$, $P(|X_n - X| > \epsilon) \xrightarrow{n\to\infty} 0$. In other words, for any fixed $\epsilon > 0$, the probability that the sequence deviates from the supposed limit X by more than ϵ becomes vanishingly small.

Convergence in Distribution

$X_n \xrightarrow[n\to\infty]{\text{d}} X$ if, for every $x \in \mathbb{R}$ such that $P(X = x) = 0$, we have $P(X_n \leq x) \xrightarrow[n\to\infty]{} P(X \leq x)$. This does not require all of the random variables to be defined on the same probability space.

Convergence in probability implies convergence in distribution! However, convergence in distribution does not imply convergence in probability.

The **CLT** is a statement about convergence in distribution (specifically to the standard normal distribution).

Note: for *none* of the modes of convergence mentioned thus far is it implied that $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$. On the other hand, for *all* of the modes of convergence mentioned thus far, it's the case that if X_n converges in some way to X, then $f(X_n)$ converges in the same way to $f(X)$ [where f is a continuous function].

2 Lecture

Recap

The Chernoff bound asserts that

$$P(Y \geq b) \leq \min_s \left(e^{-sb}\ \mathbb{E}[e^{sY}]\right)$$

Note that $\mathbb{E}[e^{sY}] = M_Y(s)$.

The MGF of the standard normal $X \sim \mathcal{N}(0,1)$ is $M_X(s) = e^{-s^2/2}$. From Chernoff, we have $P(X \geq b) \leq e^{-b^2/2}$.

Weak Law of Large Numbers

WLLN: If $X_1, ..., X_n$ are i.i.d. random variables with mean μ and finite variance σ^2, then

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty$$

where $\frac{1}{n}\sum_{i=1}^{n} X_i$ is the empirical mean M_n. Formally, for any $\epsilon, \delta > 0$ (arbitrarily small, but positive), there exists an $n(\epsilon, \delta)$ such that $P(|M_n - \mu| \geq \epsilon) < \delta$ for all $n > n(\epsilon, \delta)$.

ϵ captures "accuracy" level. δ captures "confidence" level. $n(\epsilon, \delta)$ is a threshold value for target accuracy and confidence level.

Note that $\mathbb{E}[M_n] = \mu$, while $var(M_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$. The Chebyshev inequality tells us

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty$$

Hence Chebyshev is a proof for the weak law of large numbers.

In layman's terms: if we compute M_n for large enough n, then almost always $|M_n - \mu| \leq \epsilon$. We say M_n *converges in probability* to μ. ($M_n \to \mu$.)

Example. $X_1, ..., X_n$ are i.i.d. Uniform$[-1, 1]$. If $Y_n = X_n/n$, does Y_n converge in probability? To what?

Y_n should converge in probability to 0, since X_n is something between -1 and 1 while n will only get larger and larger. Working it out, we have

$$\begin{aligned} Y_n \leq y &\implies X_n \leq ny \\ F_n(y) = F_X(ny) &\implies f_{Y_n}(y) = n f_{X_n}(ny) \end{aligned}$$

Then $P(|Y_n - 0| > \epsilon) = 0$ if $\frac{1}{n} < \epsilon$, or $n > \frac{1}{\epsilon}$.

Example. If $X_1, ..., X_n$ are i.i.d. Uniform$[0, 1]$ and $Y_n = \min(X_1, ..., X_n)$, then

$$\begin{aligned} P(|Y_n - 0| > \epsilon) &= P(X_1 > \epsilon) \cdots P(X_n > \epsilon) \\ &= (1 - \epsilon)^n \to 0 \text{ as } n \to \infty \end{aligned}$$

Example. Suppose time is discrete $(1, 2, 3, ...)$ and $Y_n = \begin{cases} 1 & \text{if arrival at time } n \\ 0 & \text{otherwise} \end{cases}$

Define $I_K = \{2^k, 2^k + 1, ..., 2^{k+1} - 1\}, \quad k = 0, 1, ...$ The first interval I_0 includes time 1. The second interval I_1 includes times 2 and 3. The third interval I_2 includes times $4, 5, 6, 7...$ essentially, we are doubling the size of our intervals as we go along.

Our arrival process is such that we will have one arrival in every interval (at any time within the interval). Formally, we will suppose there is exactly one arrival in each interval I_K and it is equally likely to be at any time within each interval.

As a sanity check, we have

- $P(Y_1 = 1) = 1$
- $P(Y_2 = 1) = P(Y_3 = 1) = 0.5$
- $P(Y_4 = 1) = P(Y_5 = 1) = P(Y_6 = 1) = P(Y_7 = 1) = 0.25$

What is the probability that $Y_n = 1$ if Y_n is in the kth interval? i.e. what is $P(Y_n = 1)$ if $n \in I_k$? As it happens, $P(Y_n = 1) = 1/2^k$ if $n \in I_k$. But now

$$\lim_{n\to\infty} P(Y_n = 1) = \lim_{k\to\infty} \frac{1}{2^k} = 0$$

Supposedly Y_n is converging to 0.

Given any finite (but also large) n, there is certain to be an infinite number of arrivals after n. And yet we are supposedly converging to 0 in some sense! Since it doesn't *really* converge, that sense must therefore be weak. This is the *weakness* of the *weak law*.

Strong Law of Large Numbers

Let $\{X_n,\ n \geq 0\}$ be a sequence of random variables with mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \text{ as } n \to \infty \text{ with probability 1}$$

...i.e. the sample mean converges to the expected value μ with probability 1.

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mu\right) = 1$$

$$Y_n \xrightarrow{\text{almost surely}} \mu$$

The WLLN says that the fraction of the excursions outside $(\mu - \epsilon, \mu + \epsilon)$ converges to 0. The SLLN says that *every* realization of the sample mean converges to μ.

Central Limit Theorem

Let $S_n = X_1 + ... + X_n$, where the X_i's are i.i.d. and drawn from Uniform$[0, 1]$. We want to know the distribution of S_n. As n gets larger, the distribution of S_n will become smoother and more spread out; notably, it will look increasingly bell-shaped. The same is also true if X is drawn from Exp(1) instead of Uniform$[0, 1]$.

As n gets large, the mean and variance of S_n converge to ∞. Therefore we need to normalize! We will define

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

where $n\mu$ is the mean of S_n and $n\sigma^2$ is the variance. Z_n will have a mean of zero and a variance of one.

The CLT says (amazingly!) that

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x) \text{ for every } x$$

where $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$, i.e. the CDF of $\mathcal{N}(0, 1)$.

In other words, $Z_n \xrightarrow{\text{converges in distribution}} \mathcal{N}(0, 1)$.

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific books. Athena Scientific, 2002.

[2] Jean Walrand. *Probability in Electrical Engineering and Computer Science: An Application-Driven Course.* Quorum Books, Westport, CT, USA, 2014.