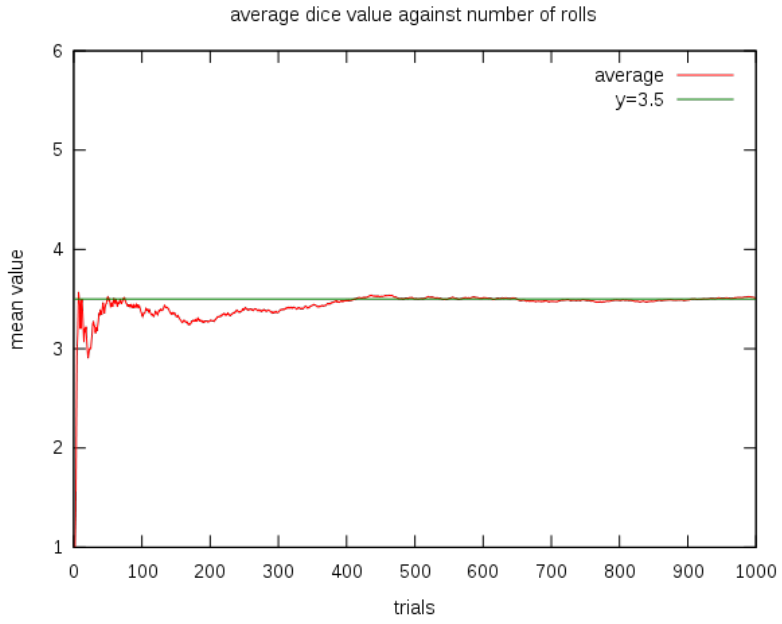

Figure 1.8: When rolling a balanced die, the average value converges to 3.5. From Wikipedia.

Thus, the sample mean values $Y(n) := (X(1) + \cdots + X(n))/n$ converge to the expected value, with probability 1. Note that the sample mean values $Y(n)$ are random variables: for each n, the value of $Y(n)$ depends on the particular realization of the random variables $X(m)$; if you repeat the experiment, the values will probably be different. However, the limit is always μ, with probability 1. We say that the convergence is *almost sure*.

Figure 1.9: Jacob Bernoulli. 1655-1705.

Almost Sure

THE NOTION OF ALMOST SURE CONVERGENCE is quite subtle. For instance, let $X(n) \in \{0,1\}$ be the outcome of the n-th flip of a fair coin. The sequence of coin flips that all yield 1 has probability zero. Indeed, the probability that the first m coin flips all yield 1 is 2^{-m}, so the probability that this hold for all m is zero.[9] Similarly, the SLLN says that the probability that the fraction of 1's remains less than 0.45 as we keep flipping coins is also zero. Let

$$A = \{\mathbf{x} = (x_1, x_2, \ldots) \in \{0,1\}^\infty | \lim_{n\to\infty} \frac{x_1 + \cdots + x_n}{n} = 0.5\}.$$

That is, A is the collection of sequences of outcomes of coin flips such that the fraction of 1's converges to 0.5 as one flips more and more coins. There are many sequences of 0's and 1's that do not have that

[9] Let's try to be precise. Define $A_n = \{\omega | X_m(\omega) = 1, m = 1, \ldots, n\}$ and $A = \{\omega | X_m(\omega) = 1, m \geq 1\}$. Then $A_1 \supset A_2 \supset A_3 \supset \ldots$ and $\cap_n A_n = A$. Hence $P(A_n) \to P(A)$. (See Section A.1.) But, $P(A_n) = 2^{-n}$. Thus, $P(A) = 0$.

property. However, the SLLN says that

$$P(\mathbf{X} = (X_1, X_2, \ldots) \in A) = 1.$$

That is, the sequence $\mathbf{X}$ of coin flips that one observes is such that that the fraction of 1's converges to 0.5, with probability 1.

This result is the basis for the very intuitive *frequentist* interpretation of probability. For instance, flip a balanced die repeatedly and let $X(n)$ be the outcome in $\{1, 2, \ldots, 6\}$ of the n-roll. Then, $Y(n)$ is the average value of the first n rolls and the SLLN says that this fraction converges to $\mu = 3.5$, as illustrated in Figure 1.8.

Convergence of Expected Value

Assume that X_n and X are such that $X_n \to X$ with probability one. Is it the case that $E(X_n) \to X$? It turns out that the answer is negative, in general. As a simple example, let $Z =_D U[0,1]$ and define $X_n = n.1\{Z \leq 1/n\}$ for $n \geq 1$ and $X = 0$. Observe that $X_n = 0$ if $n > 1/Z$. Thus, $X_n = 0$ for n large enough, so that $X_n \to 0$, with probability one. Also, X_n is equal to n with probability $1/n$ (when $Z \leq 1/n$) and equal to zero otherwise, so that $E(X_n) = 1$, for all n. (See Figure 1.10.)

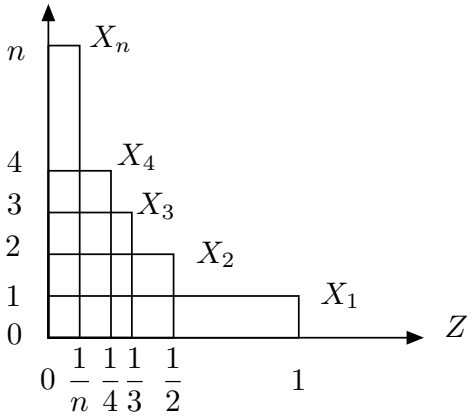

Figure 1.10: Random variables X_n such that $X_n \to 0$ and $E(X_n) \to 1$.

Thus, we have

$$X_n \to X \text{ w.p. } 1 \text{ and } E(X_n) = 1 \not\to E(X) = 0.$$

We say that, in general, one can not interchange the expectation and the limit. Thus, in general,

$$\lim_{n\to\infty} E(X_n) \neq E(\lim_{n\to\infty} X_n).$$

In our previous example, the left-hand side is 1 and the right-hand side is 0.

There are two cases when the limit of the expected values is the expected value of the limit. We state these as a theorem that we do not prove here, but that we use later.

Theorem 1.4 *DCT and MCT*

Let X_n and X be such that $X_n \to X$, with probability one.

(a) Dominated Convergence Theorem (DCT):

$$\text{If } |X_n| \leq Y \text{ with } E(Y) < \infty, \text{ then } E(X_n) \to E(X);$$

(b) Monotone Convergence Theorem (MCT):

$$\text{If } 0 \leq X_n \leq X_{n+1}, \forall n, \text{ then } E(X_n) \to E(X).$$

Note that the condition of MCT or of DCT are not satisfied by our example.

As an example of *MCT*, let $X = \sum_{n=0}^{\infty} 1_{A_n}$ for some events A_n. Then

$$E(X) = \sum_n E(1_{A_n}) = \sum_{n=0}^{\infty} P(A_n).$$

In particular,

$$\text{if } \sum_{n=0}^{\infty} P(A_n) < \infty, \text{ then } P(X < \infty) = 1,$$

so that ω belongs only to finitely many A_n, with probability one. This is the Borel-Cantelli Lemma that we proved in Theorem A.1 using the continuity of probability.

Weak Law of Large Numbers (WLLN)

THE SLLN IS NOT THAT SIMPLE TO PROVE. However, a weaker result, the *Weak Law of Large Numbers (WLLN)* is easy to show and provides some intuition. We prove the SLLN later in this section.

Before we explain the WLLN, we need a brief detour to review some simple inequalities. Recall that the *variance* of a random variable X is defined as

$$\text{var}(X) = E((X - E(X))^2).$$

Thus, the variance measures how far the random variable is away from its mean, on average. This is a measure of uncertainty, or variability across different realizations. Note that[10]

$$\text{var}(X) = E(X^2) - E(X)^2.$$

Chebyshev's inequality provides another indication that the variance measures the deviation from the mean.

This inequality is as follows.

Theorem 1.5 *Chebyshev's inequality* *Let X be a random variable with $E(X) = \mu$ and $var(X) < \infty$. Then*

$$P(|X - \mu| \geq \epsilon) \leq \frac{var(X)}{\epsilon^2}. \tag{1.10}$$

To see why this result holds, note that

$$1\{|X - \mu| \geq \epsilon\} \leq \frac{|X - \mu|^2}{\epsilon^2}, \tag{1.11}$$

as you can verify by considering the cases $|X - \mu| \geq \epsilon$ and $|X - \mu| < \epsilon$ separately. Now recall that if two random variables V and W are

[10] Indeed, one has

$$(X - E(X))^2 = X^2 - 2XE(X) + E(X)^2$$

and the expected value is

$$E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2) - E(X)^2.$$

Figure 1.11: Pafnuty Chebyshev. 1821-1884.

such that $V \leq W$, then $E(V) \leq E(W)$. Thus, taking expectations in (1.11) gives (1.10). (We used the fact that $E(1_A) = P(A)$ since the indicator 1_A of the event A is a random variable that is equal to one with probability $P(A)$ and to zero otherwise.)

We need a definition.

Definition 1.6 *Convergence in Probability*

Let $X_n, n \geq 1$ and X be random variables defined on a common probability space. One says that X_n converges in probability to X, and one writes $X_n \xrightarrow{p} X$ if, for all $\epsilon > 0$,

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

The WLLN is the following result.[11]

[11] You probably wonder why this is called 'weak' whereas the other result is called 'strong.' The reason is that the almost sure convergence in the SLLN implies the convergence in probability (1.12). See Problem 17.

Theorem 1.7 *Weak Law of Large Numbers*

Let $\{X(n), n \geq 1\}$ be a sequence of i.i.d. random variables with mean μ. Then

$$Y_n \xrightarrow{p} \mu. \tag{1.12}$$

In the coin flipping example, this says that if we flip a fair coin many times, the fraction of heads is very unlikely to be far from 50%.

Proof:

To get the WLLN, we use Chebyshev's inequality to get

$$P(|Y(n) - \mu| \geq \epsilon) \leq \frac{\text{var}(Y(n))}{\epsilon^2}.$$

(We used the fact that the variance of a sum of independent random variables is equal to the sum of the variances. See Appendix A.) Now,

$$\text{var}(Y(n)) = \frac{\text{var}(X(1) + \cdots + X(n))}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Hence, $\text{var}(Y_n) \to 0$, and this concludes the proof of the WLLN. ♣

Figure 1.12 illustrates the SLLN and WLLN. The SLLN states that the sample means of i.i.d. random variables converge to the mean, with probability one. The WLLN says that as the number of samples increases, the fraction of realizations where the sample mean differs from the mean by some amount gets small.

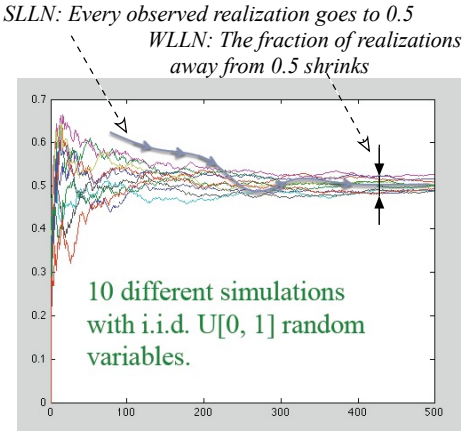

Figure 1.12: SLLN and WLLN for i.i.d. $U[0,1]$ random variables.

Strong Law of Large Numbers (SLLN)

We prove the SLLN, i.e., Theorem 1.3, under a stronger than necessary assumption that the fourth moment of the random variables is finite. That is, we assume that $E(X_n^4) < \infty$. To simplify the notation, let us first assume that $E(X_n) = 0$. Note that $X^2 < 1 + X^4$ since this

result is obvious for $X^2 < 1$ and since $X^4 > X^2$ when $X^2 > 1$. Similarly, $|X|^3 < 1 + X^4$. Hence, $E(X^2) < \infty$ and $E(|X|^3) < \infty$. We want to show that

$$\frac{X_1 + \cdots + X_n}{n} \to 0, \text{ as } n \to \infty,$$

with probability one.

The staring point is similar to Chebyshev's inequality. We note that[12]

$$P(|\frac{X_1 + \cdots + X_n}{n}| > \epsilon) \leq \frac{E((X_1 + \cdots + X_n)^4)}{\epsilon^4 n^4}.$$

[12] Observe that we use the fourth power instead of the second. The reason for this choice is that it results in a bound of order β/n^2, which is summable over n. This fact enables us to use Borel-Cantelli. The second power gives a bound of order β/n that is not summable.

Now,

$$E((X_1 + \cdots + X_n)^4) = E(\sum_{m=1}^{n} X_m^4 + \sum_{(k,m)\in A} X_k^2 X_m^2 + \sum_{(k,m,j)\in B} X_k^2 X_m X_j + \sum_{(k,m)\in A} X_k^3 X_m + \sum_{(k,m,i,j)\in C} X_k X_m X_i X_j)$$

where A is the set of pairs of distinct elements in $\{1, \ldots, n\}$, B is the set of triplets of different elements in that set and C is the set of quadruplets of different elements in the same set.

Since the X_m are zero-mean and independent, the mean values of terms like $X_i^2 X_j X_m$, $X_k^3 X_m$ and $X_k X_m X_i X_j$ are equal to zero. Hence,

$$E((X_1 + \cdots + X_n)^4) = nE(X_1^4) + n(n-1)E(X_1^2)^2 \leq \alpha n^2$$

where $\alpha = \max\{E(X_1^4) - E(X_1^2)^2, E(X_1^2)^2\}$.

We conclude that

$$\sum_n P(|\frac{X_1 + \cdots + X_n}{n}| > \epsilon) < \infty.$$

Using the Borel-Cantelli Lemma A.1, we have that

$$P(|\frac{X_1 + \cdots + X_n}{n}| > \epsilon, \text{ i.o.}) = 0.$$

This fact allows us to conclude that for almost all realizations there is some finite n_0 such that

$$|\frac{X_1 + \cdots + X_n}{n}| \leq \epsilon, \forall n \geq n_0.$$

Since $\epsilon > 0$ is arbitrary, this proves that

$$\frac{X_1 + \cdots + X_n}{n} \to 0, \text{ as } n \to \infty,$$

with probability one.

THE CASE WHEN $E(X_n) \neq 0$ is easily reduced to the previous case by subtracting the mean values. We leave that verification to the readers.

1.7 *Appendix: Proof of Big Theorem

For completeness, this section presents the proofs of the more difficult results of this chapter.[13]

[13] The sections marked $*$ can be omitted without loss of continuity.

Proof of Theorem 1.2 (a)

Let m_j be the expected return time to state j. That is,

$$m_j = E[T_j | X(0) = j] \text{ with } T_j = \min\{n > 0 | X(n) = j\}.$$

We show that $\pi(j) = 1/m_j, j = 1, \ldots, N$ is the unique invariant distribution if the Markov chain is irreducible.

During $n = 1, \ldots, N$ where $N \gg 1$, the Markov chain visits state j a fraction $1/m_j$ of the times. A fraction $P(j,i)$ of those times, it visits state i just after visiting state j. Thus, a fraction $(1/m_j)P(j,i)$ of the times, the Markov chain visits j then i in successive steps. By summing over j, we find the fraction of the times that the Markov chain visits i. Thus,

$$\sum_j \frac{1}{m_j} P(j,i) = \frac{1}{m_i}.$$

Hence, there is one invariant distribution π and it is given by $\pi_i = 1/m_i$, which is the fraction of time that the Markov chain spends in state i.

To show that the invariant distribution is unique, assume that there is another one, say $\phi(i)$. Start the Markov chain with that distribution. Then

$$\frac{1}{N} \sum_{n=0}^{N-1} 1\{X(n) = i\} \to \pi(i).$$

However, taking expectation, we find that the left-hand side is equal to $\phi(i)$. Thus, $\phi = \pi$ and the invariant distribution is unique.[14]

[14] Indeed, $E(\{X(n) = i\}) = P(X(n) = i) = \phi(i)$.

Proof of Theorem 1.2 (b)

If the Markov chain is irreducible but not aperiodic, then π_n may not converge to the invariant distribution π. For instance, if the Markov chain alternates between 0 and 1 and starts from 0, then $\pi_n = [1, 0]$ for n even and $\pi_n = [0, 1]$ for n odd, so that π_n does not converge to $\pi = [0.5, 0.5]$.

If the Markov chain is aperiodic, $\pi_n \to \pi$. Moreover, the convergence is geometric. We fist illustrate the argument on a simple example shown in Figure 1.13.

Consider the number of steps to go from 1 to 1. Note that

$$\{n > 0 | P^n(1,1) > 0\} = \{3, 4, 6, 7, 8, 9, 10, \ldots\}.$$

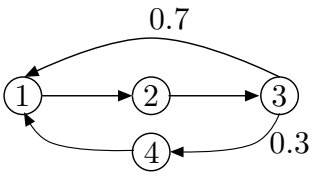

Figure 1.13: An aperiodic Markov chain.

Thus, $P^n(1,1) > 0$ if $n \geq 6$. Now, $P[X(2) = 1|X(0) = 2] > 0$, so that $P[X(n) = 1|X(0) = 2] > 0$ for $n \geq 8$. Indeed, if $n \geq 8$, then X can go from 2 to 1 in two steps and then from 1 to 1 in $n-2$ steps. The argument is similar for the other states and we find that there is some $M > 0$ and some $p > 0$ such that

$$P[X(M) = 1|X(0) = i] \geq p, i = 1, 2, 3, 4.$$

Now, consider two copies of the Markov chain: $\{X(n), n \geq 0\}$ and $\{Y(n), n \geq 0\}$. One chooses $X(0)$ with distribution π_0 and $Y(0)$ with the invariant distribution π. The two Markov chains evolve independently initially. We define

$$\tau = \min\{n > 0|X(n) = Y(n)\}.$$

In view of the observation above,

$$P(X(M) = 1 \text{ and } Y(M) = 1) \geq p^2.$$

Thus, $P(\tau > M) \leq p^2$. If $\tau > M$, then the Markov chain have not met yet by time M. Using the same argument as before, we see that they have a probability at least p^2 of meeting in the next M steps. Thus,

$$P(\tau > kM) \leq p^{2k}.$$

Now, modify $X(n)$ by gluing it to $Y(n)$ after time τ. This coupling operation does not change the fact that $X(n)$ still evolves according to the transition matrix P, so that $P(X(n) = i) = \pi_n(i)$ where $\pi_n = \pi_0 P^n$.

Now,

$$\sum_i |P(X(n) = i) - P(Y(n) = i)| \leq P(X(n) \neq Y(n)) \leq P(\tau > n).$$

Hence,

$$\sum_i |\pi_n(i) - \pi(i)| \leq P(\tau > n),$$

and this implies that

$$\sum_i |\pi_n(i) - \pi(i)| \leq p^{2k} \text{ if } n > kM.$$

To extend this argument to a general aperiodic Markov chain, we need the fact that for each state i there is some integer n_i such that $P^n(i,i) > 0$ for all $n \geq n_i$. We prove that fact as Lemma 1.10 in the following section.

Periodicity

We start with a property of the set of return times of an irreducible Markov chain.

Lemma 1.8 *Fix a state i and let $S := \{n > 0 | P^n(i,i) > 0\}$ and $d = g.c.d.(S)$. There must be two integers n and $n+d$ in the set S.*

Proof:

The trick is clever. We first illustrate it on an example. Assume $S = \{9, 15, 21, \ldots\}$ with $d = g.c.d.(S) = 3$. There must be $a, b \in S$ with $g.c.d.\{a, b\} = 3$. Otherwise, the gcd of S would not be 3. Here, we can choose $a = 15$ and $b = 21$. Now, consider the following operations:

$$(a,b) = (15,21) \to (6,15) \to (6,9) \to (3,6) \to (3,3).$$

At each step, we go from (x,y) with $x \leq y$ to the ordered pair of $\{x, y-x\}$. Note that at each step, each term in the pair (x,y) is an integer linear combination of a and b. For instance, $(6,15) = (b-a, a)$. Then, $(6,9) = (b-a, a-(b-a)) = (b-a, 2a-b)$, and so on. Eventually, we must get to $(3,3)$. Indeed, the terms are always decreasing until we get to zero. Assume we get to (x,x) with $x \neq 3$. At the previous step, we had $(x, 2x)$. The step before must have been $(x, 3x)$, and so on. Going back all the way to (a,b), we see that a and b are both multiples of x. But then, $g.c.d.\{a,b\} = x$, a contradiction.

From this construction, since at each step the terms are integer linear combinations of a and b, we see that

$$3 = ma + nb$$

for some integers m and n. Thus,

$$3 = m^+a + n^+b - m^-a - n^-b$$

where $m^+ = \max\{m, 0\}$ and $m = m^+ - m^-$, and similarly for n^+ and n^-. Now we can choose

$$N = m^-a + n^-b \text{ and } N + 3 = m^+a + n^+b.$$

The last step of the argument is to notice that if $a, b \in S$, then $\alpha a + \beta b \in S$ for any integers α and β that are not both zero. This fact follows from the definition of S as the return times from i to i. Hence, both N and $N+3$ are in S.

The proof for a general set S with gcd equal to d is identical. ♣

This result enables us to show that the period of a Markov chain is well-defined.

Lemma 1.9 *For an irreducible Markov chain, $d(i)$ defined in (1.7) has the same value for all states.*

Proof: Pick $j \neq i$. We show that $d(j) \leq d(i)$. This suffices to prove the lemma, since by symmetry one also has $d(i) \leq d(j)$.

By irreducibility, $P^m(j,i) > 0$ for some m and $P^n(i,j) > 0$ for some n. Now, by definition of $d(i)$ and by the previous lemma, there is some integer N such that $P^N(i,i) > 0$ and $P^{N+d(i)}(i,i) > 0$. But then,

$$P^{m+N+n}(j,j) > 0 \text{ and } P^{m+N+d(i)+n}(j,j) > 0.$$

This implies that the integers $K := n + N + m$ and $K + d(i)$ are both in $S := \{n > 0 | P^n(j,j) > 0\}$. Clearly, this shows that

$$d(j) := g.c.d.(S) \leq d(i).$$

♣

The following fact is then crucial for our prove of convergence.

Lemma 1.10 *Let X be an irreducible aperiodic Markov chain. Let $S = \{n > 0 | P^n(i,i) > 0\}$. Then, there is some n_i such that $n \in S$, for all $n \geq n_i$.*

Proof:

We know from Lemma 1.8 that there is some integer N such that $N, N+1 \in S$. We claim that

$$n \in S, \forall n > N^2.$$

To see this, first note that for $m > N - 1$ one has

$$\begin{aligned}
&mN + 0 = mN,\\
&mN + 1 = (m-1)N + (N+1),\\
&mN + 2 = (m-2)N + 2(N+1),\\
&\ldots,\\
&mN + N - 1 = (m - N + 1)N + (N-1)(N+1).
\end{aligned}$$

Now, for $n > N^2$ one can write

$$n = mN + k$$

for some $k \in \{0, 1, \ldots, N-1\}$ and $m > N - 1$. Thus, n is an integer linear combination of N and $N+1$ that are both in S, so that $n \in S$.

♣

1.8 Summary

The key notions and results are as follows:

- Markov Chains: states, transition probabilities, irreducible, aperiodic, invariant distribution;
- Big Theorem: irreducible implies unique invariant distribution equal to the long term fraction of time in the states; convergence to invariant distribution if irreducible and aperiodic;