Chapter 8

Sums of Random Variables: Bounds and Limits

So far in this course, we have focused mostly on pairs of random variables X and Y. Many experiments of interest generate more than two random variables for each outcome. When we consider we consider $n \geq 2$ random variables $X_1, \ldots, X_n$, we describe their probabilistic behavior using a joint Cumulative distribution function (CDF) of the form

$$F_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}[\{X_1 \leq x_1, \ldots, X_n \leq x_n\}],$$

which is the natural extension of the joint CDF for pairs of random variables. When the joint random variables are discrete, we define the joint probability mass function (PMF) as

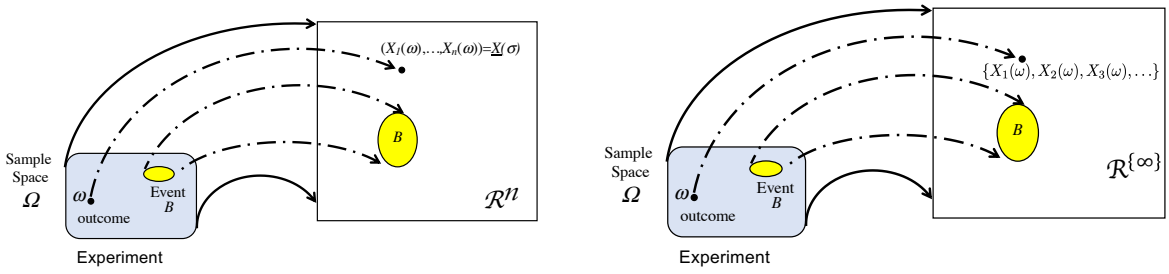
$$P_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \mathbb{P}[\{X_1 = x_1, \ldots, X_n = x_n\}].$$

When the random variables are continuous, we define the joint probability density function (PDF) as

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{X_1,\ldots,X_n}(x_1,\ldots,x_n).$$

All of the basic properties that we established for CDFs, PMFs and PDFs for pairs of random variables extend naturally to CDFs, PMFs and PDFs of n random variables.

In this chapter, we study experiments that generate a countably infinite collection of random variables. Such collections are often called discrete time random processes, as the index of the random variables can be mapped to the countable natural numbers. Figure 8.1 compares experiments that generate random vectors, which we have discussed previously, to ones that generate a countable collection of random variables. Formally, each element $X_k(\omega)$ of the collection $\{X_1, X_2, X_3, \ldots\}$ is a random variable, a measurable function from the sample space Ω to the real numbers. Such collections are often called random processes or stochastic processes. A random process is an indexed collection $\{X_t, t \in T\}$ of random variables generated by a single experiment. When the index T is countable and can be mapped to the natural numbers $\mathcal{N}$, we refer to such processes as discrete-time or discrete-index random processes. Such processes are generalizations of the concept of random vectors introduced in earlier chapters, as shown in Figure 8.1.



Random vectors | Infinite sequence of random variables

Figure 8.1: Experiments generate infinite sequences of random variables.

The study of general random processes is a subject of a more advanced course, and requires tools that we will not introduce in this course. There are special cases which we can address with simple extensions of

the methodology we have described in previous chapters. Like random vectors, for general random processes one would have to define the joint probability mass functions for finite set of distinct indices $k_1, k_2, \ldots, k_n$, of the form $P_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n)$ if the random variables are discrete. Equivalently, one would need to define joint probability density functions $f_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n)$. Computing such joint densities is cumbersome and hard to describe.

In this chapter, we focus on the special case where the collection of random variables $\{X_1, X_2, X_3, \ldots\}$ are mutually independent. This implies that, for any finite set of distinct indices $k_1, k_2, \ldots, k_n$, and subsets $A_1, A_2, \ldots A_n \subset \Re$, we have

$$\mathbb{P}[\{X_{k_1} \in A_1\}, \{X_{k_2} \in A_2\}, \ldots, \{X_{k_n} \in A_n\}] = \mathbb{P}[\{X_{k_1} \in A_1\}]\mathbb{P}[\{X_{k_1} \in A_1\}] \cdots \mathbb{P}[\{X_{k_n} \in A_n\}].$$

When the random variables are discrete, the joint probability mass functions factor as

$$P_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n) = P_{X_{k_1}}(x_1)P_{X_{k_2}}(x_2) \cdot P_{X_{k_n}}(x_n).$$

For continuous random variables, the joint densities factor as

$$f_{X_{k_1},X_{k_2},\ldots,X_{k_n}}(x_1, x_2, \ldots, x_n) = f_{X_{k_1}}(x_1)f_{X_{k_2}}(x_2) \cdot f_{X_{k_n}}(x_n).$$

This independence property will allow us to analyze properties of the collection of random variables using the tools we have developed for the analysis of pairs of random variables in earlier chapters.

Of particular interest is the case where the collection $\{X_1, X_2, X_3, \ldots\}$ corresponds to outputs of repeating an experiment independently, with an infinite number of trials. For instance, let X_i correspond to the output of a Bernoulli trial, with parameter p that represents the probability that $X_i = 1$. The empirical theory of probability suggests that p should be the fraction of experiments that result in an outcome $X_i = 1$. From the results of the last chapter, the maximum likelihood estimate of p given observations of the outcomes of the first N experiments is $\frac{\sum_{i=1}^{N} X_i}{N}$. What happens as the number of experiments N increases to infinity? In the limit, we would expect that this estimate, which is a derived random variable, would converge in some sense to the correct value p. We will analyze the behavior of such sequences of random variables and make precise in what manner do such sequences converge.

8.1 Independent, Identically Distributed Random Variables

A collection of random variables $\{X_n, n \in \mathcal{N}\}$ is referred to as an independent, identically distributed collection of random variables if the random variables $X_1, X_2, \ldots$ are mutually independent, and the marginal cumulative distribution function of each random variable is the same for each random variable. That is, $F_{X_k}(x) = F_{X_j}(x)$ for any $j, k \in \mathcal{N}$. We use the short-hand notation **i.i.d.** to represent independent and identically distributed in the rest of this chapter.

Let $\{X_n, n \in \mathcal{N}\}$ be an i.i.d. collection of random variables, each of which has finite mean μ and finite variance σ^2. Define a sequence of dependent random variables S_n using partial sums as:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Using linearity of expectation and the i.i.d. property, we establish the following:

$$\mathbb{E}[S_n] = \sum_{j=1}^{n} E[X_j] = n\mu.$$

What about the covariance of S_n? This is also computed readily, as

$$\begin{aligned}\mathsf{Var}S_n &= \mathbb{E}[(S_n - \mathbb{E}[S_n])^2] = \mathbb{E}\big[\big(\sum_{j=1}^{n}(X_j-\mu)\big)^2\big] \\ &= \mathbb{E}\big[\sum_{j=1}^{n}\sum_{k=1}^{n}(X_j-\mu)(X_i-\mu)\big] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}\big[(X_j-\mu)(X_i-\mu)\big] \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n}\mathsf{Cov}[X_j,X_k] = \sum_{j=1}^{n}\mathsf{Var}[X_j] = n\sigma^2\end{aligned}$$

where the last equality follows because the X_j are i.i.d., hence $\mathsf{Cov}[X_j, X_k] = 0$ if $k \neq j$, and $\mathsf{Cov}[X_j, X_k] = \mathsf{Var}[X_j] = \sigma^2$ if $k = j$.

Notice that, as n grows, $\mathbb{E}[S_n]$ and $\mathsf{Var}[S_n]$ both grow linearly with n. Thus, we don't expect any type of convergence for the sequence S_n. Let's define instead the variables $M_n = \frac{S_n}{n}$, the average of the first n random variables X_k. Then,

$$\mathbb{E}[M_n] = \frac{\mathbb{E}[S_n]}{n} = \mu,$$

and

$$\mathsf{Var}[M_n] = \big(\frac{1}{n}\big)^2\mathsf{Var}[S_n] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Note now that, as n increases, the variables M_n have the same mean, and the variance of the random variables decreases. The distribution of M_n becomes more concentrated about its average μ.

Example 8.1
Let X be an exponential random variable with $\lambda = 1$. Thus, $\mathbb{E}[X] = \frac{1}{\lambda} = 1$. Let M_n denote the sample mean of n independent samples of X. How many samples are needed so that the variance of the sample mean is less than or equal to 0.01?

From the properties of exponential random variables, $\mathsf{Var}[X] = \frac{1}{\lambda^2} = 1$. Hence, for the average of n samples, $\mathsf{Var}[M_n] = \frac{\mathsf{Var}X}{n}$. This means that we need at least 100 samples for the variance to be 0.01 or less.

At this point, we don't know much about the probability distribution of M_n. Indeed, since M_n is a sum of independent random variables, its probability density function is an n-fold convolution of the densities of the scaled random variables X_j/n. In order to make statements concerning the probability of events related to M_n, we discuss next some estimates of such probabilities based on only mean and variance information.

8.2 Useful inequalities for Random Variables

In order to analyze notions of convergence of random variables, it is useful to bound the errors between the limit random variable and elements of the sequence using inequalities that do not require knowledge of the full distribution of the random variables. Below, we present a few useful inequalities:

8.2.1 Markov inequality

Suppose that X is a non-negative random variable with known finite mean, and we want to obtain some bounds on the probability distribution function of X. The **Markov Inequality** is given by

$$\mathbb{P}[\{X \geq a\}] = \int_a^\infty f_X(x)\,dx \leq \frac{\mathbb{E}[X]}{a}.$$

How do show the Markov inequality is true? The steps below illustrate the argument when X is a continuous random variable with finite expected value. Since X is non-negative, the density $f_X(x)$ is zero for $x < 0$.

$$\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_{0}^{\infty} x f_X(x)\,dx \\
&= \int_{0}^{a} x f_X(x)\,dx + \int_{a}^{\infty} x f_X(x)\,dx \\
&\geq \int_{a}^{\infty} x f_X(x)\,dx \qquad \text{(Drop the first term, non-negative)} \\
&\geq a \int_{a}^{\infty} f_X(x)\,dx = a\mathbb{P}[\{X \geq a\}] \qquad (x \geq c \text{ in the integrand})
\end{aligned}$$

The Markov inequality follows by dividing both sides by a.

The above argument can be generalized as follows: Let $g(x) \geq 0$ everywhere, and let $g(x) > a > 0$ for all $x \in A$, for a subset A of the real line $\Re$. Then,

$$\begin{aligned}
E[g(X)] &= \int_{x\in A} g(x) f_X(x)\,dx + \int_{x\notin A} g(x) f_X(x)\,dx \\
&\geq \int_{x\in A} g(x) f_X(x)\,dx \geq a \int_{x\in A} f_X(x)\,dx \\
&= a\mathbb{P}[\{X \in A\}].
\end{aligned}$$

Hence, $\mathbb{P}[\{X \in A\}] \leq \frac{E[g(X)]}{a}$.

8.2.2 Chebyshev inequality

Suppose that the mean μ and variance σ^2 of a random variable X are known, and we would like to bound the probability that the variable is far from its mean. The Chebyshev inequality states that

$$\mathbb{P}[\{|X - \mu| \geq a\}] \leq \frac{\sigma^2}{a^2}.$$

The Chebyshev inequality can be derived from the Markov Inequality, by defining the non-negative random variable $Y = (X - \mu)^2$. Since $\mathbb{E}[Y] = \mathsf{Var}[X]$ is finite, the Markov inequality states that

$$\mathbb{P}[\{Y \geq a^2\}] \leq \frac{\mathbb{E}[Y]}{a^2} = \frac{\sigma^2}{a^2}.$$

In terms of equivalent events,

$$\mathbb{P}[\{|X - \mu| \geq a\}] = \mathbb{P}[\{Y \geq a^2\}] \leq \frac{\sigma^2}{a^2},$$

which shows the Chebyshev Inequality.

A different way of writing the Chebyshev inequality is as follows: Let $a = a'\sigma$. Then,

$$\mathbb{P}[\{|X - \mu| \geq a\}]\mathbb{P}[\{|X - \mu| \geq a'\sigma\}] =\leq \frac{\sigma^2}{a'^2\sigma^2} = \frac{1}{a'^2}.$$

This can be interpreted as in terms of number of standard deviations away from the mean. The probability that X is more than a' standard deviations away from its mean is less than $\frac{1}{a'^2}$.

The above can be generalized for any random variable X such that $\mathbb{E}[(X - \mu)^n]$ is finite for some even number n, as

$$\mathbb{P}[\{|X - m| \geq a\}] = \mathbb{P}[\{|X - m|^n \geq a^n\}] \leq \frac{\mathbb{E}[|X - m|^n]}{a^n}$$

or, more generally, for any real, nonnegative, even function $g(x)$ which is non-decreasing for $x > 0$, and has finite expectation. Then,

$$\mathbb{P}[\{g(X) \geq g(a)\}] \leq \frac{\mathbb{E}[g(X)]}{g(a)}.$$

Example 8.2
A random variable W, which represents the waiting time to be served at a restaurant, is uniformly distributed in the interval from 0 to 10 minutes. Estimate a bound on the probability that the wait is at least 8 minutes.

Note that, in this case, we know the exact probability of the event $\{W \geq 8\}$, because we have the density of W: Hence, $\mathbb{P}[\{W \geq 8\}] = 0.2$. What if we estimated this using either the Markov inequality or the Chebyshev inequality? We know that $\mathbb{E}[W] = 5$, and $\text{Var}[W] = \frac{100}{12} = \frac{25}{3}$. We also know that $W \geq 0$. Hence, the Markov inequality indicates that

$$\mathbb{P}[\{W \geq 8\}] \leq \frac{\mathbb{E}[W]}{8} = \frac{5}{8},$$

which is much larger than 0.2. It shows that the bound can be loose.

What about the Chebyshev inequality? It states:

$$\mathbb{P}[\{|W - 5| \geq 3\}] \leq \frac{\frac{25}{3}}{9} = \frac{25}{27}.$$

If we divide by 2 to represent the one-sided probability that $W > 8$, we have

$$\mathbb{P}[\{W \geq 8\}] \leq \frac{25}{54},$$

which is closer to 0.2, but still a loose bound.

Example 8.3
Assume X is Gaussian, with mean 0 and variance 1. Then, $\mathbb{P}[\{|X| > a\}] = 2Q(a)$, where $Q(\cdot)$ is the standard Gaussian complementary cumulative distribution function. We can compare, as a function of a, the estimate generated by the Chebyshev inequality and the true value $2Q(a)$, as:

Value of a	Chebyshev Inequality	2Q(a)
a = 2	0.25	0.0455
a = 3	0.111	0.0027
a = 4	0.0625	0.0000633
a = 5	0.04	0.0000006

The values illustrate the conservative nature of the Chebyshev inequality.

Example 8.4
Chebyshev's Inequality can provide a tight bound for some distributions. Consider the discrete random variable X with range in $R_X = \{-1, 1\}$ such that $P(1) = 0.5, P(-1) = 0.5$. Then, $\mathbb{E}[X] = 0, \text{Var}[X] = 1$. Therefore, Chebyshev's Inequality states that

$$\mathbb{P}[\{|X - \mathbb{E}[X]| \geq 1\}] \leq 1.$$

However, we know that $\mathbb{P}[\{|X - \mathbb{E}[X]| \geq 1\}] = 1$ in this example, so the bound is equal to the actual probability.

8.2.3 Chernoff and Jensen Inequalities

There are other bounds on probabilities of random variables that are useful to know. We discuss them briefly here without proof.

Given a random variable X, define a new random variable Y_ϵ as:

$$Y_\epsilon = \begin{cases} 1 & X \geq \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

That is, Y is the indicator random variable that $X \geq \epsilon$.

Then, for all $t \geq 0$, the following inequality holds:

$$e^{tX} \geq e^{t\epsilon} Y.$$

Thus,

$$\mathbb{E}\left[e^{tX}\right] \geq \mathbb{E}\left[e^{t\epsilon}Y\right] = e^{t\epsilon}\mathbb{P}[\{X \geq \epsilon\}],$$

which implies that

$$\mathbb{P}[\{X \geq \epsilon\}] \leq e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

This bound can be tightened through the choice of t, as follows:

$$\mathbb{P}[\{X \geq \epsilon\}] \leq \min_{t \geq 0} e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

Note that this bound requires computation of $\mathbb{E}\left[e^{tX}\right]$, which is equivalent to computing the characteristic function (or moment-generating function) of X! Thus, this bound requires extensive knowledge of the full probability density function of X, and not just its mean and variance.

Another useful inequality is Jensen's inequality. A *convex* function $g(x)$ of a continuous variable x in an interval I is a function such that, for any $\alpha \in [0,1]$, any x, $y \in I$, the following is true:

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

Let X denote a random variable with probability density or probability mass function distributed over I, and let μ denote its mean, which must be in I. Then, for any convex function g, we have

$$g(\mu) \leq \mathbb{E}[g(X)].$$

One way to recognize that this is true is to note that, if X were a discrete random variable with $P_X(x) = \alpha, P_X(y) = 1-\alpha$, then the definition of g as convex implies

$$g(\alpha x + (1-\alpha)y) = g(\mathbb{E}[X]) \leq \alpha g(x) + (1-\alpha)g(y) = \mathbb{E}[g(X)].$$

This can be extended to other discrete probability mass functions, and in a limiting argument to continuous random variables X.

Jensen's inequality can be used to derive many inequalities concerning moments of random variables, such as the Cauchy-Schwartz inequality that we used to prove that the correlation coefficient between two random variables X, Y is a number with magnitude less than or equal to 1.

Example 8.5
Assume X is Binomial(n,p). Then, using the Chernoff bound, we have

$$\mathbb{P}[\{X \geq \epsilon\}] \leq \min_{t \geq 0} e^{-t\epsilon}\mathbb{E}\left[e^{tX}\right], \quad t \geq 0.$$

We can compute $\mathbb{E}\left[e^{tX}\right]$ in this case, as X is the sum of n independent Bernoulli(p) random variables $Y_1, \ldots, Y_n$. Hence,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[e^{t\sum_{k=1}^n Y_k}\right] = \prod_{k=1}^n \mathbb{E}[e^{tY_k}] = \prod_{k=1}^n (1-p+pe^t) = (1-p+pe^t)^n.$$

Let's compute a bound on $\mathbb{P}[\{X \geq \alpha n\}]$ for $1 > \alpha > p$. Then,

$$\min_{t \geq 0} e^{-t\alpha n}\mathbb{E}\left[e^{tX}\right] = \min_{t \geq 0} e^{-t\alpha n}(1-p+pe^t)^n$$

Taking the derivative with respect to t and setting it equal to 0 yields

$$\frac{d}{dt}\left(e^{-t\alpha n}(1-p+pe^t)^n\right) = -\alpha n e^{-t\alpha n}(1-p+pe^t)^n + npe^t e^{-t\alpha n}(1-p+pe^t)^{n-1} = 0.$$

Dividing by common factors yields the solution at the minimum value:

$$\alpha n(1-p+pe^t) = npe^t \Rightarrow e^t = \frac{\alpha(1-p)}{p(1-\alpha)}$$

Substituting into the bound, we get:

$$\min_{t\geq 0} e^{-t\alpha n}\mathbb{E}\left[e^{tX}\right] = \left(\frac{p(1-\alpha)}{\alpha(1-p)}\right)^{\alpha n}\left(1-p+p\frac{\alpha(1-p)}{p(1-\alpha)}\right) = \left(\frac{p(1-\alpha)}{\alpha(1-p)}\right)^{\alpha n}\frac{(1-p)}{(1-\alpha)} = \left(\frac{p}{\alpha}\right)^{\alpha n}\left(\frac{1-\alpha}{1-p}\right)^{\alpha n-1}$$

For $p = 0.5, \alpha = 0.75$ the above bound is $\mathbb{P}[\{X \geq \alpha n\}] = 2\left(\frac{1}{3}\right)^{0.75n}$, which decays fast as n increases.

8.2.4 Hoeffding's Inequality

Hoeffding's inequality provides bounds on probabilities of the averages of random variables. Let $X_1, \ldots, X_n$ be independent random variables whose range $R_{X_k} \subset [a_k, b_k]$, where $-\infty < a_k < b_k < \infty$. That is, with probability 1, $a_k \geq X_k \leq b_k$ for $k = 1, \ldots, n$. We define the sample mean of these variables by

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

Then,

$$\mathbb{P}[\{M_n - \mathbb{E}[M_n] \geq \epsilon\}] \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}},$$

$$\mathbb{P}[\{M_n - \mathbb{E}[M_n] \leq -\epsilon\}] \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}}.$$

We can combine the two bounds to get a bound that is similar to the Chebyshev bound, as

$$\mathbb{P}[\{|M_n - \mathbb{E}[M_n]| \geq \epsilon\}] \leq 2e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^n (b_k-a_k)^2}}.$$

For the special case that X_k are independent, identically distributed Bernoulli(p) random variables, $a_k = 0, b_k = 1$, and thus $\sum_{k=1}^n (b_k - a_k)^2 = n$. In this case, Hoeffding's inequality yields

$$\mathbb{P}[\{|M_n - p| \geq \epsilon\}] \leq 2e^{-2n\epsilon^2}.$$

Example 8.6

Let's apply Hoeffding's inequality to the previous example, where X is Binomial(n, p), so that X is the sum of n independent Bernoulli(p) random variables $Y_1, \ldots, Y_n$. We want to compute $\mathbb{P}[\{X \geq \alpha n\}]$ for $1 > \alpha > p$. Note that $M_n = \frac{X}{n}$, so $\mathbb{P}[\{X \geq \alpha n\}] = \mathbb{P}[\{M_n \geq \alpha\}] = \mathbb{P}[\{M_n - p \geq \alpha - p\}$. Using Hoeffding's inequality, we have

$$\mathbb{P}[\{M_n - p \geq \alpha - p\} \leq e^{-2n(\alpha-p)^2}.$$

For $\alpha = 0.75, p = 0.5$, this bound becomes $\mathbb{P}[\{M_n - p \geq 0.25\} \leq e^{-\frac{n}{8}}$.

8.3 The Law of Large Numbers

The law of large numbers has a central role in probability and statistics. It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value. It is consistent with the frequency interpretation of the concept of probability, where the probability of an event is the fraction of times when the event occurs if the experiment were repeated independently an infinite number of times. There are two main versions of the Law of Large Numbers: the weak law of large numbers and the strong law of large numbers. The differences are subtle, and we will highlight some of the

We state the weak law of large numbers first, and then prove it.

Theorem 8.1 (Weak Law of Large Numbers)
Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite means $\mathbb{E}[X_n] = \mu$, and define the sequence of sample means $\{M_n\}$ as

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

Then, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[\{\, |M_n - \mu| > \varepsilon \,\}] = 0.$$

The proof of this theorem in its general case is subtle, and requires a truncation argument. We will instead show this using the additional assumption that $\mathsf{Var}[X_n] = \sigma^2 < \infty$. In this case, we have already shown in Section 8.1 that $\mathbb{E}[M_n] = \mu, \mathsf{Var}[M_n] = \frac{\sigma^2}{n}$. Using Chebyshev's inequality, we have that,

$$\mathbb{P}[\{|M_n - \mu| \geq \varepsilon\}] \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Taking the limit of this as $n \to \infty$ establishes the weak law of large numbers.

As mentioned earlier, the weak law applies in the case of i.i.d. random variables, but it also applies in some other cases. For instance, if the X_n have finite bounded variances, and are uncorrelated, the law still holds. Even if the variances grow unbounded with n, as long as the variance of the averages M_n goes to zero as $n \to \infty$, the same argument can be applied to establish the weak law of large numbers.

The type of convergence used in the weak law of large numbers is convergence in probability. A sequence of random variables $\{M_n\}$ converges to a limiting random variable M in probability if and only if

$$\lim_{n\to\infty} \mathbb{P}[\{|M_n - M| > \varepsilon\}] = 0.$$

When the random variables X_n have finite variance, bounded by σ^2, we can show the averages M_n converge to their limit in mean square also, which means

$$\lim_{n\to\infty} \mathbb{E}[(M_n - \mu)^2] = 0.$$

This is trivial to show as we know the variance of M_n goes to zero, and the mean is μ.

Example 8.7
Assume X is a Bernoulli random variable, with probability p that $X = 1$. Let X_k be a repetition of the same experiment, for $k = 1, 2, \ldots$. From our results in estimation, we know that the maximum likelihood estimate of p given n observations X_k is given by

$$\widehat{p}_{ML}(\{X_k, k = 1, 2, \ldots, n\}) = \frac{\sum_{k=1}^{n} X_k}{n}.$$

which is the sample average discussed above. By the weak Law of Large Numbers,

$$\mathbb{P}[\{|\frac{\sum_{k=1}^{n} X_k}{n} - \widehat{p}_{ML}(\{X_k, k = 1, 2, \ldots, n\})| \geq \varepsilon\}] \leq \frac{\sigma^2}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2},$$

which converges to zero as $n \to \infty$.

Example 8.8
One of the problems with MMSE estimation that we discussed in Section 7.3 is that the integrals are hard to compute. For instance, in Example 7.6, we had to compute the following integral to get the conditional density of X given $Y = y$: $\int_0^{1000} \frac{2x}{10^6} \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx$.

In general, suppose we have a function $g(x)$, and we wanted to compute $\int_a^b g(x)\, dx$, but $g(x)$ was a continuous function that was hard to integrate. We can compute the integral approximately using the weak Law of Large Numbers as follows: Let $\{X_n\}$ be an i.i.d. sequence of random variables, uniformly distributed in $[a, b]$. Let $Y_n = g(X_n)$. Then, $\{Y_n\}$ is also an i.i.d. sequence, and $\mathbb{E}[Y] = \int_a^b \frac{g(x)}{b-a}\, dx$. Given that $[a, b]$ is a bounded interval and $g(x)$ is continuous, we can show that $\mathsf{Var}[Y_n] = \sigma_Y^2 < \infty$.

By the weak Law of Large Numbers, the average $\frac{Y_1+Y_2+\ldots+Y_n}{n}$ is close to $\mathbb{E}[Y]$. Hence, an approximation for the integral is

$$\int_a^b g(x)\,dx \approx (b-a)\frac{Y_1+Y_2+\ldots+Y_n}{n}.$$

Furthermore, we can compute the probability that the error is significant using the Chebyshev inequality. This probabilistic technique is known as the Monte Carlo method of integration.

The statement of the weak law of large numbers is a statement about probabilities, averaged over all the outcomes in the experiment. It does not guarantee that, for any outcome $\omega \in \Omega$ that generates a sequence of realizations of random variables $X_1(\omega), X_2(\omega), \ldots$, the average of those random variables will be close to $\mathbb{E}[X] = \mu$. It does not even guarantee that the set of outcomes for which the average does not converge to μ has zero probability of occurring. For that, we need the Strong Law of Large Numbers, stated next:

Theorem 8.2 (Strong Law of Large Numbers)
Let $\{X_n\}$ be a sequence of independent, identically distributed random variables with finite mean μ. Define the sequence of sample means $\{M_n\}$ as

$$M_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

then,

$$\mathbb{P}\left[\{\omega \in \Omega : \lim_{n\to\infty} M_n(\omega) = \mu\}\right] = 1.$$

The type of convergence in the strong law of large numbers is known as almost sure convergence. It states that the probability of an outcome where the sequence does not converge is zero. The proof is more complex than that of the weak law and is beyond the scope of our course. The strong law requires independence of the random variables X_k, whereas the weak law can be established using uncorrelated assumptions.

The main difference between the strong law of large numbers and the weak law of large numbers is where the limit is placed in the statement: The weak law states:

$$\lim_{n\to\infty} \mathbb{P}[\{\, |M_n - \mu| > \varepsilon \,\}] = 0,$$

whereas the strong law states:

$$\mathbb{P}\left[\{\omega \in \Omega : \lim_{n\to\infty} M_n = \mu\}\right] = 1.$$

Thus, the strong law states that, for any $\varepsilon > 0$, the probability of the event $\{|M_n - \mu| > \varepsilon$ for at most a finite $n\}$ is equal to 1.

8.4 The Central Limit Theorem

The law of large numbers characterizes that the sample averages M_n converge to a deterministic quantity, the mean $\mathbb{E}[X] = \mu$. Basically, it states that the cumulative distribution function $F_{M_n}(z)$ converges to a unit step function:

$$F_{M_n}(z) = \begin{cases} 0 & z < \mu \\ 1 & z \geq \mu. \end{cases}$$

It is often of interest to characterize the error $M_n - \mu$. We know from our previous analysis that, if the sequence $\{X_k\}$ is i.i.d., with finite mean μ and finite variance σ^2, the error $M_n - \mu$ has 0 mean, and variance $\frac{\sigma^2}{n}$. If we define a scaled variable $Y_n = \frac{\sqrt{n}}{\sigma}(M_n - \mu)$, the variables Y_n have zero mean and variance 1 for all n. We can express Y_n in terms of the partial sums $S_n = nM_n$ as

$$Y_n = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

The Central Limit Theorem states that, as n increases, the cumulative distribution functions of Y_n converge to a special form, as stated below:

Theorem 8.3 (Central Limit Theorem)
Consider a sequence of independent, identically distributed random variables $\{X_n\}$ with finite mean μ and finite variance σ^2. Denote the partial sum S_n and the partial average M_n as

$$S_n = \sum_{i=1}^{n} X_i; \; M_n = \frac{1}{n} S_n.$$

Define the new random sequence $\{Y_n\}$ as

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then, for any real. number y, the sequence of cumulative distribution functions $F_{Y_n}(y)$ converges to $\Phi(y)$, the cumulative distribution function of a standard Gaussian random variable with mean 0 and variance 1.

The surprising part of the Central Limit Theorem (CLT) is that the distribution of the individual random variables can be arbitrary. This is why Gaussian random variables are used so often in probabilistic analysis, since they approximately model sums of many independent effects. Note also the scaling used in the Central Limit Theorem: S_n has mean $n\mu$ and variance $n\sigma^2$. Hence, Y_n is measured in terms of units of standard deviation away from the mean, a similar scaling that we used when computing probabilities of Gaussian random variables.

We sketch a brief proof the the CLT by computing what are known as characteristic functions, which are the Fourier transform of the probability density functions of continuous random variables, or equivalently the Fourier transform of the of the generalized probability mass functions (expressed as the sum of $\delta(\cdot)$ functions) for discrete random variables. Since density functions integrate to 1 and probability mass functions sum to 1, the characteristic function transform will be well-defined for all $j\omega$, with $j = \sqrt{-1}$.

The characteristic function of a random variable X is

$$\Psi_W(\omega) = \mathbb{E}[e^{j\omega X}] = \begin{cases} \int_{-\infty}^{\infty} e^{j\omega x} f_X(x)\, dx & X \text{ continuous,} \\ \sum_{x_k \in R_X} e^{j\omega x_k} P(x_k) & X \text{ discrete.} \end{cases}$$

Note that

$$Y_n = \frac{1}{\sigma_X \sqrt{n}} \sum_{k=1}^{n} (X_k - \mu_x)$$

is a sum of independent, zero-mean random variables. There is a convergence result in probability called Lévy's continuity theorem, which states that, if the characteristic functions of a sequence of random variables Y_n converge pointwise as $n \to \infty$ to a function $\psi(\omega)$ which is continuous at $\omega = 0$, then the CDFs of Y_n converge pointwise to the CDF of a random variable Y with characteristic function $\psi(\omega)$. We will use this result to prove the CLT using characteristic functions.

The characteristic function of Y_n is given by:

$$\begin{aligned} \Psi_{Y_n}(\omega) = \mathbb{E}[e^{j\omega Y_n}] &= \mathbb{E}[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} \sum_{k=1}^{n} (X_k - \mu)}] \\ &= \mathbb{E}\left[\prod_{k=1}^{n} e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_k - \mu)}\right] \\ &= \prod_{k=1}^{n} \mathbb{E}\left[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_k - \mu)}\right] \quad \text{(independence)} \\ &= \left(\mathbb{E}\left[e^{j\omega \frac{1}{\sigma_X \sqrt{n}} (X_1 - \mu)}\right]\right)^n \quad \text{(identically distributed)} \end{aligned}$$

where the last equalities follows from the independent, identically distributed assumption. We expand the exponential in the expression using a Taylor series as:

$$e^{j\omega \frac{X_1 - \mu}{\sigma\sqrt{n}}} = 1 + \frac{j\omega(X_1 - \mu)}{\sigma\sqrt{n}} - \frac{\omega^2 (X_1 - \mu)^2}{2\sigma^2 n} + \ldots$$

For large n, we neglect terms beyond the first three terms to get the approximation:

$$\mathbb{E}[e^{j\omega\frac{X_1-\mu}{\sigma\sqrt{n}}}] \approx 1 + \frac{j\omega\mathbb{E}[X_1-\mu]}{\sigma\sqrt{n}} - \frac{\omega^2\mathbb{E}[(X_1-\mu)^2]}{2\sigma^2 n}$$
$$\approx 1 - \frac{\omega^2}{2n}$$

because $\mathbb{E}[X_1-\mu]=0, \mathbb{E}[(X_1-\mu)^2]=\sigma^2$. Thus,

$$\Psi_{Y_n}(\omega) \approx (1-\frac{\omega^2}{2n})^n$$

and, taking limits as $n\to\infty$, we get

$$\lim_{n\to\infty}\Psi_{Y_n}(s) = e^{-\omega^2/2}$$

Let Z be a zero mean, unit variance Gaussian random variable. Then,

$$\Psi_Z(\omega) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+j\omega z}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}+j\omega z+\frac{\omega^2}{2}-\frac{\omega^2}{2}}dz = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(z-j\omega z)^2}{2}-\frac{\omega^2}{2}}dz = e^{\frac{-\omega^2}{2}}.$$

Thus, the characteristic functions of Y_n converge for each ω to the characteristic function of a zero-mean, unit variance Gaussian random variable for all values. By Lévy's continuity theorem, this implies that the CDF of Y_n converges to the CDF of a Gaussian$(0,1)$ random variable.

The CLT implies that, given any i.i.d. sequence of random variables, we can compute probabilities of events relating to the sum of the random variables approximately using a Gaussian distribution. That is,

$$\mathbb{P}[\{(X_1+X_2+\ldots+X_n)\le a\}] \approx \Phi\left(\frac{a-n\mu}{\sqrt{n}\sigma}\right)$$

and

$$\mathbb{P}[\{\frac{1}{n}(X_1+X_2+\ldots+X_n)\le b\}] \approx \Phi\left(\frac{b-\mu}{\frac{\sigma}{\sqrt{n}}}\right).$$

As a rule of thumb, these approximation are very accurate as long as $\frac{|a-n\mu_X|}{\sqrt{n}\sigma}$ is less than 3.

Example 8.9

Assume we have a disk drive that takes X milliseconds for each disk access time, where X is a random variable, uniformly distributed in [0,12]. Assume one must access disk 12 times independently, and define the total access time $T = X_1 + \ldots + X_{12}$. Then, $\mathbb{E}[T]12\mathbb{E}[X] = 72$ msec, and $\mathsf{Var}[T] = 12\mathsf{Var}[X] = 12\cdot\frac{12^2}{12} = 144$. Therefore, the standard deviation of the sum is 12. We want to compute the probability that the total wait time is greater than 75 seconds.

We approximate this with the CLT, since T is the sum of i.i.d. random variables.

$$\mathbb{P}[T>75] = 1 - F_T(75) \approx 1 - \Phi\left(\frac{75-72}{12}\right) = 1-\Phi(0.25) = Q(0.25).$$

What about the probability that $T<48$? This is

$$F_T(48) \approx \Phi(\frac{48-72}{12}) = \Phi(-2) = Q(2).$$

Note that, to compute this exactly, we would need the probability density of T, which would require performing 12 convolutions.

Example 8.10

A Modem transmits 10^4 bits, where each bit is i.i.d. with probability $p=0.5$. We would like to estimate the probability that we get more that 5100 one bits. We also want to estimate the probability that the number of one bits we receive is in the interval (4900,5100].

The total number of one bits received, T is the sum of 10^4 independent Bernoulli random variables. We know this is a Binomial $(10^4, 0.5)$ random variable, but computing the quantities asked involve summing between 100 and 200 binomial terms. We approximate this using. the CLT as follows:

$$\mathbb{E}[T] = 10^4 p = 5000; \quad \mathsf{Var}[T] = 10^4 p(1-p) = 2500; \quad \sigma_T = 50.$$

With this approximation, we quickly estimate

$$\mathbb{P}[\{T > 5100\}] = 1 - F_T(5100) \approx 1 - \Phi(\frac{5100 - 5000}{50}) = 1 - \Phi(2) = Q(2).$$

$$\mathbb{P}[\{T \in (4900, 5100]\}] = F_T(5100) - F_T(4900) \approx \Phi(2) - \Phi(-2) = \Phi(2) - Q(2).$$