Contents

1. Lindeberg-Feller CLT ................................................................. 1
2. Examples for Lindeberg-Feller CLT ............................................. 4
3. Non-Gaussian limit theory ........................................................... 7
4. Stable laws and Convergence of types .......................................... 12
5. Infinite divisibility ...................................................................... 15
6. Conditional expectation ............................................................... 20
7. Conditional distribution ............................................................... 26
8. Introduction to martingales .......................................................... 28
9. Optional Stopping Theorem .......................................................... 31
10. Martingale Convergence Theorem ............................................... 34
11. Uniform Integrability ................................................................. 37
12. UI Martingales .......................................................................... 40
13. Exchangeability ......................................................................... 43
14. $L^2$-martingales and some inequalities ........................................ 46
15. Markov Chains: definitions and examples ..................................... 50
16. Stationary and reversible measures .............................................. 54
17. Existence/uniqueness of stationary measures ................................. 57
18. Strong Markov property and density of returns .............................. 60
19. Convergence to equilibrium ......................................................... 65

Note: Editing in progress, changes to be expected at any time!
1. Lindeberg-Feller CLT

In 275A we saw the proof of the following Central Limit Theorem:

**Theorem 1.1 (CLT)** Let \( X_1, X_2, \ldots \) be i.i.d. with \( E(X_1^2) < \infty \). Denote \( \mu := E(X_1) \) and \( \sigma^2 := \text{Var}(X_1) \) and set \( S_n := X_1 + \cdots + X_n \). Then

\[
\frac{S_n - n\mu}{\sqrt{n}} \xrightarrow{\text{law}} \mathcal{N}(0, \sigma^2). \tag{1.1}
\]

This theorem is beautiful but its conditions may sometimes be a bit too restrictive in applications. Here are two natural reasons for generalization:

1. There is no need to work with an infinite sequence of random variables; it should suffice to have independent random variables for each \( n \).
2. The random variables entering the sum \( S_n \) may not need to be identically distributed, just independent is enough.

These naturally lead to the consideration of triangular arrays, which are collections of random variables \( \{X_{n,m}\}_{1 \leq m \leq n} \), where independence is assumed only for each fixed \( n \). For these we have:

**Theorem 1.2 (Lindeberg-Feller CLT)** Let \( \{X_{n,m}\}_{1 \leq m \leq n} \) be independent for each \( n \) and such that \( E(X_{n,m}) = 0 \) and \( E(X_{n,m}^2) < \infty \). Set \( S_n := \sum_{m=1}^n X_{n,m} \) and suppose that the following two “Lindeberg-Feller conditions” hold:

1. There is \( \sigma^2 \in [0, \infty) \) such that
   \[
   \sum_{m=1}^n E(X_{n,m}^2) \xrightarrow{n \to \infty} \sigma^2. \tag{1.2}
   \]
2. For each \( \varepsilon > 0 \),
   \[
   \sum_{m=1}^n E(X_{n,m}^2 1_{|X_{n,m}| > \varepsilon}) \xrightarrow{n \to \infty} 0. \tag{1.3}
   \]

Then \( S_n \xrightarrow{\text{law}} \mathcal{N}(0, \sigma^2) \).

Before we delve into the proof, let us check that this implies the above Theorem 1.1. If \( X_1, X_2, \ldots \) are i.i.d. with the stated properties, we will naturally take \( X_{n,m} := \frac{1}{\sqrt{n}}(X_m - \mu) \). Then \( E(X_{n,m}^2) = \sigma^2/n \) and (LF1) holds trivially while for (LF2) we get to check that

\[
E(X_{n,1}^2 1_{|X_1| > \varepsilon \sqrt{n}}) \xrightarrow{n \to \infty} 0
\]

for any \( \varepsilon > 0 \). This follows by the Dominated Convergence Theorem and the fact that \( E(X_1^2) < \infty \).

**Proof of Theorem 1.2.** The proof will ultimately reduce to similar arguments as for the CLT above, but we have to impose some truncation first. For this we note that (LF2) implies

\[
\exists \varepsilon_n \downarrow 0 \text{ such that } K_n := \frac{1}{\varepsilon_n^2} \sum_{m=1}^n E(X_{n,m}^2 1_{|X_{n,m}| > \varepsilon_n}) \xrightarrow{n \to \infty} 0. \tag{1.5}
\]
Now we define $X_{n,m} := X_{n,m}1_{|X_{n,m}| \leq \varepsilon_n}$ and set $S_n := \sum_{m=1}^{n} X_{n,m}$. Since
\[ P(|S_n - \overline{S}_n| > \varepsilon_n) \leq P(\exists m: |X_{n,m}| > \varepsilon_n) \leq \sum_{m=1}^{n} P(|X_{n,m}| > \varepsilon_n) \leq \frac{1}{\varepsilon_n^2} \sum_{m=1}^{n} E(X_{n,m}^2 1_{|X_{n,m}| > \varepsilon_n}) \tag{1.6} \]
by Chebyshev’s inequality, and $\varepsilon_n \downarrow 0$, the random variables $S_n$ and $\overline{S}_n$ have the same distributional limit, provided one of the limits exists. So we may focus on proving this for $\overline{S}_n$ instead of $S_n$. For later use we also observe that by $E(S_n) = 0$, the triangle inequality and Chebyshev again,
\[ |E\overline{S}_n| = |E(\overline{S}_n - S_n)| \leq \sum_{m=1}^{n} E(|X_{n,m} - \overline{X}_{n,m}|) \leq \frac{1}{\varepsilon_n} \sum_{m=1}^{n} E(X_{n,m}^2 1_{|X_{n,m}| > \varepsilon_n}) \tag{1.7} \]
and so, invoking also (LF1),
\[ E\overline{S}_n \xrightarrow{n \to \infty} 0 \quad \text{and} \quad \text{Var}(\overline{S}_n) \xrightarrow{n \to \infty} \sigma^2 \tag{1.8} \]
as well. We can now introduce $\phi_{n,m}(t) := E(e^{it(X_{n,m} - \overline{X}_{n,m})})$ and note that
\[ E(e^{it(\overline{S}_n - E\overline{S}_n)}) = \prod_{m=1}^{n} \phi_{n,m}(t). \tag{1.9} \]
In light of the first part of (1.8), it suffices to show that this converges to $e^{-\frac{t^2}{2}\sigma^2}$ for this we have to deal with the product of the characteristic functions. A natural instinct is to take a logarithm and convert the product into a sum, but this is hindered by the fact that the quantities $\phi_{n,m}(t)$ are in general complex-valued and so one has to be particularly careful about working with the complex logarithm. We will overcome this by replacing $\phi_{n,m}(t)$ by its second-order Taylor polynomial, $1 - \frac{t^2}{2} \text{Var}(X_{n,m})$, which is blatantly real valued. To see what kind of error estimate we may need for this, we first note a general bound:

**Lemma 1.3** Let $z_1, \ldots, z_n, w_1, \ldots, w_n \in \mathbb{C} : |z| \leq 1$. Then
\[ \left| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \right| \leq \sum_{i=1}^{n} |z_i - w_i|. \tag{1.10} \]

**Proof.** This holds trivially for $n = 1$ and for $n > 1$ is proved by applying the induction step
\[ \left| \prod_{i=1}^{n} z_i - \prod_{i=1}^{n} w_i \right| \leq |z_1| \left| \prod_{i=2}^{n} z_i - \prod_{i=2}^{n} w_i \right| + |z_1 - w_1| \left| \prod_{i=1}^{n} z_i \right| \tag{1.11} \]
and bounding both $|z_1|$ and the last product over $z_i$ by one. □

The error bound will then be provided by:
Lemma 1.4
\[ \sum_{m=1}^{n} \left| \phi_{n,m}(t) - \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) \right| \xrightarrow{n \to \infty} 0 \quad (1.12) \]

Proof. Abbreviate \( \hat{X}_{n,m} := X_{n,m} - EX_{n,m} \) and note that \( \text{Var}(X_{n,m}) = E(\hat{X}_{n,m}^2) \). By Taylor’s theorem we have
\[ \phi_{n,m}(t) = 1 + i0 - t^2 \int_0^1 uE(\hat{X}_{n,m}^2 e^{i(1-u)\hat{X}_{n,m}}) \, du \quad (1.13) \]
Hence,
\[ \phi_{n,m}(t) - \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) = t^2 \int_0^1 uE(\hat{X}_{n,m}^2(1 - e^{i(1-u)\hat{X}_{n,m}})) \, du \quad (1.14) \]
A key point now, and the main reason why we had to truncate first, is that \(|\hat{X}_{n,m}| \leq 2\varepsilon_n\) implies the uniform bound
\[ |1 - e^{i(1-u)\hat{X}_{n,m}}| = 2|\sin((1-u)\hat{X}_{n,m})| \leq \max_{0 \leq x \leq |\varepsilon_n|} 2\sin(x) \quad (1.15) \]
which only depends on \( n \) but not on \( m \). Hence we can apply the absolute value to (1.14) and sum the result over \( m \) to get
\[ \sum_{m=1}^{n} \left| \phi_{n,m}(t) - \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) \right| \leq \frac{1}{2} t^2 \left( \max_{0 \leq x \leq |\varepsilon_n|} 2\sin(x) \right) \sum_{m=1}^{n} E(\hat{X}_{n,m}^2). \quad (1.16) \]
Since \( E(\hat{X}_{n,m}^2) = \text{Var}(\mathcal{X}_{n,m}) \leq E(X_{n,m}^2) \leq E(X_{n,m}^2) \), (LF1) ensures the sum on the right is bounded in \( n \). The maximum goes to zero and so the result follows.

To conclude the proof of the theorem, we now notice that \( \text{Var}(\mathcal{X}_{n,m}) \leq \varepsilon_n^2 \) and so once \( \delta_n := \frac{1}{2} t^2 \varepsilon_n^2 < 1 \) we are permitted to write
\[ \prod_{m=1}^{n} \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) = \exp\left\{ \sum_{m=1}^{n} \log \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) \right\}. \quad (1.17) \]
The bound
\[ |\log(1+z) - z| \leq \frac{|z|^2}{1 - |z|}, \quad |z| < 1, \quad (1.18) \]
then yields
\[ \left| \sum_{m=1}^{n} \log \left( 1 - \frac{t^2}{2} \text{Var}(\mathcal{X}_{n,m}) \right) + \frac{t^2}{2} \sum_{m=1}^{n} \text{Var}(\mathcal{X}_{n,m}) \right| \leq \frac{\delta_n}{1 - \delta_n} \sum_{m=1}^{n} \text{Var}(\mathcal{X}_{n,m}) \quad (1.19) \]
But the sum of the variances is \( \text{Var}(\mathcal{S}_n) \), which by (1.8) tends to \( \sigma^2 \). As \( \delta_n \downarrow 0, \) the two lemmas then imply
\[ E(e^{it(\mathcal{S}_n - E\mathcal{S}_n)}) \xrightarrow{n \to \infty} e^{-\frac{1}{2} \sigma^2 t^2}, \quad t \in \mathbb{R}. \quad (1.20) \]
Since \( E\mathcal{S}_n \to 0 \), we are done.
2. Examples for Lindeberg-Feller CLT

We will now proceed to discuss some examples where Theorem 1.2 is relevant. In fact, we will use a slightly different form, which is originally due to Lindeberg:

**Theorem 2.1** Let \( k_n \) be a sequence of integers such that \( k_n \to \infty \) and let \( \{X_{n,m} : 1 \leq m \leq k_n\} \) be independent for each \( n \). Let \( S_n := \sum_{m=1}^{k_n} X_{n,m} \) and suppose that for each \( \varepsilon > 0 \),

\[
\frac{1}{\text{Var}(S_n)} \sum_{m=1}^{k_n} E\left(X_{n,m}^2 \mathbb{1}_{|X_{n,m}| > \varepsilon \sqrt{\text{Var}(S_n)}}\right) \to 0. \tag{2.1}
\]

Then

\[
\frac{S_n}{\sqrt{\text{Var}(S_n)}} \overset{\text{law}}{\to} N(0, 1). \tag{2.2}
\]

**Proof.** It suffices to notice that the proof of Theorem 1.2 never used that the array was exactly triangular. Relabeling \( X_{n,m}/\sqrt{\text{Var}(S_n)} \) into \( X_{n,m} \), (LF1) then holds trivially with \( \sigma^2 = 1 \) while (LF2) is a rewrite of (2.1). \( \square \)

**CLT with non-standard scaling:** Our first example will concern the situation when the CLT fails, but just barely. Let \( X_1, X_2, \ldots \) be i.i.d. with the following properties

\[
X_1 \overset{\text{law}}{=} -X_1 \quad \text{and} \quad P(|X_1| > x) = x^{-2} \quad \text{for} \quad x > 1. \tag{2.3}
\]

In this case \( E(X_1^2) = \infty \) and so the CLT does not apply. However, if we use a truncation, we can still prove a Gaussian limit theorem albeit under a different scaling.

Consider a positive sequence \( a_n \) and use it to define the random variables

\[
X_{n,m} := (X_m \wedge a_n) \vee (-a_n). \tag{2.4}
\]

Their sum will be denoted by

\[
\overline{S}_n := \sum_{m=1}^{n} X_{n,m}. \tag{2.5}
\]

Since \( EX_{n,m} = 0 \) due to the symmetry of \( X_{n,m} \) and the symmetric nature of the truncation, a computation tells us

\[
\text{Var}(\overline{S}_n) = n \text{Var}(X_{n,1}) = nE(X_{n,1}^2)
\]

\[
= n \int_{1}^{a_n} 2xP(|X_1| \leq x)dx
\]

\[
= n \int_{1}^{a_n} \frac{2}{x}dx
\]

\[
= 2n \log a_n. \tag{2.6}
\]

Now assume that \( a_n \) is such that

\[
\frac{a_n}{\sqrt{\text{Var}(\overline{S}_n)}} \to 0. \tag{2.7}
\]
Then \( |X_{n,m}| < \varepsilon \sqrt{\text{Var}(S_n)} \) with probability one once \( n \) is large enough and so the (only) condition of the above theorem applies. Hence we get

\[
\frac{\overline{S}_n}{\sqrt{\text{Var}(S_n)}} \xrightarrow{\text{law}} N(0,1) \quad (2.8)
\]

The issue is now to extract a meaningful statement for \( S_n := X_1 + \cdots + X_n \). For this we need to ensure that \( \overline{S}_n \) can be replaced by \( S_n \), which will happen e.g. once \( \max_{k \leq n} |X_k| \leq a_n \). By the union bound, the complementary event has probability

\[
nP(|X_1| > a_n) = \frac{n}{a_n^2} \quad (2.9)
\]

Hence, if \( a_n^2/n \to \infty \) while simultaneously (2.17) holds, we have a limit statement for \( S_n \) as well. Setting \( a_n := \sqrt{n \log \log n} \), both of these conditions hold whereby we get

\[
\frac{S_n}{\sqrt{n \log n}} \xrightarrow{\text{law}} N(0,1) \quad (2.10)
\]

There is a Gaussian limit, but one has to scale by \( \sqrt{n \log n} \) instead of just \( \sqrt{n} \).

**Coupon collector problem:** Our other example concerns the Coupon Collector Problem in which the CLT actually fails because the second Lindeberg-Feller condition is not satisfied.

Let \( X_1, X_2, \ldots \) be i.i.d. uniform on \( \{1, 2, \ldots, n\} \). As the values of the sequence are revealed, new not-yet-seen values keep coming up. The times when this happens are defined as follows: Set \( T_0 := 0 \) and, for \( i = 1, \ldots, n-1 \),

\[
T_{i+1} := \inf\{m > T_i : |\{X_1, \ldots, X_m\}| > i\} \quad (2.11)
\]

The objective is to describe the statistics of \( T_n \), i.e., the first “time” when all elements in the set of possible values have been seen, in the limit as \( n \to \infty \).

The uniform nature of the variables \( X_i \) implies the following convenient structural facts about the family \( \{T_1, \ldots, T_n\} \):

- \( \{T_{i+1} - T_i : i = 0, \ldots, n-1\} \) are independent, and
- \( T_{i+1} - T_i \) has the law of a geometric random variable with parameter \( \frac{n-i}{n} \).

The latter comes from the fact that, once \( i \) values are known — which happens at “time” \( T_i \) — finding a value not yet seen amounts to independent trials with success probability \( \frac{n-i}{n} \). The former is checked formally by a similar argument.

So if \( Y_p \) is geometric with parameter \( p \) (and different occurrences of this random variable are regarded as independent), we have

\[
T_n \xrightarrow{\text{law}} \sum_{i=0}^{n-1} Y_{n-i} \quad (2.12)
\]

This permits us to study the expectation: a direct calculation shows that \( EX_p = 1/p \) and so, using the asymptotic form of the harmonic series,

\[
ET_n = \sum_{i=0}^{n-1} \frac{n}{n-i} = \sum_{k=1}^{n} \frac{1}{k} = \log n + \gamma + o(1), \quad n \to \infty \quad (2.13)
\]
where $\gamma$ is the Euler constant. This, as it turns out, yields the leading order behavior of $T_n$; the question now are the fluctuations.

We first write

$$T_n - ET_n \xrightarrow{\text{law}} \sum_{i=0}^{n-1} \left( \frac{Y_{n-i}}{n} - EY_{n-i} \right)$$ \hspace{1cm} (2.14)

Next we note that a calculation shows that

$$\Var(X_p) = \frac{1-p}{p^2}.$$ \hspace{1cm} (2.15)

So we get

$$\Var(T_n) = \sum_{i=0}^{n-1} \frac{i/n}{(n-i)^2} = n^2 \sum_{k=1}^{n-k} \frac{n-k}{k^2}$$ \hspace{1cm} (2.16)

A moment’s thought now reveals that

$$\frac{1}{n^2} \Var(T_n) \xrightarrow{n \to \infty} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$ \hspace{1cm} (2.17)

Hence, the fluctuations of $T_n$ around $ET_n$ are of order $n$; our remaining goal is to extract a distributional limit for $\frac{1}{n}(T_n - ET_n)$. However, this limit is not Gaussian. This is because

$$pX_p \xrightarrow{\text{law}} \exp(1)$$ \hspace{1cm} (2.18)

which yields

$$\frac{1}{n} \left( Y_{\frac{n}{k}} - EY_{\frac{n}{k}} \right) \xrightarrow{\text{law}} \frac{1}{k} (Z - EZ)$$ \hspace{1cm} (2.19)

with $Z$ exponential with parameter one. In particular, denoting the variable on the left by $\tilde{Y}_k$, thanks to the continuity of the law of $Z$ we have

$$E(\tilde{Y}_k^2 1_{|\tilde{Y}_k| > \epsilon}) \xrightarrow{n \to \infty} \frac{1}{k^2} E((Z - EZ)^2 1_{|Z - EZ| > \epsilon}) > 0$$ \hspace{1cm} (2.20)

and so (2.1) is not satisfied. In fact, even the conclusion of Theorem 2.1 fails since similar reasoning actually implies

$$\frac{T_n - ET_n}{n} \xrightarrow{\text{law}} \sum_{k=1}^{\infty} \frac{1}{k} (Z_k - EZ_k)$$ \hspace{1cm} (2.21)

where $Z_k$ are i.i.d. exponential with parameter one. Here the infinite series converges a.s. by, e.g., the Kolmogorov 3-series theorem. The limit object has a non Gaussian law.
3. Non-Gaussian Limit Theory

The example of marginal-CLT convergence we discussed in the previous section has a natural generalization. Consider i.i.d. random variables $X_1, X_2, \ldots$ with

$$X_1 \overset{\text{law}}{=} -X_1 \quad \text{and} \quad P(|X_1| > x) = x^{-\alpha} \quad \text{for} \quad x > 1. \quad (3.1)$$

where $\alpha \in (0, 2)$. Notice that $E|X_1| < \infty$ if and only if $\alpha > 1$ while $E(X_1^2) = \infty$ in all cases. Let $S_n := X_1 + \cdots + X_n$ as usual let us try to derive a limit law for $S_n$.

By independence we have

$$E(e^{itS_n}) = [E(e^{itX_1})]^n \quad (3.2)$$

and so we need a good representation for the ch.f. of $X_1$. The explicit form of the law of $X_1$ yields

$$E(e^{itX_1}) = 1 - \alpha \int_1^\infty \frac{1 - \cos(tx)}{x^{\alpha+1}} \, dx \quad (3.3)$$

Under the assumption that $\alpha \in (0, 2)$, the integral has a well defined (and finite) limit as the lower limit tends to zero. Denoting the integral by $C_\alpha(t)$, scaling $t$ by $n^{1/\alpha}$ thus yields

$$E(e^{itS_n/n^{1/\alpha}}) = \left[1 - \frac{1}{n} \alpha |t|^{\alpha} C_\alpha(t/n^{1/\alpha})\right]^n \rightarrow e^{-c|t|^\alpha} \quad (3.4)$$

where $C_\alpha$ is the limit value of the integral. We conclude:

**Theorem 3.1** For i.i.d. random variables $X_1, X_2, \ldots$ with the specific law (3.1) with $\alpha \in (0, \infty)$, and $S_n := X_1 + \cdots + X_n$, we have

$$S_n \overset{\text{law}}{\rightarrow} Y_\alpha \quad (3.5)$$

where $Y_\alpha$ is the random variable with ch.f. $e^{-c|t|^\alpha}$ with

$$c := \alpha \int_{(0, \infty)} \frac{1 - \cos(x)}{x^{\alpha+1}} \, dx. \quad (3.6)$$

In particular, $t \mapsto e^{-c|t|^\alpha}$ is a characteristic function for each $c > 0$ and each $\alpha \in (0, 2)$.

Notice that this naturally extends the CLT (Gaussian limit law) to a one parameter family of limit laws. Of course, the drawback of this derivation is that the law we started with was very special. However, it turns out that the general case is not considerably different. To discuss that in detail, we first make:

**Definition 3.2** A function $L: (0, \infty) \rightarrow (0, \infty)$ is said to be slowly varying at $+\infty$ if,

$$\lim_{t \rightarrow \infty} \frac{L(at)}{L(t)} = 1, \quad a > 0. \quad (3.7)$$
Examples of slowly varying functions are, e.g., \( L(x) = \log(x) \) or \( L(x) = e^{\sqrt{\log(x)}} \). A function of the form \( x^\alpha L(x) \), where \( L(x) \) is slowly varying, is said to be regularly varying with exponent \( \alpha \).

The result generalizing the example above is now as follows:

**Theorem 3.3** Let \( X_1, X_2, \ldots \) be i.i.d. random variables for which there are numbers \( \alpha \in (0, 2) \) and \( \theta \in [0, 1] \) such that the following two conditions hold:

1. \( P(X_1 > x)/P(|X_1| > x) \xrightarrow{x \to \infty} \theta \),
2. \( P(|X_1| > x) = x^{-\alpha}L(x) \) for \( L \) slowly varying at \(+\infty\).

Define

\[
a_n := \inf\{x > 0: nP(|X_1| > x) \leq 1\} \quad (3.8)
\]

\[
b_n := nE(1_{\{|X_1| \leq a_n\}}) \quad (3.9)
\]

Then \( S_n := X_1 + \cdots + X_n \) obeys

\[
S_n - b_n \xrightarrow{\text{law}} \frac{a_n}{n \to \infty} Y
\]

where \( Y \) is a random variable with the characteristic function

\[
E(e^{itY}) = \exp\left\{\int (e^{itx} - 1 - it1_{[-1,1]}(x)) (\theta1_{\{x > 0\}} + (1 - \theta)1_{\{x < 0\}}) \frac{\alpha dx}{|x|^\alpha + 1}\right\} \quad (3.11)
\]

As for the central limit theorem proof is based on asymptotic analysis of the characteristic function of \( S_n \) in terms of characteristic function of \( X_1 \). Before we commence the actual proof, let us note three additional properties that are consequences of conditions (1-2) above:

**Lemma 3.4** Under (1-2) above we have:

3. \( nP(|X_1| > a_n) \leq 1 \) and \( nP(|X_1| > a_n) \xrightarrow{n \to \infty} 1 \).

4. For each \( t > 0 \),

\[
nP(X_1 > ta_n) \xrightarrow{n \to \infty} \theta t^\alpha \quad \text{and} \quad nP(X_1 < -ta_n) \xrightarrow{n \to \infty} (1 - \theta) t^\alpha
\]

5. For each \( \delta > 0 \) there are \( C > 0 \) and \( t_0 < \infty \) such that

\[
\frac{P(|X_1| > xt)}{P(|X_1| > t)} \leq Cx^{-\alpha - \delta}, \quad x \in (0, 1], t \geq t_0.
\]

**Proof.** Assume without loss of generality that \( n \) is so large that \( a_n > 0 \). Then for any \( \epsilon > 0 \),

\[
nP(|X_1| > a_n) \leq 1 \leq nP(|X_1| > a_n - \epsilon a_n), \quad (3.14)
\]

where the left inequality is a consequence of the right continuity of \( x \mapsto P(|X_1| > x) \). By (2), the ratio of the right and left-hand sides tends to \((1 - \epsilon)^{\alpha}\) which can be made arbitrarily close to one by taking \( \epsilon \) small enough. This proves (3).

For (4) we just write

\[
nP(X_1 > ta_n) = nP(X_1 > a_n) \frac{P(X_1 > ta_n)}{P(|X_1| > a_n)} \xrightarrow{n \to \infty} \theta t^\alpha
\]

The first term on the right tends to one by (3), the second term tends to \( \theta \) by (1) and the last term tends to \( t^\alpha \) by (2). This proves the first half of (4); the second half is completely analogous.
For the claim (5), set \( t_0 \) to be so large that
\[
\frac{P(|X_1| > \frac{t}{2})}{P(|X_1| > t)} \leq 2^{\alpha + \delta}, \quad t \geq t_0.
\] (3.16)
This is possible because as \( t \to \infty \) the ratio tends to \( 2^{\alpha} \) by (2). Now assume that \( 2^{-n} \leq x < 2^{-n+1} \) and, noting that \( t \geq t_0 \), let \( k \) be the maximal integer less than \( n \) such that \( 2^{-k} \geq t_0 \). Then
\[
\frac{P(|X_1| > xt)}{P(|X_1| > t)} \leq \frac{P(|X_1| > 2^{-n}t)}{P(|X_1| > t)} < \frac{P(|X_1| > 2^{-n}t)}{P(|X_1| > 2^{-k}t)} \prod_{\ell=0}^{k-1} \frac{P(|X_1| > 2^{-\ell-1}t)}{P(|X_1| > 2^{-\ell}t)}
\] (3.17)
By (3.16) and our choice of \( k \), every term in the product on the right-hand side is less than \( 2^{\alpha + \delta} \). Invoking the bounds \( P(|X_1| > 2^{-n}t) \leq 1 \) and \( P(|X_1| > 2^{-k}t) \geq P(|X_1| > t_0) \), we thus get
\[
\frac{P(|X_1| > xt)}{P(|X_1| > t)} \leq \frac{1}{P(|X_1| > t_0)} 2^{k(\alpha + \delta)}
\] (3.18)
The claim follows by using that \( k \leq n \) and \( 2^n \leq 2x^{-1} \). \( \square \)

We now proceed to give the main ingredients of the proof of Theorem 3.1. The key point is a use of a cutoff at scale \( a_n \). Fix \( \varepsilon \in (0, 1) \) and define the random variables
\[
\overline{X}_{n,k} := X_k 1\{|X_k| \leq \varepsilon a_n\} \quad \text{and} \quad \overline{S}_n := \sum_{k=1}^n \overline{X}_{n,k}.
\] (3.19)
Then we have:

**Lemma 3.5** (Small-value variance) For each \( \delta \in (0, 2 - \alpha) \) there is \( C < \infty \) such that
\[
\limsup_{n \to \infty} \text{Var} \left( \frac{\overline{S}_n}{a_n} \right) \leq C \varepsilon^{2 - \alpha - \delta}.
\] (3.20)

**Proof.** Since \( |\overline{X}_1| \leq |X_1| \wedge (\varepsilon a_n) \), we have
\[
\text{Var} \left( \frac{\overline{S}_n}{a_n} \right) \leq nE \left( \frac{\overline{X}_{n,1}^2}{a_n^2} \right) \leq nE \left( \left( \frac{|X_1| \wedge (\varepsilon a_n)}{a_n} \right)^2 \right)
= n \int_0^\varepsilon 2yP(|X_1| > xa_n)\,dy
\] (3.21)
\[
\leq nP(|X_1| > a_n) \int_0^\varepsilon 2Cy^{1-\alpha-\delta} \,dy.
\] The claim now follows using (3). \( \square \)

Note that the right-hand side of (3.20) tends to zero as \( \varepsilon \downarrow 0 \). Hence, unlike for the CLT, the “small” values (at scale \( a_n \)) contribute little to the total sum \( S_n \). We are thus let to consider the opposite range of values and define
\[
\widehat{X}_{n,k} := X_k 1\{|X_k| > \varepsilon a_n\} \quad \text{and} \quad \widehat{S}_n := \sum_{k=1}^n \widehat{X}_{n,k}.
\] (3.22)
We then have:
Lemma 3.6 (Large-value limit)  For any \( \varepsilon > 0 \) and any \( t \in \mathbb{R} \),
\[
E(e^{i S_n/a_n}) \xrightarrow{n \to \infty} \exp \left\{ \int \left( e^{itx} - 1 \right) \left( \theta 1_{\{x > \varepsilon\}} + (1 - \theta) 1_{\{x < -\varepsilon\}} \right) \frac{\alpha dx}{|x|^\alpha + 1} \right\}
\]  (3.23)

Proof. Consider measures \( \mu_{n}^{\varepsilon, +} \) and \( \mu_{n}^{\varepsilon, -} \) defined as follows
\[
\mu_{n}^{\varepsilon, +}(A) := P(\hat{X}_{n,1}/a_n \in A \cap \mathbb{R}_+)
\]  (3.24)

Obviously, \( \mu_{n}^{\varepsilon, +} \) is supported on \((\varepsilon, \infty)\) while \( \mu_{n}^{\varepsilon, -} \) is supported on \((-\infty, -\varepsilon)\). We claim that
\[
\mu_{n}^{\varepsilon, +} + \mu_{n}^{\varepsilon, -} \xrightarrow{w \ n \to \infty} \left( \theta 1_{\{x > \varepsilon\}} + (1 - \theta) 1_{\{x < -\varepsilon\}} \right) \frac{\alpha dx}{|x|^\alpha + 1},
\]  (3.25)

where \( dx \) denotes the Lebesgue measure. A slight difference compared to earlier use of weak convergence is that the measures are not necessarily probability measures. However, we quickly note that both \( \{ \mu_{n}^{\varepsilon, +} \} \) and \( \{ \mu_{n}^{\varepsilon, -} \} \) are tight; indeed, once \( R > \varepsilon \),
\[
\mu_{n}^{\varepsilon, \pm}([-R, R]) \leq nP(|X_1| > R a_n) \xrightarrow{n \to \infty} R^{-\alpha}
\]  (3.26)

and the right-hand side tends to zero as \( R \to \infty \). To show weak convergence, it then suffices to look at the convergence of the distribution functions. Focusing on \( \mu_{n}^{\varepsilon, +} \), for \( x > \varepsilon \) we then get
\[
\mu_{n}^{\varepsilon, +}((-\infty, x]) = nP(\varepsilon a_n < X_1 \leq x a_n)
\]  (4)
\[
= nP(X_1 > \varepsilon a_n) - nP(X_1 > x a_n)
\]
\[
= \int_{(-\infty,x]} \theta (e^{-\alpha} - x^{-\alpha}) \frac{\alpha dx}{|x|^\alpha + 1}.
\]  (3.27)

The measure on the right-hand side is non-degenerate and so, by proper normalization, we conclude weak convergence \( \mu_{n}^{\varepsilon, +} \to \theta 1_{\{x > \varepsilon\}} \frac{\alpha dx}{|x|^\alpha + 1} \). A similar argument yields \( \mu_{n}^{\varepsilon, -} \to \theta 1_{\{x < -\varepsilon\}} \frac{\alpha dx}{|x|^\alpha + 1} \) and thus proves (3.25).

In order to prove the main assertion of the lemma, we write
\[
E(e^{i S_n/a_n}) = \left( 1 + \frac{nE(e^{i \hat{X}_{n,1}/a_n} - 1)}{n} \right)^n
\]  (3.28)

A comparison of various objects shows
\[
nE(e^{i \hat{X}_{n,1}/a_n} - 1) = \int (e^{itx} - 1)(\mu_{n}^{\varepsilon, +}(dx) + \mu_{n}^{\varepsilon, -}(dx)).
\]  (3.29)

Since \( x \mapsto e^{itx} - 1 \) is bounded, the integral converges to that with respect to the measure on the right of (3.25). The convergence claim (3.23) now follows form the fact that if \( z_n \in \mathbb{C} \) are such that \( z_n \to z \in \mathbb{C}, \) then \( (1 + z_n/n)^n \to e^z \).

Noting that the characteristic function on the right of (3.23) shares much of the structure with the desired limiting expression — just the term \( itx \{[-1,1] \}(x) \) is missing and there is the truncation by \( \varepsilon \) — we might be tempted to finish the proof by taking \( \varepsilon \downarrow 0 \). This will work for \( \alpha \in (0,1) \) but
will fail in the opposite range of $\alpha$’s as the limiting integral would not converge. An additional centering term is thus necessary. To see what kind of term we need, let us note that

$$S_n - b_n = S_n - E \tilde{S}_n + \tilde{S}_n - nE(\tilde{X}_{n,1} 1_{\{|\tilde{X}_{n,1}| \leq \alpha_n\}}).$$

(3.30)

We thus have to prove:

**Lemma 3.7** (Centering term) We have

$$\frac{n}{a_n} E(\tilde{X}_{n,1} 1_{\{|\tilde{X}_{n,1}| \leq \alpha_n\}}) \xrightarrow{n \to \infty} \int x 1_{[-1,1]}(x) \left( \theta 1_{\{x > \epsilon\}} + (1 - \theta) 1_{\{x < -\epsilon\}} \right) \frac{\alpha dx}{|x|^\alpha + 1}.$$  

(3.31)

**Proof.** Using the notation from the proof of the previous lemma, we have

$$\frac{n}{a_n} E(\tilde{X}_{n,1} 1_{\{|\tilde{X}_{n,1}| \leq \alpha_n\}}) = \int x 1_{[-1,1]}(x) (\mu_n^{e^+}(dx) + \mu_n^{e^-}(dx)).$$

(3.32)

The function $x \mapsto x 1_{[-1,1]}(x)$ is bounded and continuous a.e. with respect to the weak limit of $\mu_n^{e^+} + \mu_n^{e^-}$. It follows that the integral converges to that with respect to the limit on the right-hand side of (3.25). □

**Proof of Theorem 3.1.** Combining the previous two lemmas, we get that

$$\frac{1}{a_n} \left( \tilde{S}_n - nE(\tilde{X}_{n,1} 1_{\{|\tilde{X}_{n,1}| \leq \alpha_n\}}) \right) \xrightarrow{law} Y_\epsilon,$$

(3.33)

where $Y_\epsilon$ is a random variable whose characteristic function is as in (3.11), except that $x$ is restricted outside $(-\epsilon, \epsilon)^c$. As $\alpha < 2$ and $e^{ix} - 1 - it 1_{[-1,1]}(x)$ vanishes proportionally to $x^2$ as $x \to 0$, this characteristic function tends to the exponential on the right of (3.11) as $\epsilon \downarrow 0$. This exponential is in turn continuous in $t$ at $t = 0$ and so, by Levy Continuity Theorem, it is a characteristic function of a random variable $Y$ such that $Y_\epsilon \xrightarrow{law} Y$.

To see that the stated facts suffice to prove the desired claim, we need the following lemma:

**Lemma 3.8** Let $S_n$ be a sequence of random variables which, for each $\epsilon > 0$, admits the decomposition $S_n = Y_{n,\epsilon} + Z_{n,\epsilon}$, where

1. for each $\epsilon > 0$, $Y_{n,\epsilon} \xrightarrow{law} Y_\epsilon$ as $n \to \infty$,
2. $Y_\epsilon \xrightarrow{law} Y$ as $\epsilon \downarrow 0$, and
3. $\lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} E(Z_{n,\epsilon}^2) = 0$.

Prove that $S_n \Rightarrow Y$.

The proof of the lemma is relegated to HW1. The decomposition of the sum into two parts is provided by (3.30). □
4. STABLE LAWS AND CONVERGENCE OF TYPES

The law obtained in Theorem 3.1 is hard to describe more explicitly than just by its characteristic function. For those seeking a version without an integral, here is one:

\[
E(e^{itY}) = \begin{cases} 
\exp\left\{it\mu - \sigma^\alpha |t|^\alpha \left(1 - i\beta \text{sign}(t) \tan\left(\frac{\pi \alpha}{2}\right)\right)\right\}, & \text{if } \alpha \neq 1, \\
\exp\left\{it\mu - \sigma |t|^\beta \left(1 - i\beta \text{sign}(t) \frac{\sigma}{\sqrt{2}} \log |t|\right)\right\}, & \text{if } \alpha = 1, 
\end{cases}
\]

(4.1)

Here the parameters are called as follows:

- \(\alpha\) = index of stability,
- \(\beta\) = skewness, \(\beta := 2\theta - 1\),
- \(\sigma\) = scale,
- \(\mu\) = centering or shift.

Notice that when \(\alpha = 2\), the value of \(\beta\) is immaterial and we simply get the characteristic function of \(\mathcal{N}(\mu, \sigma^2)\). For \(\alpha < 2\), the values of the parameters \(\mu\) and \(\sigma\) are not simple to extract from the integral expression (and they bear little relation to the moments as these generally do not exist).

If we call a random variable with these parameters as \(S_{\alpha}(\sigma, \beta, \mu)\), we obtain the following distributional identity for independent versions thereof:

\[
S_{\alpha}(\sigma_1, \beta_1, \mu_1) + S_{\alpha}(\sigma_2, \beta_2, \mu_2) \overset{\text{law}}{=} S_{\alpha}(\sigma, \beta, \mu),
\]

(4.2)

where

\[
\sigma^\alpha := \sigma_1^\alpha + \sigma_2^\alpha, \quad \mu := \mu_1 + \mu_2, \quad \beta := \frac{\sigma_1^\alpha \beta_1 + \sigma_2^\alpha \beta_2}{\sigma_1^\alpha + \sigma_2^\alpha},
\]

(4.3)

so the parameters \(\mu\) and \(\sigma\) indeed do behave similarly as for addition of independent Gaussians. Similarly, one has the following relations under shift and scaling:

\[
S_{\alpha}(\sigma, \beta, \mu) \overset{\text{law}}{=} \mu + S_{\alpha}(\sigma, \beta, 0)
\]

(4.4)

and

\[
cS_{\alpha}(\sigma, \beta, \mu) \overset{\text{law}}{=} \begin{cases} 
S_{\alpha}(|c|\sigma, \text{sign}(c)\beta, c\mu), & \text{if } \alpha \neq 1, \\
S_{\alpha}(|c|\sigma, \text{sign}(c)\beta, c\mu - \frac{c}{\sqrt{2}} \beta \log |c|), & \text{if } \alpha = 1.
\end{cases}
\]

(4.5)

In particular,

\[
-S_{\alpha}(\sigma, \beta, \mu) \overset{\text{law}}{=} S_{\alpha}(\sigma, -\beta, -\mu)
\]

(4.6)

The extra term \(\frac{c}{\sqrt{2}} \beta \log |c|\) when \(\alpha = 1\) is an indication that the parametrization is somewhat singular at this value of \(\alpha\). Finally, we note that when \(\beta = 0\) and \(\mu = 0\), the law of \(S_{\alpha}\) is symmetric while for \(\beta = \pm 1\) it is said to be totally skewed.

The above laws actually fall into the category of stable laws. These are defined for instance as follows:

**Definition 4.1** A random variable \(Y\) is said to be stable — and it is said to have a stable law — if there is a sequence \(Y_1, Y_2, \ldots\) of i.i.d. random variables and two sequences \(\{a_n\}\) and \(\{b_n\}\) with \(a_n > 0\) such that for each \(n \geq 1\),

\[
Y \overset{\text{law}}{=} \frac{Y_1 + \cdots + Y_n}{a_n} - b_n
\]

(4.7)
The key point of this definition is that it permits us to characterize all possible limit laws for centered and scaled partial sums of a sequence of i.i.d. random variables.

**Theorem 4.2**  
*Y* is a distribution limit of the sequence $\frac{X_1 + \cdots + X_n}{a_n} - b_n$, for some i.i.d. $X_1, X_2, \ldots$ and sequences $b_n$ and $a_n > 0$, if and only if *Y* is stable.

**Proof.** The “if” part of the statement is trivial from the definition of stable random variables. For the “only if” part we denote $S_n = X_1 + \cdots + X_n$ and assume that

$$Y_n := \frac{S_n - b_n}{a_n} \xrightarrow{n \to \infty} Y$$

(4.8)

Fixing $k \geq 1$, we then have

$$Y_{nk} = \frac{1}{a_{nk}} \sum_{j=1}^{k} (S_{nj} - S_{n(j-1)}) - b_{nk} = \frac{a_n}{a_{nk}} \sum_{j=1}^{k} \left( \frac{S_{nj} - S_{n(j-1)}}{a_n} - b_n \right) + \frac{ka_n}{a_{nk}} b_n - b_{nk}$$

(4.9)

The random variables in the parentheses are independent and each of them converges in law to $Y$. Their joint law thus also converges to $k$ independent copies of $Y$. The key problem is that we have no information about the convergence of the factor $a_n/a_{nk}$ and the sequence $\frac{ka_n}{a_{nk}} b_n - b_{nk}$. For this we use:

**Theorem 4.3** (Convergence of types)  
Suppose $X_n \xrightarrow{\text{law}} X$ with $X$ non-degenerate and let $\alpha_n > 0$ and $\beta_n$ be real numbers such that $\alpha_n X_n + \beta_n \xrightarrow{\text{law}} Y$ for some $Y$ non-degenerate. Then there are $\alpha, \beta \in \mathbb{R}$ such that

1. $\alpha_n \to \alpha$ and $\beta_n \to \beta$,
2. $Y \xrightarrow{\text{law}} \alpha X + \beta$.

With this theorem in hand, we instantly get that $\frac{a_n}{a_{nk}} \to 1/A_k$ and $\frac{ka_n}{a_{nk}} b_n - b_{nk} \to -B_k$ and, moreover, $Y \xrightarrow{\text{law}} (Y_1 + \cdots + Y_k)/A_k - B_k$, where $Y_1, \ldots, Y_k$ are independent copies of $Y$. Since this holds for any $k \geq 1$, we conclude that $Y$ is stable. \(\square\)

It remains to prove the above “Convergence of types” result. The proof will be based on arguments concerning characteristic functions. The following properties, whose proof is left as homework, will be needed:

**Lemma 4.4**  
Let $X$ be a random variable and let $\varphi_X(t) := E(e^{itX})$.

1. If $|\varphi_X(t)| = 1$ then $X = c$ a.s. for some $c \in \mathbb{R}$.
2. If $|\varphi_X(t)| \to 1$ for all $t \in (-\delta, \delta)$ with $\delta > 0$, and $X_n \xrightarrow{\text{law}} -X_n$, then $X_n \xrightarrow{\text{law}} 0$.
3. If $|\varphi_X(at)| = |\varphi_X(t)|$ for all $t \in \mathbb{R}$, then either $X = c$ a.s. or $a = \pm 1$.

**Proof of Theorem 4.3.** Denote $\varphi_n(t) := E(e^{itX_n})$ and $\psi_n(t) := E(e^{it(a_n X_n + \beta_n)})$. Obviously,

$$\psi_n(t) = e^{itb_n} \varphi_n(at).$$

(4.10)
We know that \( \varphi_n(t) \to \varphi(t) := E(e^{itX}) \) and \( \psi_n \to \psi(t) := E(e^{itY}) \) with the convergence uniform on any compact subinterval of \( \mathbb{R} \).

First we claim that \( a_n \) remains bounded uniformly away from 0 and \( \infty \). (Remember that \( a_n > 0 \).) Indeed, if \( \liminf_{n \to \infty} a_n = 0 \), then \( a_{n_k} \to 0 \) for some sequence \( n_k \to \infty \). By (4.10) we then get \( |\psi_{n_k}(t)| \to 1 \) and so \( |\psi(t)| = 1 \). Thanks to Lemma 4.4, \( Y = c \) a.s., in contradiction with assumed non-degeneracy of \( Y \). On the other hand, if \( a_{n_k} \to \infty \) for some sequence \( n_k \to \infty \), then (4.10) similarly yields \( |\varphi(t)| = 1 \) contradicting non-degeneracy of \( X \).

Since \( a_n \) is bounded and uniformly positive, we may choose a converging subsequence \( a_{n_k} \to a \in (0, \infty) \). Then (4.10) gives \( |\psi(t)| = |\varphi(at)| \). So if we had two subsequences with different limits, \( a_1 \) and \( a_2 \), we would have \( |\varphi(a_1 t)| = |\varphi(a_2 t)| \). In particular, there would be \( a \in (0, \infty) \setminus \{1\} \) such that \( |\varphi(at)| = |\varphi(t)| \) for all \( t \). By Lemma 4.4(3), this is in conflict with assumed non-degeneracy of \( X \). Hence all subsequential limits of \( a_n \) are the same and so \( a_n \) converges.

Let \( a := \lim_{n \to \infty} a_n \) and note that, at least for \( t \) close to zero (so that \( \varphi_n(a_n t) \neq 0 \)),

\[
e^{it\beta_n} = \frac{\psi_n(t)}{\varphi_n(a_n t)}.
\]

(4.11)

Since (by the fact that \( a_n \to a \)) both the numerator and denominator converge to functions that are close to one in an open neighborhood of 0, so does \( e^{it\beta_n} \). This rules out the possibility that \( \beta_n \) is unbounded because then a sequence of \( t \)'s could be found tending to zero so that \( e^{it\beta_n} \to -1 \). So we may examine subsequential limits. Denoting by \( \beta \) a subsequential limit point, we then have

\[
e^{it\beta} = \frac{\psi(t)}{\varphi(at)} \quad \text{whenever} \quad \varphi(at) \neq 0.
\]

(4.12)

But this identifies \( t \to e^{it\beta} \) in an open neighborhood of 0. Taking the complex log, we extract a unique \( \beta \). Hence also \( \beta_n \to \beta \), proving (1). For (2) we invoke (4.12) one more time.

There are numerous other characterizations of stable laws, for instance: \( X \) is stable if for each \( \alpha, \beta > 0 \) there is \( \gamma > 0 \) such that \( \gamma X \overset{\text{law}}{=} \alpha X + \beta X' \), where \( X' \) is an independent copy of \( X \). Some of these will be explored in the homework, others may be found in a monograph “Stable Non-Gaussian Random Processes” by G. Samoradnitsky and M.S. Taqqu.
5. Infinite Divisibility

The stable laws permit us to characterize the limit laws for partial sums $S_n$ of an infinite sequence of i.i.d. random variables. We may wish to relax the conditions to include triangular arrays with independence guaranteed only for each row. More explicitly, we wish the study of limits points of sequences $X_{n,1} + \cdots + X_{n,n}$ where $\{X_{n,k} : k = 1, \ldots, n\}$ are i.i.d. for each $n$, but no specific relation for different $n$’s. (The scaling and normalization can now be absorbed into these random variables.) The class of possible limits is then described by:

**Definition 5.1** A random variable $X$ is said to be infinitely divisible if for each $n \geq 1$ there are i.i.d. $X_{n,1}, \ldots, X_{n,n}$ such that $X \xrightarrow{\text{law}} X_{n,1} + \cdots + X_{n,n}$.

The formal statement of the connection is then:

**Proposition 5.2** Let $\{X_{n,k}\}$ be such that $X_{n,1}, \ldots, X_{n,n}$ are i.i.d. for each $n \geq 1$. Any distributional limit of $X_{n,1} + \cdots + X_{n,n}$ is then infinitely divisible.

**Proof:** Let $S_n := X_{n,1} + \cdots + X_{n,n}$ and assume $S_n \xrightarrow{\text{law}} Y$. Write

$$S_{nk} = Y_{n,1} + \cdots + Y_{n,n} \text{ where } Y_{n,l} := \sum_{j=n(l-1)+1}^{nl} X_{n,j}. \quad (5.1)$$

Then $Y_{n,l}$ are i.i.d. We claim that $\{Y_{n,1} : n \geq 1\}$ is tight. This is seen from

$$P(S_{kn} > kt) \geq P(Y_{n,1}, \ldots, Y_{n,n} > t) = P(Y_{n,1} > t)^k \quad (5.2)$$

and the fact that $\{S_{kn} : n \geq 1\}$ is tight by assumed convergence to $Y$. (A similar argument handles also the negative tail of $Y_{n,1}$.) We may thus extract a subsequential limit $Y_{n,m} \xrightarrow{\text{law}} Z_1$. Thanks to independence, $(Y_{n,1}, \ldots, Y_{n,k}) \xrightarrow{\text{law}} (Z_1, \ldots, Z_k)$ and thus $Y \xrightarrow{\text{law}} Z_1 + \cdots + Z_k$, where $Z_1, \ldots, Z_k$ are i.i.d. As this holds for all $k \geq 1$, we conclude that $Y$ is infinitely divisible. \[ \square \]

The following theorem gives us a full characterization of infinitely-divisible laws:

**Theorem 5.3** (Levy-Khinchin formula) $X$ is infinitely divisible if and only if there is $\mu \in \mathbb{R}$, $\sigma^2 \in [0, \infty)$ and a finite Borel measure $\nu$ on $\mathbb{R}$ with $\nu(\{0\}) = 0$ such that

$$E(e^{itX}) = \exp\left\{it\mu - \frac{\sigma^2}{2} t^2 + \int \left( e^{itx} - 1 - \frac{itx}{1+x^2} \right) \frac{1+x^2}{x^2} \nu(dx) \right\}. \quad (5.3)$$

Moreover, the objects $(\mu, \sigma, \nu)$ with the above properties are determined uniquely.

We start by proving the “only if” part of the statement. First we recall a general fact:

**Lemma 5.4** Let $\varphi_Z(t) := E(e^{itZ})$ for some random variable $Z$. There is a constant $c < \infty$ such that for all $R > 0$,

$$P(|Z| > R) \leq cR \int_0^{1/R} \left[ 1 - \Re \varphi_Z(t) \right] dt \quad (5.4)$$
Proof. This is an argument from the proof of the Levy Continuity Theorem: Write
\[
\frac{1}{u} \int_0^u E \left( 1 - \cos(uZ) \right) du = E \left( 1 - \frac{\sin(uZ)}{uZ} \right) \\
\geq E \left( 1_{\{u|Z| \geq 1\}} \left( 1 - \frac{\sin(uZ)}{uZ} \right) \right) \\
\geq \left[ \inf_{s \geq 1} \left( 1 - \frac{\sin(s)}{s} \right) \right] P(u|Z| \geq 1) .
\]
(5.5)
The infimum is positive so we may write it as \(1/c\) for some \(c \in (0, \infty)\). The claim follows by noting that \(E(\cos uZ) = \Re \varphi_Z(u)\) and setting \(u := 1/R\). \(\square\)

Suppose that \(X\) is infinitely divisible. Then \(X \overset{\text{law}}{=} X_{n,1} + \cdots + X_{n,n}\) with \(X_{n,1}, \ldots, X_{n,n}\) i.i.d. Denote
\[
\varphi(t) := E(e^{itX}) \quad \text{and} \quad \varphi_n(t) := E(e^{itX_{n,1}}) \quad (5.6)
\]
Clearly, we have
\[
\varphi(t) = [\varphi_n(t)]^n , \quad t \in \mathbb{R} ,
\]
(5.7) for all \(n \geq 1\). A key fact that opens the door to many useful consequences comes in:

**Lemma 5.5** We have \(X_{n,1} \overset{\text{law}}{\to} 0\) and so \(\varphi_n(t) \to 1\) uniformly on compact intervals of \(t\).

**Proof.** Let \(\delta > 0\) be such that \(\Re \varphi(t) > \frac{1}{2}\) for \(t \in (-\delta, \delta)\). Then (5.7) implies \(|\varphi_n(t)| \to 1\) uniformly \(t \in (-\delta, \delta)\). In particular, a branch of the complex logarithm that is analytic in \(\{z \in \mathbb{C}: \Re z > 0\}\) with \(\log 1 = 0\) can be applied to (5.7) with the result
\[
\log \varphi(t) = n \log \varphi_n(t) , \quad -\delta < t < \delta .
\]
(5.8) Dividing by \(n\), the left-hand side tends to zero and so \(\log \varphi_n(t) \to 0\) as well. The convergence is uniform by uniform continuity of \(\varphi\). Using that our branch of \(z \mapsto \log z\) is a one-to-one function, we infer \(\varphi_n(t) \to 1\) for \(t \in (-\delta, \delta)\). The above tightness lemma subsequently ensures that \(\{X_{n,1} : n \geq 1\}\) and, since any weakly converging subsequence must tend to zero (by Lemma 4.4(1)), we get \(X_{n,1} \overset{\text{law}}{\to} 0\). The claim now follows. \(\square\)

Now we get:

**Corollary 5.6** We have
\begin{enumerate}
\item \(\varphi(t) \neq 0\) for every \(t \in \mathbb{R}\),
\item \(t \mapsto \log \varphi(t)\) exists as a continuous function on \(\mathbb{R}\),
\item \(n[\varphi_n(t) - 1] \to \log \varphi(t)\) as \(n \to \infty\), and
\item there is \(\varepsilon_n(t)\) such that
\[
\varphi(t) = \exp \left\{ (1 + \varepsilon_n(t))n[\varphi_n(t) - 1] \right\} \quad (5.9)
\]
with \(\varepsilon_n(t) \to 0\) locally uniformly in \(t\).
\end{enumerate}
Proof. Let $g$ the complex function defined on the unit disc in $\mathbb{C}$ by
\[ g(z) := \frac{\log(1 + z) - z}{z} \] (5.10)
This function is analytic on $\{z: |z| < 1\}$ and obeys $g(z) \to 0$ as $z \to 0$. So, once $|\varphi_n(t) - 1| < 1$, we can write
\[ \varphi(t) = (1 + [\varphi_n(t) - 1])^n = \exp\left\{ (1 + g(\varphi_n(t) - 1) n[\varphi_n(t) - 1]\right\}. \] (5.11)
The argument of the exponential is always a finite complex number (whenever the formula applies) and so we get that $\varphi(t) \neq 0$ for all $t$.

Let us examine the curve $t \mapsto \varphi(t)$ in the complex plane. This curves starts of at 1 when $t = 0$ and then moves continuously around while avoiding 0. It follows that, for any compact interval of $t$’s, we can find an analytic continuation of $z \mapsto \log z$ to a “tube” neighborhood of the curve. In particular, $\log \varphi(t)$ is well defined by this continuation for all $t \in \mathbb{R}$. Moreover, since $g(\varphi_n(t) - 1) \to 0$ uniformly on compact intervals of $t$ and $t \mapsto n[\varphi_n(t) - 1]$ is continuous (and so there is no ambiguity in the choice of the imaginary of the log) we must have that
\[ (1 + g(\varphi_n(t) - 1) n[\varphi_n(t) - 1] = \log \varphi(t) \] (5.12)
once $n$ is sufficiently large. Hence we have $n[\varphi_n(t) - 1] \to \log \varphi(t)$ and also the asymptotic expression in part (4) of the claim. \[ \Box \]
With the above facts in hand, we can start the actual proof of the “only if” part of the Levy-Khinchin formula. Drawing inspiration on the proof of stable convergence, consider the measure
\[ \mu_n(A) := P(X_{n,1} \in A) \] (5.13)
This is relevant because
\[ n[\varphi_n(t) - 1] = \int (e^{itx} - 1) \mu_n(dx) \] (5.14)
Our goal is to extract a weak limit of the measures $\{n \mu_n\}$. First we note that this family is tight:

**Lemma 5.7 (Tightness)** There is $c < \infty$ such that for each $R > 0$,
\[ \limsup_{n \to \infty} n \mu_n([-R, R]^c) \leq cR \int_0^{1/R} \log \frac{1}{|\varphi(t)|} dt. \] (5.15)

**Proof.** Apply the identity in Lemma 5.4 for $Z := X_{n,1}$ and note that, by the previous corollary,
\[ n[1 - \Re \varphi_n(t)] \xrightarrow{n \to \infty} \Re \log \varphi(t) = \log \frac{1}{|\varphi(t)|} \] (5.16)
locally uniformly in $t$. \[ \Box \]
Notice that this lemma indeed guarantees tightness because the $R \to \infty$ limit of the right-hand side is $\log 1/|\varphi(0)| = 0$ by the Mean-Value Theorem and continuity of $t \mapsto \varphi(t)$. Once we know tightness of $n \mu_n$, we have to wonder where the total mass of this measure, which is $n$, goes as $n \to \infty$. It turns out that it concentrates at zero. A uniform estimate on how the concentration occurs is given in:
Lemma 5.8 (Collapse of mass to zero) We have
\[
\limsup_{n \to \infty} \int_{[-1,1]} x^2 n \mu_n(dx) < \infty. \tag{5.17}
\]

Proof. This follows because we \(1 - \cos u \geq cu^2\) whenever \(|u| \leq 1\) and so we can write
\[
n \left[1 - \Re \varphi_n(1)\right] \geq n E \left(1 - \cos(X_{n,1})\right) \geq n E(1_{|X_{n,1}| \leq 1}(1 - \cos(X_{n,1})) \geq n E(X_{n,1}^2 1_{|X_{n,1}| \leq 1}) \tag{5.18}
\]
The right-hand side is now easily identified with the integral under consideration; the left-hand side tends to \(\log 1/|\varphi(1)|\) as before. \(\square\)

Proof of Theorem 5.3, “only if” part. Introduce the measure
\[
\nu_n(dx) := \frac{x^2}{1 + x^2} n \mu_n(dx). \tag{5.19}
\]
Thanks to the above lemmas and the boundedness of \(x \mapsto \frac{x^2}{1 + x^2}\), the measures \(\{\nu_n\}\) are tight with a uniform bound on the total mass. We may thus extract a subsequence \(n_k\) so that \(\nu_{n_k} \xrightarrow{\text{law}} \nu\). Next we write
\[
n(\varphi_n(t) - 1) = \int \left(e^{ix} - 1\right) \frac{1 + x^2}{x^2} \nu_n(dx)
= \int \left(e^{ix} - 1 - \frac{ix}{1 + x^2}\right) \frac{1 + x^2}{x^2} \nu_n(dx) + it \int \frac{x}{1 + x^2} n \mu_n(dx) \tag{5.20}
\]
The reason for the second line is that the function under the first integral is now bounded on \(\mathbb{R}\) and so the integral converges to that with respect to \(\nu\). But the left-hand side converges to \(\log \varphi(t)\) and so also the second integral on the right must converge to some quantity called \(\mu\). We conclude that, for all \(t \in \mathbb{R}\),
\[
\log \varphi(t) = \int \left(e^{ix} - 1 - \frac{ix}{1 + x^2}\right) \frac{1 + x^2}{x^2} \nu(dx) + it \mu. \tag{5.21}
\]
This is almost the desired expression except that we do not necessarily have \(\nu(\{0\}) = 0\) (although \(\nu_n\) had no point mass at zero, this could change after taking weak limits). Splitting the point mass at zero from the integral and calling it \(\sigma^2\), we get an extra term \(\frac{1}{2} \sigma^2 t^2\), with \(\frac{1}{2} t^2\) being the limit value of the integrand at \(x = 0\). The “only if” part of the claim follows. \(\square\)

Proof of Theorem 5.3, “if” part. Here we will proceed by a direct construction. Suppose \(\mu \in \mathbb{R}\), \(\sigma^2 \in [0, \infty)\) and a finite measure \(\nu\) with no mass at zero is given. Define
\[
\lambda(dx) := \frac{1 + x^2}{x^2} \nu(dx) \tag{5.22}
\]
and, given \(\varepsilon > 0\), define
\[
\lambda_{\varepsilon}(dx) := 1_{(-\varepsilon, \varepsilon)}(x) \lambda(dx) \tag{5.23}
\]
and
\[
I_\epsilon(t) := \int \left( e^{itx} - 1 - \frac{itx}{1 + x^2} \right) \lambda_\epsilon(dx) \tag{5.24}
\]

Note that \( \lambda_\epsilon(\mathbb{R}) < \infty \) and so also \( I_\epsilon(t) \) is finite for all \( t \).

Consider now i.i.d. random variables \( Y_1, Y_2, \ldots \) with
\[
P(Y_1 \in A) = \frac{\lambda_\epsilon(A)}{\lambda_\epsilon(\mathbb{R})} \tag{5.25}
\]

and let \( N_\epsilon \) be Poisson with parameter \( \lambda_\epsilon(\mathbb{R}) \), independent of the \( Y_1, Y_2, \ldots \). Define
\[
X_\epsilon := \sum_{j=1}^{N_\epsilon} Y_j. \tag{5.26}
\]

We now compute the characteristic function of \( Y_\epsilon \). Due to independence, we will do the integrals over \( N_\epsilon \) after the integrals over \( Y_1, Y_2, \ldots \). Denoting \( \varphi(t) := E(e^{itY_1}) \), this yields
\[
E(e^{itX_\epsilon}) = E(\varphi(t)^{N_\epsilon}) = \sum_{n=0}^{\infty} \frac{\lambda_\epsilon(\mathbb{R})^n}{n!} e^{-\lambda_\epsilon(\mathbb{R})} \varphi(t)^n = \exp \left\{ \lambda_\epsilon(\mathbb{R}) [\varphi(t) - 1] \right\} \tag{5.27}
\]

Comparing all involved objects, we then conclude that
\[
E(e^{itX_\epsilon}) = e^{I_\epsilon(t)}. \tag{5.28}
\]

But the fact that \( \nu \) is finite and the integrand in the Levy-Khinchin formula is bounded implies
\[
I_\epsilon(t) \to I(t) := \int \left( e^{itx} - 1 - \frac{itx}{1 + x^2} \right) \lambda(dx) \tag{5.29}
\]

As \( t \mapsto I(t) \) is also continuous, the Levy Continuity Theorem guarantees that \( X_\epsilon \xrightarrow{\text{law}} X \) with the characteristic function of \( X \) given as \( e^{I(t)} \).

It remains to show that \( X \) is infinitely divisible. But this follows trivially from the fact that the same construction as above (with \( \nu \) rescaled by \( 1/k \)) ensures that \( e^{I(t)/k} \) is a characteristic function of a random variable. Then \( X \) is a sum of \( k \) independent copies of this random variable and so it is infinite divisible as desired. \( \square \)

We remark that the construction underlying the proof of the “if” part naturally appears in the context of the Poisson point processes. In fact, \( \lambda \) is then the so called intensity measure and \( X \) corresponds to the associated compound Poisson process.
6. Conditional Expectation

Having discussed at length the limit theory for sums of independent random variables we will now move on to deal with dependent random variables. An important tool in this is conditioning. In this lecture we will discuss the simplest example of this technique, leading to the definition of conditional expectation.

**Definition 6.1** Consider a probability space \((\Omega, \mathcal{F}, P)\) and a random variable \(X \in L^1(\Omega, \mathcal{F}, P)\). Let \(\mathcal{G} \subset \mathcal{F}\) be a sigma-algebra. We say that the random variable \(Y\) is a conditional expectation of \(X\) given \(\mathcal{G}\) if

1. \(Y\) is \(\mathcal{G}\)-measurable, and
2. we have \(E(Y1_A) = E(X1_A)\) for every \(A \in \mathcal{G}\).

The last condition in particular implies (choosing \(A := \{Y \geq 0\}\)) that \(Y\) is integrable.

We first check that the random variable \(Y\) is determined uniquely by properties (1-2), that is, assuming it exists in the first place.

**Lemma 6.2** If \(Y\) and \(Y'\) satisfy (1-2), then \(Y = Y' \) a.s.

**Proof.** Let \(A_{\varepsilon} := \{Y \geq Y' + \varepsilon\}\). Then \(A_{\varepsilon} \in \mathcal{G}\) because both \(Y\) and \(Y'\) are \(\mathcal{G}\)-measurable by (1), and so it thus \(Y - Y'\). Therefore, by (2),

\[
E(Y1_{A_{\varepsilon}}) = E(X1_{A_{\varepsilon}}) = E(Y'1_{A_{\varepsilon}}). \tag{6.1}
\]

All these numbers are finite and so we may subtract the left and right-hand sides to get \(E((Y - Y')1_{A_{\varepsilon}}) = 0\). But \(Y - Y' \geq \varepsilon\) on \(A_{\varepsilon}\) and so we get \(P(A_{\varepsilon}) = 0\). Taking union over \(\varepsilon \in \{2^{-n} : n \geq 1\}\) yields \(Y \geq Y'\) a.s. The other inequality follows by symmetry. \(\Box\)

Since the conditional expectation of \(X\) given \(\mathcal{G}\) is determined up to a modification on a null set, we will henceforth write \(E(X|\mathcal{G})\) to denote any version of this function. However, we first have to address existence:

**Lemma 6.3** For each \(X \in L^1\), the conditional expectation \(E(X|\mathcal{G})\) exists.

The proof will be based on the well known Lebesgue decomposition of measures. Given two measures \(\mu, \nu\) on the same measurable space, recall that \(\mu\) is said to be absolutely continuous with respect to \(\nu\) — with the notation \(\mu \ll \nu\) — if \(\nu(A) = 0\) implies \(\mu(A) = 0\). If \(\mu\) and \(\nu\) are in turn supported on disjoint sets, i.e., there exists \(A\) such that \(\mu(A) = 0\) and \(\nu(A^c) = 0\), then we say that these measures are mutually singular, with notation \(\mu \perp \nu\).

**Theorem 6.4** (Lebesgue decomposition, Radon-Nikodym Theorem) Let \(\mu, \nu\) be \(\sigma\)-finite measures on \((\Omega, \mathcal{F})\). Then there are measures \(\mu_1, \mu_2\) so that

\[
\mu = \mu_1 + \mu_2, \quad \mu_1 \ll \nu, \quad \mu_2 \perp \nu. \tag{6.2}
\]

Moreover, there exists an \(\mathcal{F}\)-measurable, non-negative \(f : \Omega \to \mathbb{R}\) such that

\[
\mu_1(A) = \int_A f(x) \nu(dx), \quad A \in \mathcal{F}. \tag{6.3}
\]
We remark that the function \( f \) from the second part of this theorem is called the Radon-Nikodym derivative of \( \mu \) with respect to \( \nu \), sometimes also denoted \( \frac{d\mu}{d\nu} \).

There is a classical proof of the Lebesgue decomposition which is based on the concept of a signed measure and the Hahn and Jordan decompositions. The proof we will present will in turn be based on Hilbert space techniques.

Recall that a linear vector space \( \mathcal{H} \) is a Hilbert space if it complete in the norm \( \| \cdot \| \) associated with the scalar product \( (\cdot, \cdot) \) on \( \mathcal{H} \). An example of a Hilbert space is \( L^2(\Omega, \mathcal{F}, \mu) \). A linear map \( \varphi : \mathcal{H} \to \mathbb{R} \) (assume \( \mathcal{H} \) is over reals) is a continuous linear functional if \( |\varphi(x)| \leq C\|x\| \) for all \( x \in \mathcal{H} \). Relatively straightforward functional-theoretic arguments then yield:

**Theorem 6.5 (Riesz representation)** Let \( \mathcal{H} \) be a Hilbert space and \( \varphi \in \mathcal{H}^* \) a continuous linear functional. Then there is a \( a \in \mathcal{H} \) such that

\[
\varphi(x) = \langle a, x \rangle, \quad x \in \mathcal{H}.
\]  

*(Sketch of main ideas.)* First we realize that \( \mathcal{B} := \{ x \in \mathcal{H} : \varphi(x) = 0 \} \) is a closed linear subspace of \( \mathcal{H} \). The case \( \varphi = 0 \) is handled trivially, so we may suppose \( \varphi \neq 0 \). Then \( \mathcal{A} \neq \mathcal{H} \) and so there exists some \( x \in \mathcal{H} \setminus \mathcal{B} \). Considering the orthogonal decomposition of \( x = a + b \), where \( b \in \mathcal{B} \) and \( a \perp \mathcal{B} \), it is then checked that \( a \) has the stated property. \( \square \)

With this fact in hand, we can now give:

**Proof of Theorem 6.4.** The \( \sigma \)-finiteness assumption permits us to assume that both \( \mu \) and \( \nu \) are finite measures. (Otherwise partition the space into parts where this is true.) Consider the Hilbert space \( \mathcal{H} := L^2(\Omega, \mathcal{F}, \mu + \nu) \) and define the functional

\[
\varphi(f) := \int f(x) \mu(dx), \quad f \in \mathcal{H}. 
\]  

Since \( \mu \) is finite, we have by Cauchy-Schwarz that

\[
|\varphi(f)| \leq \mu(\Omega)^{1/2} \|f\|_{L^2(\mu)} \leq C \|f\|_{\mathcal{H}}. 
\]  

Since \( \varphi \) is obviously linear, the above expression defines a continuous linear functional. The Riesz representation ensures the existence of an \( h \in \mathcal{H} \) such that

\[
\int f(x) \mu(dx) = \int f(x)h(x)[\mu + \nu](dx), \quad f \in L^2(\Omega, \mathcal{F}, \mu + \nu). 
\]  

We now make a couple of observations: First,

\[
\nu(h \geq 1) = 0. 
\]  

Indeed, plugging \( f := 1_{\{h \geq 1\}} \) implies \( \mu(h \geq 1) = \varphi(h \geq 1) \geq \nu(h \geq 1) \) implying the claim. Similarly, we observe that

\[
\mu(h > 1) = 0, 
\]  

because taking \( f := 1_{\{h \geq 1 + \varepsilon\}} \) yields \( \mu(h \geq 1 + \varepsilon) \geq \varphi(h \geq 1 + \varepsilon) \) and so \( \mu(h \geq 1 + \varepsilon) = 0 \) for all \( \varepsilon > 0 \); the claim is obtained by taking a union along a sequence of \( \varepsilon \downarrow 0 \). Finally, we also claim that

\[
[\mu + \nu](h < 0) = 0. 
\]
Indeed, taking \( f := 1_{\{ h < -\varepsilon \}} \) yields \( -\varepsilon [\mu + \nu](h < -\varepsilon) \geq 0 \) implying that \( [\mu + \nu](h < -\varepsilon) = 0 \); taking a union along a sequence \( \varepsilon \downarrow 0 \) then finishes the claim as well.

With these observations in hand, we now define

\[
\mu_1(dx) := 1_{\{0 \leq h(x) < 1\}} \mu(dx) \quad \text{and} \quad \mu_2(dx) := 1_{\{h(x) = 1\}} \mu(dx). \tag{6.11}
\]

Then \( \mu_1 + \mu_2 = \mu \) and \( \mu_2 \perp \nu \) by the above claims. For the proof of the remaining properties, we rewrite (6.7) as

\[
\int f(x) [1 - h(x)] \mu(dx) = \int f(x) h(x) \nu(dx), \quad f \in L^2(\Omega, \mathcal{F}, \mu + \nu) \tag{6.12}
\]

On the left hand side we can now replace \( \mu \) by \( \mu_1 \); setting

\[
f(x) := 1_{\{0 \leq h(x) < 1 - \varepsilon\}} \frac{1}{1 - h(x)} 1_A \tag{6.13}
\]

which is bounded and thus in \( L^2(\Omega, \mathcal{F}, \mu + \nu) \) for \( \varepsilon > 0 \), and taking \( \varepsilon \downarrow 0 \) with the help of the Monotone Convergence Theorem then yields

\[
\mu_1(A) = \int_A \frac{h(x)}{1 - h(x)} 1_{\{0 \leq h(x) < 1\}} \nu(dx), \quad A \in \mathcal{F}. \tag{6.14}
\]

This is the desired expression. In particular, \( \mu_1(A) = 0 \) if \( \nu(A) = 0 \), so \( \mu_1 \ll \nu \) as claimed. \( \square \)

We are now also ready to conclude the existence of the conditional expectation:

**Proof of Lemma 6.3.** Suppose first that \( X \geq 0 \). Define the measure \( \mu \) on \( \mathcal{G} \) by \( \mu(A) := E(X1_A) \). This measure is \( \sigma \)-finite and it is absolutely continuous with respect to \( P \) — the restriction of the probability measure to \( \mathcal{G} \). Hence, there is \( \mathcal{G} \)-measurable \( Y \geq 0 \) such that

\[
\mu(A) = \int_A Y(\omega) P(d\omega), \quad A \in \mathcal{G}. \tag{6.15}
\]

It follows that \( Y \) satisfies the properties (1-2) of conditional expectation. This constructs \( E(X|\mathcal{G}) \) for non-negative \( X \); the general integrable \( X \) are taken care of by expanding into positive and negative parts and applying the above to each of the separately. \( \square \)

We proceed to discuss some basic properties of conditional expectation:

**Lemma 6.6** Assuming all random variables below are integrable, we have:

1. (triviality on constants) \( E(1|\mathcal{G}) = 1 \).
2. (positivity) \( X \geq 0 \Rightarrow E(X|\mathcal{G}) \geq 0 \) a.s.
3. (linearity) \( E(\alpha X + \beta Y|\mathcal{G}) = \alpha E(X|\mathcal{G}) + \beta E(Y|\mathcal{G}) \)
4. (Monotone Convergence Theorem) \( 0 \leq X_n \uparrow X \Rightarrow E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G}) \).
5. \( E(E(X|\mathcal{G})) = E(X) \).
6. If \( X \) is \( \mathcal{G} \) measurable, then \( E(X|\mathcal{G}) = X \) a.s.

**Proof.** These properties are direct applications of the definition and (for (4)) also the Monotone Convergence Theorem. We leave the details to the reader. \( \square \)

A consequence of (3) is then:
Corollary 6.7 (Conditional Fatou lemma) Let $X_n \geq 0$ be integrable. Then
\[
\liminf_{n \to \infty} E(X_n|\mathcal{G}) \geq E\left(\liminf_{n \to \infty} X_n |\mathcal{G}\right).
\] (6.16)

A consequence of (5) is in turn:

Lemma 6.8 If $X \in L^1(\Omega, \mathcal{F}, P)$, then
\[
|E(X|\mathcal{G})| \leq E(|X||\mathcal{G})
\] (6.17)
and
\[
\|E(X|\mathcal{G})\|_{L^1(\Omega, \mathcal{G}, P)} \leq \|X\|_{L^1(\Omega, \mathcal{F}, P)}.
\] (6.18)

In particular, $X \mapsto E(X|\mathcal{G})$ is a continuous linear map — in fact, a contraction — of $L^1(\Omega, \mathcal{F}, P)$ onto $L^1(\Omega, \mathcal{G}, P)$.

Proof. We have $-|X| \leq X \leq |X|$ and so
\[
-E(|X|\mathcal{G}) \leq E(X|\mathcal{G}) \leq E(|X|\mathcal{G}) \quad \text{a.s.}
\] (6.19)
This proves (6.17); taking expectation then yields (6.18). Since $X \mapsto E(X|\mathcal{G})$ is obviously linear, it defines a linear map of $L^1(\Omega, \mathcal{F}, P)$ onto $L^1(\Omega, \mathcal{G}, P)$. The operator norm of this map is less than one so this map is in fact a contraction. The map is onto as it an identity on $L^1(\Omega, \mathcal{F}, P)$. \(\square\)

A standard argument also yields:

Lemma 6.9 (Conditional Cauchy-Schwarz inequality) Suppose $X, Y \in L^2(\Omega, \mathcal{F}, P)$ and suppose $\mathcal{G} \subset \mathcal{F}$ is a sigma algebra. Then
\[
[E(XY|\mathcal{G})]^2 \leq E(X^2|\mathcal{G})E(Y^2|\mathcal{G}) \quad \text{a.s.}
\] (6.20)

Proof. Consider the function
\[
\lambda \mapsto E\left((X - \lambda Y)^2|\mathcal{G}\right).
\] (6.21)
Let $\Omega'$ be the event that this function is non-negative for all $\lambda \in \mathbb{Q}$. Then $\Omega' \in \mathcal{G}$ with $P(\Omega') = 1$. Expanding the square and using that the discriminant of the quadratic polynomial which is non-negative on all rationals is non-positive, we get that the inequality in (6.20) holds on $\Omega'$. \(\square\)

Lemma 6.10 (Conditional Jensen’s inequality) Let $\phi: \mathbb{R} \to \mathbb{R}$ be convex and bounded from below. Then for any $X$ such that $\phi(X) \in L^1$,
\[
E\left(\phi(X)\right|\mathcal{G}\right) \geq \phi\left(E(X|\mathcal{G})\right) \quad \text{a.s.}
\] (6.22)

Proof. Define $\mathcal{L} := \{(a, b) \in \mathbb{R}^2 : \phi(x) > ax + b \, \forall x \in \mathbb{R}\}$. Supposing that $\phi$ is not linear and bounded from below, the set $\mathcal{L}$ is non-empty and open in the topology of $\mathbb{R}^2$. In particular, $\mathcal{L} \cap \mathbb{Q}^2$ is dense in $\mathcal{L}$. Then
\[
\phi(x) = \sup_{(a, b) \in \mathcal{L} \cap \mathbb{Q}^2} (ax + b), \quad x \in \mathbb{R}.
\] (6.23)
Since $\phi(X) \geq aX + b$ for each $(a, b) \in \mathcal{L}$, and since $\phi(X)$ and thus — thanks to our conditions on $\phi$ — also $X$ is integrable, we thus have
\[
E\left(\phi(X)\right|\mathcal{G}\right) \geq aE(X|\mathcal{G}) + b \quad \text{a.s.}
\] (6.24)
Let $\Omega_{a,b}$ be the event where the inequality holds. Then $P(\Omega_{a,b}) = 1$. Define $\Omega' := \cap_{(a,b) \in L \cap Q^2} \Omega_{a,b}$ and note that $P(\Omega') = 1$. The proof is completed by taking supremum over $(a,b) \in L \cap Q^2$ in (6.24) which implies that the desired inequality holds on $\Omega'$.

A consequence of the conditional Jensen inequality is a generalization of Lemma 6.8:

**Lemma 6.11** For any $p \in [1, \infty]$, the map $X \mapsto E(X|\mathcal{G})$ is a continuous linear transform of $L^p(\Omega, \mathcal{F}, P)$ onto $L^p(\Omega, \mathcal{G}, P)$. In fact, this map is a contraction.

**Proof.** Let $p \in [1, \infty)$. By Jensen’s inequality we have

$$\left( E(|X|^p|\mathcal{G}) \right)^{1/p} \leq E(\left| X \right|^p|\mathcal{G}) \quad \text{a.s.}$$

(6.25)

Taking expectations and raising both sides to $1/p$ yields the result.

For $p := \infty$ we notice that $|X| \leq \|X\|_{\infty}$ implies

$$\left| E(X|\mathcal{G}) \right| \leq \|X\|_{\infty} \quad \text{a.s.}$$

(6.26)

and so the result follows in this case as well. □

As a homework assignment, we also established:

**Lemma 6.12** (Conditional Hölder inequality) Let $p, q \in [1, \infty]$ be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then for any sigma algebra $\mathcal{G} \subset \mathcal{F}$, all $X \in L^p(\Omega, \mathcal{F}, P)$ and all $Y \in L^q(\Omega, \mathcal{F}, P)$,

$$\left| E(XY|\mathcal{G}) \right| \leq E\left( \left| X \right|^p|\mathcal{G} \right)^{1/p} E\left( \left| Y \right|^q|\mathcal{G} \right)^{1/q} \quad \text{a.s.}$$

(6.27)

Finally, we will examine the dependence of the conditional expectation on the underlying sigma algebra. First we note this useful principle:

**Lemma 6.13** ("Smaller always wins" principle) Let $\mathcal{G}_1 \subseteq \mathcal{G}_2$ be sub sigma-algebras of $\mathcal{F}$. Then for any $X \in L^1(\Omega, \mathcal{F}, P)$

$$E\left( E\left( X|\mathcal{G}_1 \right) |\mathcal{G}_2 \right) = E\left( X|\mathcal{G}_1 \right) \quad \text{a.s.}$$

(6.28)

as well as

$$E\left( E\left( X|\mathcal{G}_2 \right) |\mathcal{G}_1 \right) = E\left( X|\mathcal{G}_1 \right) \quad \text{a.s.}$$

(6.29)

In short, sequential conditioning always reduces to that with respect to the smaller sigma algebra.

**Proof.** For the first identity, note that $E(X|\mathcal{G}_1)$ is $\mathcal{G}_2$-measurable (because it is already $\mathcal{G}_1$-measurable) and the expectation of both sides agree on any event $A \in \mathcal{G}_2$. The proof of uniqueness of the conditional expectation thus applies to show equality a.s.

For the second identity we note that both sides are automatically $\mathcal{G}_1$-measurable. So picking $A \in \mathcal{G}_1$ and applying condition (2) from the definition of conditional expectation, we get

$$E(X1_A)_{A \in \mathcal{G}_2} = E\left( E\left( X|\mathcal{G}_2 \right) 1_A \right)_{A \in \mathcal{G}_2} = E\left( E\left( E\left( X|\mathcal{G}_2 \right) |\mathcal{G}_1 \right) 1_A \right)$$

(6.30)

Again, by the uniqueness argument, this implies the second identity as well. □

Finally, we observe the following useful fact:
Lemma 6.14  Suppose that $\sigma(X)$ and $\mathcal{G}$ are independent. Then

$$E(X|\mathcal{G}) = E(X) \quad \text{a.s.} \quad (6.31)$$

Proof. Constants are universally measurable so condition (1) in the definition of conditional expectation holds. For condition (2) we notice that for any $A \in \mathcal{G}$, the stated independence means

$$E(X1_A) = E(X)E(1_A) = E(E(X)1_A). \quad (6.32)$$

Hence, $E(X)$ serves as a version of $E(X|\mathcal{G})$. \qed
7. Conditional Distribution

The conditional expectation shares so many properties with expectation that one may wonder if it can also be realized as an integral with respect to a measure. This naturally yields to the consideration of conditional distribution and conditional probability. We start with the former object first leaving the latter to a later time in the course.

**Theorem 7.1** Let $X$ be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subset \mathcal{F}$ be a sigma algebra. Then there exists a map $\mu_X : \Omega \times B(\mathbb{R}) \to [0, 1]$ such that

1. for each $\omega \in \Omega$, $B \mapsto \mu_X(\omega, B)$ is a probability measure on $B(\mathbb{R})$,
2. for each $B \in B(\mathbb{R})$, $\omega \mapsto \mu_X(\omega, B)$ is $\mathcal{G}$-measurable, and
3. whenever $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable with $f(X) \in L^1(\Omega, \mathcal{F}, P)$, then

$$E\left(f(X) \mid \mathcal{G}\right)(\omega) = \int f(x) \mu_X(\omega, dx)$$

for a.e. $\omega$.

We will call $\mu_X(\omega, \cdot)$ the conditional distribution of $X$ given $\mathcal{G}$. Also note that the null set in (7.1) is permitted to depend on $f$.

**Proof.** For each $q \in \mathbb{Q}$, let us pick a version of $E(1_{\{X \leq q\}} \mid \mathcal{G})$ and use these objects to define the following sets: First, for $q < q'$ rational we set

$$\Omega_{q, q'} := \{ \omega \in \Omega : E(1_{\{X \leq q\}} \mid \mathcal{G})(\omega) \leq E(1_{\{X \leq q'\}} \mid \mathcal{G})(\omega) \}. \quad (7.2)$$

Next, for $q \in \mathbb{Q}$ we define

$$\Omega_q := \{ \omega \in \Omega : E(1_{\{X \leq q\}} \mid \mathcal{G})(\omega) = \inf_{q' > q, q' \in \mathbb{Q}} E(1_{\{X \leq q'\}} \mid \mathcal{G})(\omega) \}. \quad (7.3)$$

In addition, we also define

$$\Omega_{+\infty} := \{ \omega \in \Omega : \sup_{q \in \mathbb{Q}} E(1_{\{X \leq q'\}} \mid \mathcal{G})(\omega) = 1 \}$$

$$\Omega_{-\infty} := \{ \omega \in \Omega : \inf_{q \in \mathbb{Q}} E(1_{\{X \leq q'\}} \mid \mathcal{G})(\omega) = 0 \} \quad (7.4)$$

By the properties of the conditional expectation, all of these are full-measure events in $\mathcal{G}$. So, if we define

$$\Omega' := \left( \bigcap_{q' > q} \Omega_{q, q'} \right) \cap \left( \bigcap_{q \in \mathbb{Q}} \Omega_q \right) \cap \Omega_{+\infty} \cap \Omega_{-\infty}, \quad (7.5)$$

also $\Omega' \in \mathcal{G}$ with $P(\Omega') = 1$.

We now move to the construction of $\mu_X$. First set, for $\omega \in \Omega$ and $q \in \mathbb{Q}$,

$$F(\omega, q) := \begin{cases} E(1_{\{X \leq q\}} \mid \mathcal{G})(\omega), & \text{if } \omega \in \Omega', \\ 1_{\{q \geq 0\}}, & \text{otherwise}. \end{cases} \quad (7.6)$$
Then $q \mapsto F(\omega, q)$ is non-decreasing, right-continuous on $\mathbb{Q}$ with limits 1, resp., 0 as $q \to +\infty$, resp., $q \to -\infty$. We may thus extend it to a function on $\Omega \times \mathbb{R}$ by

$$F(\omega, x) := \inf_{q > x, q \in \mathbb{Q}} F(\omega, q)$$

(7.7)

and get a function which is, for each $\omega$, non-decreasing and right-continuous on $\mathbb{R}$ with the said limits at $\pm \infty$. In particular, $x \mapsto F(\omega, x)$ is a distribution function and so, by the existence theorems from 275A, there is a Borel probability measure $\mu_X(\omega, \cdot)$ such that

$$F(\omega, x) = \mu_X(\omega, (-\infty, x]), \quad x \in \mathbb{R}.$$  

(7.8)

It now remains to verify the state properties. Property (1) is ensured by the very construction. For property (2), we consider the class of sets

$$\mathcal{L} := \{ A \in \mathcal{B}(\mathbb{R}) : \omega \mapsto \mu_X(\omega, A) \text{ is } \mathcal{G} \text{-measurable} \}$$

(7.9)

It is easy to check that $\mathcal{L}$ contains $\Omega, \emptyset$, is closed under finite unions and complements, and also increasing limits and hence it is a sigma algebra. In light of (7.8) and the fact that $\omega \mapsto F(\omega, x)$ is $\mathcal{G}$-measurable, $\mathcal{L}$ contains all sets of the form $(-\infty, q]$ for $q \in \mathbb{Q}$. It follows that $\mathcal{L} = \mathcal{B}(\mathbb{R})$ thus proving (2). Condition (3) is verified similarly: First we note that (7.1) holds for $f$’s of the form $f(x) := 1_{\{x \leq q\}}$ for $q \in \mathbb{Q}$. The class of non-negative $f$’s satisfying (7.1) is closed under additions a monotone limits and hence (by the Monotone Class Theorem) it contains all non-negative measurable functions. The extension to integrable functions is then immediate. □
8. INTRODUCTION TO MARTINGALES

The concept of conditional expectation is useful in analyzing dependent random variables. Of course, one still desires that this dependence has a certain structure. We will explore two specific cases that go under the respective banners “Martingales” and “Markov chains.” Focusing now on the former, we begin with some definitions.

**Definition 8.1** Given a probability space \((\Omega, \mathcal{F}, P)\), a sequence \(\{\mathcal{F}_n\}_{n \geq 0}\) of sub-sigma-algebras \(\mathcal{F}_n \subset \mathcal{F}\) is said to be a filtration if \(\mathcal{F}_n \subset \mathcal{F}_{n+1}\) for all \(n \geq 0\). Given random variables \(M_n\), the pair \(\{M_n, \mathcal{F}_n\}_{n \geq 0}\) is then called a martingale if \(M_n\) is \(\mathcal{F}_n\)-measurable for each \(n \geq 0\) and

\[
E(M_{n+1}|\mathcal{F}_n) = M_n \quad \text{a.s. for each } n \geq 0. \tag{8.1}
\]

If instead an inequality holds here we call the sequence \(\{M_n, \mathcal{F}_n\}_{n \geq 0}\)
- a submartingale if \(E(M_{n+1}|\mathcal{F}_n) \geq M_n\) for all \(n \geq 0\),
- a supermartingale if \(E(M_{n+1}|\mathcal{F}_n) \leq M_n\) for all \(n \geq 0\).

A martingale is thus a submartingale and a supermartingale.

Here are some natural examples:

**Example 1 (Additive martingale):** Let \(X_1, X_2, \ldots\) be independent with \(X_i \in L^1\) and \(EX_i = 0\). Set \(M_n := X_1 + \cdots + X_n\) and \(\mathcal{F}_n := \sigma(X_1, \ldots, X_n)\). Then \(\{\mathcal{F}_n\}\) is a filtration and \(M_n\) is trivially \(\mathcal{F}_n\)-measurable. Moreover,

\[
E(M_{n+1}|\mathcal{F}_n) = M_n + E(X_{n+1}|\mathcal{F}_n) = M_n \tag{8.2}
\]

because \(X_{n+1}\) is independent of \(\mathcal{F}_n\) and thus \(E(X_{n+1}|\mathcal{F}_n) = E(X_{n+1}) = 0\). Hence \(\{M_n, \mathcal{F}_n\}\) is a martingale. Clearly, if instead \(E(X_i) \geq 0\), then it is a submartingale.

**Example 2 (Multiplicative martingale):** Let \(X_1, X_2, \ldots\) be independent with \(X_i \geq 0\) and \(EX_i = 1\). Set \(M_n := X_1 \cdots X_n\) and \(\mathcal{F}_n := \sigma(X_1, \ldots, X_n)\). Then again \(M_n\) is \(\mathcal{F}_n\)-measurable and

\[
E(M_{n+1}|\mathcal{F}_n) = M_n E(X_{n+1}|\mathcal{F}_n) = M_n. \tag{8.3}
\]

Hence \(\{M_n, \mathcal{F}_n\}\) is again a martingale.

**Example 3 (Progressive conditioning):** Let \(X \in L^1\) and let \(\{\mathcal{F}_n\}\) be any filtration. Set \(M_n := E(X|\mathcal{F}_n)\). Then

\[
E(M_{n+1}|\mathcal{F}_n) = E(E(X|\mathcal{F}_{n+1})|\mathcal{F}_n) = E(X|\mathcal{F}_n) = M_n \tag{8.4}
\]

by the “Smaller always wins” principle. Hence \(\{M_n, \mathcal{F}_n\}\) is a martingale. It is in fact an interesting question which martingales can be obtained this way. We will answer this when we talk about uniform integrability.

The next two examples concern systems that are of much independent interest in probability; yet they are good to introduce here because they give rise to natural martingales.

**Example 4 (Galton-Watson branching process):** In mid 19th century, there was some interest to develop a theory of family trees — particularly, in connection with royal families. Two statisticians (by back then standards) Galton and Watson devised a model that nowadays bears their...
name. In this model, there are generations containing a certain number of currently-alive individuals. The dynamics is such that, at each unit of time, each individual produces a certain number of off-spring, which is sampled independently from a common law with probabilities \( \{ p(n) : n \geq 0 \} \). (In particular, if there is no off-spring, the lineage of that individual dies out.)

We will define the problem as follows. Consider a family of i.i.d. integer-valued random variables \( \{ \xi_{n,k} : n, k \geq 0 \} \) with law determined by
\[
P(\xi_{n,k} = m) = p(m), \quad m \geq 0.
\] (8.5)

Define, inductively, random variables \( \{ S_n \} \) as follows:
\[
S_0 := 1 \quad \text{and} \quad S_{n+1} := \begin{cases} 0, & \text{if } S_n = 0, \\ \xi_{n+1,1} + \cdots + \xi_{n+1,S_n}, & \text{if } S_n > 0. \end{cases}
\] (8.6)

It is easy to verify that the dynamics does what we described verbally above. The additional assumption \( S_0 := 1 \) means that there is one individual at time zero.

Consider now the filtration \( \mathcal{F}_n := \sigma(\xi_{m,k} : 0 \leq m \leq n, k \geq 0) \) and let us compute:
\[
E(S_{n+1} | \mathcal{F}_n) = \begin{cases} 0, & \text{on } \{ S_n = 0 \}, \\ [E(\xi_{1,1})]S_n, & \text{on } \{ S_n > 0 \}. \end{cases}
\] (8.7)

Thus, denoting \( \mu := E(\xi_{1,1}) \) we get that \( M_n := \mu^{-n}S_n \) is a martingale. The value \( \mu = 1 \) is obviously special because \( S_n \) is then itself a martingale, and we will later see this is indeed the case even for the behavior of the model.

Example 5 (Polya’s Urn): Our next example concerns the following problem: Consider an urn that has, initially, \( r \) red balls and \( g \) green balls in it. We now proceed as follows: Sample a ball from the urn and replace it back along with another ball of the same color. Repeating the dynamics, the question is what is status of the urn in the long run.

Let \( R_n \) denote the number of red balls in the urn at time \( n \) and let \( G_n \) denote the corresponding number of green balls. Obviously,
\[
R_n + G_n = r + g + n.
\] (8.8)

Now consider the random variable
\[
M_n := \frac{R_n}{R_n + G_n} = \frac{R_n}{r + g + n}
\] (8.9)

which denotes the fraction of red balls in the urn. The dynamics can then be encoded using a sequence \( U_1, U_2, \ldots \) of i.i.d. uniform random variables on \([0, 1]\) as follows:
\[
R_{n+1} := (R_n + 1)1_{\{ U_{n+1} \leq M_n \}} + R_n 1_{\{ U_{n+1} > M_n \}}.
\] (8.10)
We claim that \( \{M_n, \mathcal{F}_n\} \) is a martingale for the filtration \( \mathcal{F}_n := \sigma(U_1, \ldots, U_n) \). Indeed, since \( R_n \) is \( \mathcal{F}_n \)-measurable, we have

\[
E(M_{n+1} | \mathcal{F}_n) = \frac{R_n + 1}{r + g + n + 1} E(1_{U_{n+1} \leq M_n} | \mathcal{F}_n) + \frac{R_n}{r + g + n + 1} E(1_{U_{n+1} > M_n} | \mathcal{F}_n)
\]

(8.11)

Notice that, unlike the earlier examples this martingale is non-negative and bounded.

We finish by noting certain stability of (sub)martingales under composition with a convex function:

**Lemma 8.2** Let \( \varphi : \mathbb{R} \to \mathbb{R} \) be convex and bounded from below and obey \( |\varphi(x)| \leq C|x| \) for some \( C > 0 \). Then

1. \( \{M_n, \mathcal{F}_n\} \) is a martingale \( \Rightarrow \{\varphi(M_n), \mathcal{F}_n\} \) is a submartingale.
2. \( \{M_n, \mathcal{F}_n\} \) is a submartingale and \( \varphi \) is non-decreasing \( \Rightarrow \{\varphi(M_n), \mathcal{F}_n\} \) is a submartingale.

**Proof.** The conditions ensure that \( \varphi(M_n) \) is integrable. By Jensen’s inequality

\[
E\left(\varphi(M_{n+1}) | \mathcal{F}_n\right) \geq \varphi\left(E(M_{n+1} | \mathcal{F}_n)\right)
\]

(8.12)

The right hand side equals \( \varphi(M_n) \) when \( \{M_n, \mathcal{F}_n\} \) is a martingale. If the latter is only a submartingale, then we still have

\[
\varphi\left(E(M_{n+1} | \mathcal{F}_n)\right) \geq \varphi(M_n)
\]

(8.13)

as soon as \( \varphi \) is non-decreasing. \( \square \)
9. Optional Stopping Theorem

In this section we will expand on the following property of martingales: Suppose \( \{M_n, \mathcal{F}_n\} \) is a martingale. Then
\[
E(M_{n+1}) = E(E(M_{n+1} | \mathcal{F}_n)) = E(M_n) = \overbrace{E(M_0)}^{\text{induction}}
\]
and so \( E(M_n) \) does not depend on \( n \). Similarly, for submartingales we get that \( E(M_n) \) is non-decreasing while for supermartingales \( E(M_n) \) is non-increasing. The question to address here is whether \( E(M_n) = E(M_0) \) is true (for martingales) when we replace \( n \) by a random time \( T \). Of course, not just any random time will do, but the following concept is relevant:

**Definition 9.1** A random variable \( T \) taking values in \( \{0, 1, 2, \ldots\} \cup \{\infty\} \) is called a stopping time for filtration \( \{\mathcal{F}_n\} \) if
\[
\{T \leq n\} \in \mathcal{F}_n, \quad \forall n = 0, 1, \ldots
\]

We note that equivalent concept is obtained when \( \{T \leq n\} \) is replaced by \( \{T > n\} \) or \( \{T = n\} \). However, \( \{T < n\} \) leads to a different notion (often called the optional time). We will given example of a stopping time associated with a simple random walk but for that let us define random walks first:

**Definition 9.2** A random walk is a collection of random variables \( \{S_n\} \) where \( S_n := S_0 + X_1 + \cdots + X_n \), where \( X_1, X_2, \ldots \) are i.i.d. If \( X_i \) take values \( \pm 1 \) with equal probabilities, and \( S_0 \in \mathbb{Z} \), then we refer to this as the simple (symmetric) random walk on \( \mathbb{Z} \).

Consider now the simple symmetric random walk on \( \mathbb{Z} \) started from \( S_0 := 0 \) and let \( a < 0 < b \) be two integers. Define
\[
T := \inf\{n \geq 0: S_n \not\in (a, b)\}
\]

This is the first time the random walk has left the interval \((a, b)\) — which necessarily happens by hitting either \( a \) or \( b \). Consider the filtration \( \mathcal{F}_n := \sigma(S_0, \ldots, S_n) \). To see that \( T \) is indeed a stopping time, we now observe that
\[
\{T > n\} = \bigcap_{m=0}^{n} \{S_m \not\in (a, b)\}.
\]

We now return back to the question of whether martingales evaluated at stopping times are still martingales:

**Lemma 9.3** Suppose \( \{M_n, \mathcal{F}_n\} \) is a (sub)martingale and \( T \) a stopping time for \( \{\mathcal{F}_n\} \). Then \( M_{T \wedge n} \in L^1 \) for all \( n \) and, in fact, \( \{M_{T \wedge n}, \mathcal{F}_n\} \) is also a (sub)martingale.

**Proof.** Since \( T \wedge n \) is \( \mathcal{F}_n \)-measurable, so is \( M_{T \wedge n} \). Moreover,
\[
|M_{T \wedge n}| \leq |M_1| + \cdots + |M_n|
\]
and so \( M_{T \wedge n} \in L^1 \). For the proof of (sub)martingale property, notice the following rewrite
\[
M_{T \wedge (n+1)} = M_{T \wedge n} 1_{\{T \leq n\}} + M_{n+1} 1_{\{T > n\}}
\]
The first summand on the right is \( F_n \)-measurable and so is the indicator in the second summand. Hence, For \( \{M_n, F_n\} \) martingale,

\[
E\left(M_{T \wedge (n+1)} \mid F_n\right) = M_{T \wedge n} 1_{\{T \leq n\}} + 1_{\{T > n\}} E(M_{n+1} \mid F_n) \\
= M_{T \wedge n} 1_{\{T \leq n\}} + M_n 1_{\{T > n\}} = M_{T \wedge n},
\]

where the last conclusion is checked by considering separately the cases \( T \leq n \) and \( T > n \). For submartingales the second equality is changed into \( \geq \).

**Theorem 9.4** (Doob’s Optional Stopping Theorem) Let \( \{M_n, F_n\} \) be a martingale and \( T \) a stopping time for \( \{F_n\} \). If for some constant \( K \in (0, \infty) \) any of the conditions

1. \( T \leq K \) a.s.,
2. \( |M_n| \leq K \) a.s. for all \( n \geq 0 \) and \( T < \infty \) a.s.,
3. \( |M_{n+1} - M_n| \leq K \) a.s. for all \( n \geq 0 \) and \( ET < \infty \),

are satisfied then \( M_T \in L^1 \) and \( E(M_T) = E(M_0) \).

**Proof.** By the previous lemma we have \( M_{T \wedge n} \in L^1 \) and \( E(M_{T \wedge n}) = E(M_0) \). When \( T \leq K \) a.s., setting \( n > K \) makes truncation by \( n \) superfluous and so the claim follows in situation (1).

For (2-3) we will now obtain statements by using an appropriate convergence theorem. In both cases we have \( T < \infty \) a.s. and so

\[
M_{T \wedge n} \xrightarrow{n \to \infty} M_T, \quad \text{a.s.}
\]

(9.8)

In (2) we have \( |M_{T \wedge n}| \leq K \) and so \( E(M_{T \wedge n}) \to E(M_T) \) by the Bounded Convergence Theorem. This shows that \( M_T \in L^1 \) and that \( E(M_T) = E(M_0) \).

In (3) we instead note the following bound

\[
|M_{T \wedge n}| \leq |M_0| + KT, \quad n \geq 0,
\]

(9.9)

and since the right-hand side is in \( L^1 \), the Dominated Convergence Theorem yields the corresponding conclusion as well.

To illustrate the above, consider again the simple symmetric random walk started at \( S_0 := 0 \) and define the stopping times

\[
T_a := \inf\{n \geq 0 : S_n = a\}, \quad a \in \mathbb{Z}.
\]

(9.10)

The stopping time in our earlier example above was then \( T := T_a \wedge T_b \). It is a matter of a simple exercise to show that \( T < \infty \) a.s. Since also \( S_{T \wedge n} \) is bounded between \( a \) and \( b \), the argument in the proof for case (2) implies

\[
E(S_T) = E(S_0) = 0.
\]

(9.11)

This innocuous expression can be further developed by noting that \( S_T = a \) when \( T_a < T_b \) while \( S_T = b \) when \( T_b > T_a \). Again, by \( T < \infty \) these two possibilities effectively partition the probability space and so we have

\[
0 = E(S_T) = aP(T_a < T_b) + bP(T_b < T_a) \\
= a[1 - P(T_b < T_a)] + bP(T_b < T_a)
\]

(9.12)
From here we compute
\[ P(T_b < T_a) = \frac{b}{b-a}. \] (9.13)

Notice that, by taking \( b \to \infty \) this implies
\[ P(T_a < \infty) = 1 \] (9.14)
for all \( a \in \mathbb{Z} \). In simple terms, the random walk visits every integer with probability one.

To see that some conditions are needed in the conclusion of the Optional Stopping Theorem, notice that \( T_a < \infty \text{ a.s.} \) implies \( S_{T_a} = a \text{ a.s.} \). But then \( E(S_{T_a}) = a \neq E(S_0) = 0 \) whenever \( a \neq 0 \).

Looking at the case (3), since the increments \( S_{n+1} - S_n \) are at most one, we must therefore have
\[ E(T_a) = \infty, \quad a \neq 0. \] (9.15)

The random walk visits every integer a.s. but the expected time to do so is infinite, except at the starting point, of course.
10. Martingale Convergence Theorem

Our next task is to study the behavior of the sequence $M_n$, for $\{M_n\}$ a sub-or-super martingale, as $n \to \infty$. The principal result to prove is:

**Theorem 10.1** (Doob’s Martingale Convergence Theorem) Suppose $\{X_n, \mathcal{F}_n\}$ is a supermartingale with $\sup_{n \geq 0} E(X_n^-) < \infty$. Then $X_\infty := \lim_{n \to \infty} X_n$ exists a.s. and $X_\infty \in L^1$.

To ease our way of speech, let us introduce some vernacular from the theory of discrete-time stochastic processes: A process is a family of random variables $\{X_n\}$ indexed by $n \in \{0, 1, \ldots\}$. Given a filtration $\{\mathcal{F}_n\}$, a process $C := \{C_n\}$ is said to be predictable if $C_n$ is $\mathcal{F}_{n-1}$-measurable for each $n \geq 1$. A process $X := \{X_n\}$ is said to be adapted if $X_n$ is $\mathcal{F}_n$-measurable for each $n \geq 0$. The proof of the above theorem will need the following concept:

**Definition 10.2** Given a predictable process $C := \{C_n\}$ and an adapted process $X := \{X_n\}$, we define $C \cdot X$ to be the process

$$
(C \cdot X)_n := \sum_{k=1}^{n} C_k (X_k - X_{k-1}). \tag{10.1}
$$

Obviously, $C \cdot X$ is also adapted. Moreover, for special $X$ we have the following:

**Lemma 10.3** Given a filtration $\{\mathcal{F}_n\}$, let $C$ be a predictable process. Then

1. $X$ is a martingale $\Rightarrow$ $C \cdot X$ is a martingale
2. $X$ is a sub/supermartingale and $C_n \geq 0 \Rightarrow C \cdot X$ is a sub/supermartingale

**Proof.** Let $Y_n := (C \cdot X)_n - (C \cdot X)_{n-1}$ denote the increment of $C \cdot X$ process. To show (1) it suffices to prove $E(Y_{n+1} | \mathcal{F}_n) = 0$, for (2) we need to show that $E(Y_{n+1} | \mathcal{F}_n) \geq 0$. Clearly,

$$
Y_n = C_n (X_n - X_{n-1}) \tag{10.2}
$$

Then, since $C_n$ and $X_{n-1}$ are both $\mathcal{F}_{n-1}$-measurable,

$$
E(Y_{n+1} | \mathcal{F}_n) = C_n \left[ E(X_{n+1} | \mathcal{F}_n) - X_n \right] \tag{10.3}
$$

If $X$ is a martingale, then the bracket vanishes. If $X$ is a submartingale, the bracket is non-negative and so the whole expression is non-negative once $C_n \geq 0$. □

The object $C \cdot X$ has two natural interpretations. First, it is a discrete version of a stochastic integral $\int C_t \, dX_t$ which we will talk about next quarter. Second, it has the interpretation of the net winning in a game where $C_n$ is a bet placed at time $n$ and $X_n - X_{n-1}$ represents the net winning per unit bet. The above lemma shows that in a fair game — which corresponds to $X$ being a martingale — there is no predictable strategy — by which we mean a strategy that does not have a peek into the future — to guarantee positive expected winning.

Returning to our main line of thought, the key tool in the proof of the above theorem is the control of the number of crossings of an interval by a process $X$. Let $a < b$ be two reals. Define the following sequence of stopping times $T'_0 < T'_1 < T'_1 < T'_2 < T'_2 < \ldots$ as follows: $T'_0 := 0$ and,
inductively, for \( i \geq 1, \)
\[
T_i := \inf\{ n > T_{i-1}' : X_n < a \}
\]
\[
T_i' := \inf\{ n > T_i : X_n > b \}
\tag{10.4}
\]
Obviously, \( T_i \) is the \( i \)-th time the process runs below \( a \) while \( T_i' \) is the time right after that when the process went above \( b \). Define
\[
U_n[a, b] := \max\{ k \geq 0 : T_k' \leq n \}.
\tag{10.5}
\]
This denotes the number of completed upcrossings of interval \([a, b]\) by time \( n \). Clearly \( U_n[a, b] \leq n \) and \( n \mapsto U_n[a, b] \) is increasing in \( n \).

**Lemma 10.4 (Upcrossing inequality)** Suppose \( X \) is a supermartingale. Then
\[
EU_n[a, b] \leq \frac{E[ (X_n - a)^- ]}{b - a}
\tag{10.6}
\]

**Proof.** Define a process \( C \) as follows: \( C_0 := 0 \) and, inductively,
\[
C_{n+1} := 1_{\{C_n = 1\}} 1_{\{X_n \leq b\}} + 1_{\{C_n = 0\}} 1_{\{X_n < a\}}
\tag{10.7}
\]
It is easy to check that \( C_k = 1 \) for \( k \in (T_i, T_i') \) for some \( i \) and \( C_k = 0 \) otherwise. Now consider the process \( C \cdot X \). Writing \( U_n \) for \( U_n[a, b] \), the special form of \( C \) then yields
\[
(C \cdot X)_n = \sum_{i=1}^{U_k} (X_{T_i'} - X_{T_i}) + 1_{\{T_{i_{k+1}} < n\}} (X_n - X_{T_{i_{k+1}}})
\tag{10.8}
\]
But \( X_{T_i'} - X_{T_i} \geq b - a \) while \( X_n - X_{T_{i_{k+1}}} \geq X_n - a \geq -(X_n - a)^- \) and so
\[
(C \cdot X)_n \geq (b - a)U_n[a, b] - (X_n - a)^-
\tag{10.9}
\]
On the other hand, since \( C_n \geq 0 \) and \( X \) is a supermartingale,
\[
E((C \cdot X)_n) \leq 0.
\tag{10.10}
\]
Putting these together we get
\[
0 \leq (b - a)EU_n[a, b] - E[(X_n - a)^-]
\tag{10.11}
\]
This gives the claim by a simple manipulation. \(\square\)

**Proof of Theorem 10.1.** The monotonicity of \( n \mapsto U_n[a, b] \) permits us to define
\[
U_\infty[a, b] := \lim_{n \to \infty} U_n[a, b].
\tag{10.12}
\]
First we note that
\[
\{ \lim_{n \to \infty} X_n \text{ does not exist in } [-\infty, \infty] \} = \bigcup_{a, b \in \mathbb{Q}} \{ U_\infty[a, b] = \infty \}.
\tag{10.13}
\]
Indeed, the event on the left really reads \( \limsup_{n \to \infty} X_n > \liminf_{n \to \infty} X_n \). It is easy to check that the limsup and liminf are different if and only if there exists an interval with rational endpoints which is (up)crossed infinitely many times by the sequence.
Thanks to the assumption that $E(X_n^-)$ are uniformly bounded we have
\[
\sup_{n \geq 0} E \left[ (X_n - a)^- \right] \leq |a| + \sup_{n \geq 0} E(X_n^-) < \infty
\]
and so, by the Upcrossing Inequality, also
\[
\sup_{n \geq 0} EU_n[a, b] < \infty.
\]
Invoking the Monotone Convergence Theorem we then get that $EU_\infty[a, b] < \infty$ and thus $U_\infty[a, b] < \infty$ a.s. Hence, every event in the countable union in (10.13) is a null event and so
\[
X_n \xrightarrow{n \to \infty} X_\infty := \limsup_{n \to \infty} X_n, \quad \text{a.s.} \tag{10.16}
\]
It remains to show that $X_\infty \in L^1$. For that we note that Fatou’s Lemma ensures
\[
E(X_\infty^-) \leq \liminf_{n \to \infty} E(X_n^-) \leq \sup_{n \geq 0} E(X_n^-) < \infty
\]
and so the lower tail of $X_\infty$ is integrable. For the upper tail we write $X_n^+ = X_n + X_n^-$ and so
\[
E(X_n^+) = E(X_n) + E(X_n^-) \leq E(X_0) + \sup_{n \geq 0} E(X_n^-) < \infty
\]
where the middle inequality follows because $X$ is a supermartingale. Invoking Fatou’s Lemma, we get $E(X_\infty^+) < \infty$ as well. \hfill \square

As we will see later, the Martingale Convergence Theorem has many profound consequences. The advantage is that one only has to have a uniform bound on the integral of of the tail towards which the expectations go — the lower tail for supermartingales and upper tail for submartingales. For martingales it suffices to check either one. A useful special case is:

**Corollary 10.5** A non-negative supermartingale converges almost surely to an integrable random variable.

We can apply this to the case of Polya’s Urn: Indeed, recalling the notations $R_n$, resp., $G_n$ for the number of red, resp., green balls in the urn we have $R_n + G_n = r + g + n$, where $r$, resp., $g$ are the initial numbers. We have already checked that
\[
M_n := \frac{R_n}{r + g + n} \tag{10.19}
\]
is a martingale for a filtration introduced earlier. Since $M_n \in [0, 1]$, the above corollary applies and $M_n$ converges a.s. to a random variable $M_\infty$. In words, the fraction of the red balls in the urn tends to a limit with probability one.
11. Uniform Integrability

Having shown that martingales under rather general conditions converge, we may recall that martingales obey $EM_n = EM_0$ and ask whether $M_n \to M_\infty$ a.s. implies $EM_\infty = EM_0$. (Note that Fatou’s Lemma gives us only $EM_\infty^+ \leq EM_0^+$. ) It turns out that the concept to invoke is as in:

**Definition 11.1**  A family of random variables $\{X_\alpha\}_{\alpha \in I}$ is uniformly integrable (UI) if

$$\forall \epsilon > 0 \exists K < \infty \forall \alpha \in I : E\left(|X_\alpha|1_{\{|X_\alpha| \geq K\}}\right) < \epsilon.$$  \hspace{1cm} (11.1)

Obviously, uniform integrability implies integrability and, in fact, a uniform containment in $L^1$, expressed as a uniform bound on the $L^1$-norm:

$$\|X_\alpha\|_1 \leq K + E\left(|X_\alpha|1_{\{|X_\alpha| \geq K\}}\right) \leq K + \epsilon.$$  \hspace{1cm} (11.2)

We will now state a couple of rather straightforward consequences of uniform integrability. First a couple of trivialities:

**Lemma 11.2**  If $X \in L^1$ then $\{X\}$ is UI.

**Proof.** By the Dominated Convergence Theorem we have $E(|X|1_{\{|X| \geq K}\}) \to 0$ as $K \to \infty$. \hspace{1cm} \square

**Lemma 11.3**  If $\{X_{\alpha,j} : \alpha \in I\}$ is UI for each $j = 1, \ldots, n$, then so is $\{X_{\alpha,j} : \alpha \in I, j = 1, \ldots, n\}$.

**Proof.** The uniform bound on $E(|X_{\alpha,j}|1_{\{|X_{\alpha,j}| \geq K\}})$ for each $j$ can be made, by increasing $K$ further, into a bound that applies to all $j$ simultaneously. \hspace{1cm} \square

The next property is perhaps more substantive:

**Proposition 11.4**  We have:

$$X_n \to X \text{ in } L^1 \iff \begin{cases} X_n \to X \text{ in probability, AND} \\ \{X_n\} \text{ is UI.} \end{cases}$$  \hspace{1cm} (11.3)

For the proof we need the following simple observation whose proof is left to HW:

**Lemma 11.5**  Let $X \in L^1(\Omega, \mathcal{F}, P)$. Then

$$\forall \epsilon > 0 \exists \delta > 0 \forall A \in \mathcal{F} : P(A) < \delta \Rightarrow E(|X|1_A) < \epsilon.$$  \hspace{1cm} (11.4)

**Proof of “⇒” in Proposition 11.4.** Suppose $X_n \to X$ in $L^1$. Then

$$P(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon} \|X_n - X\|_1 \to 0$$  \hspace{1cm} (11.5)

and so $X_n \to X$ in probability. To get that $\{X_n\}$ is UI, we note that the bound

$$|X_n| \leq |X| + |X - X_n|$$  \hspace{1cm} (11.6)

implies

$$\begin{aligned} E(|X_n|1_{\{|X_n| \geq K\}}) &\leq E|X - X_n| + E(|X|1_{\{|X| \geq K\}}) \\ &\leq E|X - X_n| + E(|X|1_{\{|X| \geq K/2\}}) + E(|X|1_{\{|X - X_n| \geq K/2\}}) \end{aligned}$$  \hspace{1cm} (11.7)
Now fix \( \varepsilon > 0 \) and pick \( K \) so large that the second term is at most \( \varepsilon \). Next apply Lemma 11.5 to find \( \delta > 0 \) so that \( P(A) < \delta \) implies \( E(|X|1_A) < \varepsilon \). Since \( X_n \rightarrow X \) in \( L^1 \), there is \( n(\varepsilon) \) such that
\[
  n \geq n(\varepsilon) \quad \Rightarrow \quad E|X_n - X| < \varepsilon \quad \text{AND} \quad P(|X_n - X| \geq K) < \delta. \tag{11.8}
\]
It follows that
\[
  E(|X_n|1_{\{|X_n| \geq K\}}) \leq 3\varepsilon, \quad n \geq n(\varepsilon). \tag{11.9}
\]
But increasing \( K \) further, we can make the same to be true for \( n < n(\varepsilon) \) as well and so \( \{X_n\} \) is indeed UI.

**Proof of “\( \Leftarrow \)” in Proposition 11.4.** Assume that \( X_n \rightarrow X \) in probability with \( \{X_n\} \) UI. We will first prove that \( X \in L^1 \). For that we note that, by choosing a subsequence \( \{n_k\} \), we can ensure \( X_{n_k} \rightarrow X \) a.s. Fatou’s Lemma gives us
\[
  E(|X|1_{\{|X| \geq K\}}) \leq \liminf_{k \to \infty} E(|X_{n_k}|1_{\{|X_{n_k}| \geq K\}}), \tag{11.10}
\]
and so \( E(|X|1_{\{|X| \geq K\}}) < \varepsilon \) for \( K \) sufficiently large. A straightforward estimate now shows that, since \( \{X_n\} \) is UI and \( X \in L^1 \), then also \( \{X_n - X\} \) is UI. Hence, for any \( 0 < \varepsilon < K < \infty \) we have
\[
  E|X_n - X| \leq \varepsilon + K P(\varepsilon < |X_n - X| \leq K) + E(|X_n - X|1_{\{|X_n - X| \geq K\}}). \tag{11.11}
\]
The third term on the right can be made less than \( \varepsilon \) by choosing \( K \) sufficiently large. Since \( X_n \rightarrow X \) in probability, the second term actually tends to zero as \( n \to \infty \). The \( L^1 \)-convergence follows. \( \Box \)

We add two natural sufficient conditions for uniform integrability:

**Lemma 11.6** Suppose that \( \{X_\alpha\} \) is such that \( \sup_\alpha \|X_\alpha\|_p < \infty \) for some \( p > 1 \). Then \( \{X_\alpha\} \) is UI.

**Proof.** By Chebyshev’s inequality
\[
  E(|X_\alpha|1_{\{|X_\alpha| \geq K\}}) \leq \frac{\|X_\alpha\|_p}{K^{p-1}} \leq \frac{1}{K^{p-1}} \sup_\alpha \|X_\alpha\|_p. \tag{11.12}
\]
Choosing \( K \) large, this can be made smaller than any \( \delta > 0 \) prescribed, uniformly in \( \alpha \). \( \Box \)

**Lemma 11.7** Suppose that \( \{X_\alpha\} \) are such that \( |X_\alpha| \leq Y \) for some \( Y \in L^1 \). Then \( \{X_\alpha\} \) is UI.

**Proof.** Since \( x \mapsto x1_{\{x \geq K\}} \) is non-decreasing on \( (0, \infty) \) we have
\[
  E(|X_\alpha|1_{\{|X_\alpha| \geq K\}}) \leq E(|Y|1_{\{|Y| \geq K\}}), \tag{11.13}
\]
for all \( \alpha \). Increasing \( K \), the right-hand side can be made as small as desired thanks to \( Y \in L^1 \) (using argument in the proof of Lemma 11.2). \( \Box \)

Here is one less trivial way to produce UI families of random variables:

**Lemma 11.8** Let \( X \in L^1 \). Then \( \{E(X|G) : \mathcal{G} \subseteq \mathcal{F} \text{ sigma algebra}\} \) is UI.

**Proof.** Fix \( \varepsilon > 0 \) and let \( \delta > 0 \) be such that \( P(A) < \delta \) implies \( E(|X|1_A) < \varepsilon \). (See Lemma 11.5.) Next choose \( K > 0 \) so that \( E|X| < K\delta \). Given a sigma algebra \( \mathcal{G} \subseteq \mathcal{F} \), abbreviate \( Y := E(X|\mathcal{G}) \) and note
\[
  P(|Y| \geq K) \leq \frac{1}{K} E|Y| \leq \frac{1}{K} E|X| < \delta, \tag{11.14}
\]
where we used that $\|Y\|_1 \leq \|X\|_1$ by Lemma 6.8. Our choice of $\delta$ then yields $E(|X|1_{\{|Y| \geq K\}}) < \epsilon$. But then

$$E(|Y|1_{\{|Y| \geq K\}}) \leq E(E(|X|\|G\|1_{\{|Y| \geq K\}}) = E(|X|1_{\{|Y| \geq K\}}) < \epsilon,$$

where we used Lemma 6.8 to get the first inequality.
12. UI Martingales

Our next item of business is to apply the concept of uniform integrability to martingales. Here is a consequence of UI property:

**Lemma 12.1** For sub/supermartingales \( \{X_n\} \), the following are equivalent:

1. \( \{X_n\} \) is UI.
2. \( X_n \) converges a.s. and in \( L^1 \).
3. \( X_n \) converges in \( L^1 \).

**Proof.** We have (2) \( \Rightarrow \) (3) trivially and (3) \( \Rightarrow \) (1) by Proposition 11.4. To show (1) \( \Rightarrow \) (2), note \( \{X_n\} \) UI implies \( \sup_{n \geq 1} E|X_n| < \infty \). So \( X_n \to X_\infty \) a.s. by the Martingale Convergence Theorem. Uniform integrability then yields \( L^1 \) convergence as well. \( \square \)

Next we observe that once a martingale converges in \( L^1 \), it can be represented by conditioning from the limiting random variable.

**Lemma 12.2** If \( \{M_n, \mathcal{F}_n\} \) is a martingale and \( M_n \to M_\infty \) in \( L^1 \) (or, equivalently, if \( \{M_n\} \) is UI), then \( M_n = E(M_\infty | \mathcal{F}_n) \) a.s. for each \( n \geq 1 \).

**Proof.** By \( L^1 \)-convergence, \( E(M_k 1_A) \to E(M_\infty 1_A) \) as \( k \to \infty \) for each \( A \in \mathcal{F} \). On the other hand, for \( A \in \mathcal{F}_n \) we have \( E(M_k 1_A) = E(M_n 1_A) \) for each \( k \geq n \). Hence

\[
A \in \mathcal{F}_n \implies E(M_n 1_A) = E(M_\infty 1_A). \quad (12.1)
\]

Since \( M_n \) is \( \mathcal{F}_n \)-measurable, it serves as a version of \( E(M_\infty | \mathcal{F}_n) \). \( \square \)

We now take a different perspective and prove:

**Theorem 12.3** (Levy Forward Theorem) Let \( X \in L^1 \) and let \( \{F_n\} \) be a filtration. Denote

\[
\mathcal{F}_\infty := \sigma \left( \bigcup_{n \geq 1} \mathcal{F}_n \right). \quad (12.2)
\]

Then we have

\[
E(X | \mathcal{F}_n) \xrightarrow{n \to \infty} E(X | \mathcal{F}_\infty), \quad \text{a.s. and in } L^1. \quad (12.3)
\]

**Proof.** Splitting \( X \) into the positive and negative parts, we may assume \( X \geq 0 \). Set \( X_n := E(X | \mathcal{F}_n) \) (choosing a version for each \( n \)) and denote

\[
X_\infty := \limsup_{n \to \infty} X_n. \quad (12.4)
\]

Since \( \{X_n, \mathcal{F}_n\} \) is a UI martingale, we have \( X_n \to X_\infty \) a.s. and in \( L^1 \). We thus only have to show that \( X_\infty \) is a version of \( E(X | \mathcal{F}_\infty) \).

First \( X_\infty \) is \( \mathcal{F}_\infty \)-measurable because \( X_n \) is \( \mathcal{F}_n \)-measurable and thus \( \mathcal{F}_\infty \)-measurable. To prove the other defining property of conditional expectation, define two probability measures

\[
v_1(A) := E(E(X | \mathcal{F}_\infty) 1_A) \quad \text{and} \quad v_2(A) := E((X_\infty) 1_A) \quad (12.5)
\]
on $\mathcal{F}$. We claim that $v_1 = v_2$ on $\bigcup_{n \geq 1} \mathcal{F}_n$. Indeed, $A \in \mathcal{F}_n$ implies $E(X_k 1_A) = E(X_n 1_A)$ once $k \geq n$ and thus, by $L^1$ convergence $X_n \to X_\omega$, also $v_2(A) := E(X_\omega 1_A) = E(X_n 1_A)$. On the other hand, $E(E(X | \mathcal{F}_\omega) 1_A) = E(X_n 1_A)$ by the “smaller-always-wins” principle and thus $v_2(A) = v_1(A)$.

The class $\{ A \in \mathcal{F} : v_1(A) = v_2(A) \}$ is a $\lambda$-system that contains the $\pi$-system $\bigcup_{n \geq 1} \mathcal{F}_n$. Thus it contains $\mathcal{F}_\infty$ as well. It follows that $X_n = E(X | \mathcal{F}_\infty)$ a.s. □

The Levy Forward Theorem gives us a very elegant:

**Proof of Kolmogorov’s Zero-One Law.** Suppose $Y_1, Y_2, \ldots$ are independent and let

$$\mathcal{F}_n := \sigma(Y_{n+1}, Y_{n+2}, \ldots) \quad \text{and} \quad \mathcal{F} := \bigcap_{n \geq 1} \mathcal{F}_n$$

(12.6)

Our goal is to show that all events from $\mathcal{F}$ are trivial. Denote also

$$\mathcal{F}_n := \sigma(Y_1, \ldots, Y_n) \quad \text{and} \quad \mathcal{F}_\infty := \sigma(Y_1, Y_2, \ldots).$$

(12.7)

Since every $A \in \mathcal{F}$ is independent of every $B \in \mathcal{F}_n$, we have $E(1_A | \mathcal{F}_n) = P(A)$ a.s. On the other hand, the Levy Forward Theorem yields

$$E(1_A | \mathcal{F}_n) \xrightarrow{n \to \infty} E(1_A | \mathcal{F}_\infty) = 1_A \quad \text{a.s.}$$

(12.8)

where we used that $A$ is $\mathcal{F}_\infty$-measurable. Hence $1_A = P(A)$ a.s. and so $P(A) \in \{0, 1\}$. □

Applications naturally lead us to consider the behavior of $E(X | \mathcal{F}_n)$ for decreasing sequences of sigma algebras as well. This is addressed in:

**Theorem 12.4 (Levy Backward Theorem)** Let $X \in L^1$ and let $\{ \mathcal{F}_n \}$ be sigma algebras such that $\mathcal{F}_{n+1} \subseteq \mathcal{F}_n$ for each $n \geq 1$. Abbreviate

$$\mathcal{F}_\infty := \bigcap_{n \geq 1} \mathcal{F}_n.$$  

(12.9)

Then we have

$$E(X | \mathcal{F}_n) \xrightarrow{n \to \infty} E(X | \mathcal{F}_\infty), \quad \text{a.s. \& in } L^1.$$  

(12.10)

**Proof.** We need to temporarily reverse time to establish convergence. Fix $N \geq 1$ and denote

$$\mathcal{G}_n := \mathcal{F}_{(N-n) \vee 1} \quad \text{and} \quad Y_n := E(X | \mathcal{G}_n)$$

(12.11)

Since $\{ \mathcal{G}_n \}$ is a filtration, $\{Y_n, \mathcal{G}_n\}$ is a UI-martingale. In particular, if $U_n[a, b]$ denotes the number of completed upcrossings of $[a, b]$ by $\{Y_1, \ldots, Y_n\}$, then Doob’s Upcrossing Inequality implies

$$EU_n[a, b] \leq EY_N + |a| \leq \frac{E|X| + |a|}{b - a}$$

(12.12)

Next abbreviate $X_n := E(X | \mathcal{F}_n)$ and let $D_n[a, b]$ denote the number of downcrossing of $[a, b]$ by $\{X_1, \ldots, X_n\}$. Then

$$D_n[a, b] \leq U_n[a, b] + 1$$  

(12.13)

and so $\sup_{n \geq 1} ED_n[a, b] < \infty$. In particular, the monotone limit $D_\infty[a, b] := \lim_{n \to \infty} D_n[a, b]$ is finite a.s. thanks to the Monotone Convergence Theorem and so $\{X_1, X_2, \ldots\}$ crosses the interval $[a, b]$ only finite number of times a.s. It follows that

$$X_n \xrightarrow{n \to \infty} X_\infty := \limsup_{n \to \infty} X_n, \quad \text{a.s.}$$  

(12.14)
The convergence also occurs in $L^1$ since $\{X_n\}$ is UI.

Our remaining task is to show that $X_\infty$ is a version of $E(X|\mathcal{F}_\infty)$. First note that, thanks to the definition using \textit{limes superior} and the fact that $n \mapsto F_n$ is decreasing, $X_\infty$ is $\mathcal{F}_n$-measurable for each $n \geq 1$. Hence it is also $\mathcal{F}_\infty$-measurable. Next pick $A \in \mathcal{F}_\infty$. Since $A \in \mathcal{F}_n$, we have

$$E(X_1 A) = E(E(X|\mathcal{F}_n)1_A) \rightarrow E(X_\infty 1_A)$$ (12.15)

by $L^1$-convergence $X_n \rightarrow X_\infty$. Hence $X_\infty = E(X|\mathcal{F}_\infty)$.

As a bonus, we will give an elegant proof of the SLLN:

\textbf{Proof of the Strong Law of Large Numbers.} Suppose $\{X_n\}$ are i.i.d. with $X_1 \in L^1$. Realize these random variables on the product space $(\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), P)$, where $P$ is the product measure. Set $S_n := X_1 + \cdots + X_n$ and $\mathcal{F}_n := \sigma(S_n, S_{n+1}, \ldots)$. We claim that

$$E(X_1|\mathcal{F}_n) = \frac{S_n}{n}$$ (12.16)

This follows from the fact that

$$E(X_k|\mathcal{F}_n) = E(X_1|\mathcal{F}_n), \quad 1 \leq k \leq n,$$ (12.17)

by taking the sum over $k = 1, \ldots, n$. The identity (12.17) is in turn proved as follows: Consider the transformation $\pi_k : \mathbb{R}^N \rightarrow \mathbb{R}^N$ defined by

$$\pi_k((x_1, x_2, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots)) := (x_k, x_2, \ldots, x_{k-1}, x_1, x_{k+1}, \ldots).$$ (12.18)

Then $\pi_k$ is a measurable bijection that preserves $P$, due to the i.i.d. structure of the measure. (This is directly checked on product events and extended by the $\pi$-$\lambda$ Theorem. Note that independence alone is not enough.) Under this map we have $X_k \circ \pi_k = X_1$ and $X_1 \circ \pi_k = X_k$ but $S_n \circ \pi_k = S_n$ whenever $k \leq n$. Hence, $\pi_k^{-1}(A) = A$ is true for any $A \in \mathcal{F}_n$ and so

$$E(X_k 1_A) = E(X_1 1_A), \quad A \in \mathcal{F}_n, k \leq n.$$ (12.19)

From here we get (12.17) as desired.

With (12.16) in hand, we apply the Backward Levy Theorem (and the fact that $\mathcal{F}_n$ is non-increasing) to get

$$\frac{S_n}{n} \rightarrow E(X_1|\mathcal{F}_\infty).$$ (12.20)

But $Y := \limsup_{n \rightarrow \infty} \frac{S_n}{n}$ is tail measurable and thus, by Kolmogorov’s Zero-One Law, it is a.s. constant. Hence $E(X_1|\mathcal{F}_\infty)$ is equal to its expectation, $E(X_1|\mathcal{F}_\infty) = E(X_1)$ a.s. It follows that $\frac{S_n}{n} \rightarrow EX_1$ a.s. and in $L^1$. 

\qed
13. Exchangeability

Our next task is to explore further the property that was used in the martingale-based proof of the SLLN. For motivation, let us consider one more time the example of Polya’s Urn. We will encode a run of the urn using random variables $Z_1, Z_2, \ldots$, where

$$Z_k := 1_{\{k\text{-th sampled ball is red}\}} \quad (13.1)$$

Then $R_n = r + Z_1 + \cdots + Z_k$. Let us compute the probability of a given run $Z_1, \ldots, Z_n$. First we examine the special case

$$P(Z_1 = 1, \ldots, Z_k = 1, Z_{k+1} = 0, \ldots, Z_n = 0) = \frac{r}{r+g} \cdots \frac{r+k-1}{r+g+k-1} \frac{g}{r+g+k} \cdots \frac{g+n-k-1}{r+g+n-1} \quad (13.2)$$

However, looking at the structure of the expression we easily convince ourselves that the same expression will be arrived at for any sequence of $k$ ones and $n-k$ zeros. For all $a_1, \ldots, a_n \in \{0,1\}$ with $\sum_{i=1}^n a_i = k$ we will thus have

$$P((Z_1, \ldots, Z_n) = (a_1, \ldots, a_n)) = P((Z_1, \ldots, Z_n) = (1, \ldots, 1, 0, \ldots, 0)) \quad (13.3)$$

This means that the random variables $(Z_1, Z_2, \ldots)$ obey the conditions in the following definition:

**Definition 13.1** A finite permutation $\pi$ is a bijection $\pi: \mathbb{N} \to \mathbb{N}$ such that $\{i \in \mathbb{N}: \pi(i) \neq i\}$ is finite. We say that the random variables $(X_1, X_2, \ldots)$ are exchangeable if

$$(X_{\pi(1)}, X_{\pi(2)}, \ldots) \overset{\text{law}}{=} (X_1, X_2, \ldots) \quad (13.4)$$

holds for any finite permutation $\pi$.

A particular consequence of the zero-one nature of the variables is that the computation of probability of a given run of zero’s and one’s conditioned on the total sum reduces to purely combinatorial argument:

$$P((Z_1, \ldots, Z_n) = (a_1, \ldots, a_n) \mid \sum_{i=1}^n Z_i = k) = \binom{n}{k}^{-1} \quad \text{whenever } \sum_{i=1}^n a_i = k. \quad (13.5)$$

Reduction to the joint law of only a few of the $Z_i$’s is then easy as well: If $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and $a_1, \ldots, a_m \in \{0,1\}$ are such that $a := a_1 + \cdots + a_m \leq k$, then

$$P(Z_{i_1} = 1, \ldots, Z_{i_m} = a \mid \sum_{i=1}^n Z_i = k) = \frac{\binom{n-m}{k-a}}{\binom{n}{k}}. \quad (13.6)$$

Now let us consider the limit $n \to \infty$. Thanks to the fact that $R_n/(r+g+n-1)$ is a martingale with values in $[0,1]$, the Martingale Convergence Theorem ensures

$$\frac{R_n}{r+g+n-1} \xrightarrow{n \to \infty} U \in [0,1] \quad \text{a.s.} \quad (13.7)$$
But that means that also
\[
\frac{1}{n} \sum_{i=1}^{n} Z_i \xrightarrow{n \to \infty} U \quad \text{a.s.} \tag{13.8}
\]

Denoting \( S_n := Z_1 + \cdots + Z_n \) and setting \( \mathcal{F}_n := \sigma(S_n, S_{n+1}, \ldots) \) we thus have
\[
P(Z_{i_1} = a_1, \ldots, Z_{i_m} = a_m \mid \mathcal{F}_n) = \frac{\binom{n-m}{a}}{\binom{n}{S_n}} U^a (1 - U)^{m-a} \tag{13.9}
\]

But the Backward Martingale Theorem implies that the limit of the left-hand side is the conditional probability given \( \mathcal{F}_\infty := \bigcap_{n \geq 1} \mathcal{F}_n \) and so
\[
P(Z_{i_1} = a_1, \ldots, Z_{i_m} = a_m \mid \mathcal{F}_\infty) = U^a (1 - U)^{m-a}. \tag{13.10}
\]

In plain words, the random variable \((Z_1, Z_2, \ldots)\), conditioned on \( \mathcal{F}_\infty \), are Bernoulli\((U)\). As the only property that was needed in the proof was exchangeability, we have proved:

**Theorem 13.2** (de Finetti Theorem for Zero-One Valued Random Variables) Suppose \( Z_1, Z_2, \ldots \) are exchangeable with values in \([0, 1]\). Then there exist a random variable \( U \) taking values in \([0, 1]\) such that, conditional on \( U \), the random variables \( \{Z_i\} \) are Bernoulli\((U)\). In particular, if \( \mu \) denotes the distribution of \( U \) on \([0, 1]\), then
\[
P(Z_1 = a_1, \ldots, Z_k = a_k) = \int_{[0,1]} \mu(du) \prod_{i=1}^{k} u^{a_i} (1-u)^{1-a_i}. \tag{13.11}
\]

for any \( k \geq 1 \) and any \( a_1, \ldots, a_k \in \{0, 1\} \).

Having solved the special case of zero-one valued random variables, we can move over to the general case. For that we need a bit more formalism. Consider a family of random variables \((X_1, X_2, \ldots)\) realized as coordinate projections on the product space \((\mathbb{R}^N, \mathcal{B}(\mathbb{R}^N), P)\). We assume that these random variables are exchangeable. Let
\[
\Pi_n := \{ \pi \text{ finite permutation such that } \pi(i) = i \ \forall i > n \}. \tag{13.12}
\]

Now we can use these to define the sigma algebras
\[
\mathcal{E}_n := \{ A \in \mathcal{B}(\mathbb{R}^N) : \pi^{-1}(A) = A \ \forall \pi \in \Pi_n \} \tag{13.13}
\]

and, since \( \Pi_n \subset \Pi_{n+1} \) and thus \( \mathcal{E}_{n+1} \subset \mathcal{E}_n \), also
\[
\mathcal{E} := \bigcap_{n \geq 1} \mathcal{E}_n. \tag{13.14}
\]

We will refer to \( \mathcal{E} \) as the sigma algebra of exchangeable events.

**Theorem 13.3** (de Finetti) Let \( \{X_n\} \) be exchangeable. Then for any collection of Borel sets \( B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R}) \),
\[
E\left( \prod_{i=1}^{k} 1_{\{X_i \in B_i\}} \mid \mathcal{E} \right) = \prod_{i=1}^{k} E(1_{\{X_i \in B_i\}} \mid \mathcal{E}) \quad \text{a.s.} \tag{13.15}
\]

In particular, if \( B \mapsto \mu(\omega, B) \) denotes the conditional distribution of \( X_1 \) given \( \mathcal{E} \), then, conditional on \( \mathcal{E} \), \( \{X_n\} \) are i.i.d. with joint law \( \mu(\omega, \cdot) \).
Proof: Pick \(B_1, \ldots, B_k \in \mathcal{B}(\mathbb{R})\) and abbreviate \(A_i^j := \{X_i \in B_j\}\). Our aim is to show

\[
P(A_1 \cap \cdots \cap A_k | \mathcal{E}) = \prod_{i=1}^k P(A_i^j | \mathcal{E}),
\]

where we use the probability notation to denote the expectation of an indicator. (Incidentally, since the underlying space is standard Borel and \(\mathcal{E}\) is countably generated, the conditional probability in fact exists.) For this we first write

\[
E\left( \prod_{j=1}^k \left( \frac{1}{n} \sum_{i=1}^n 1_{A_i^j} \right) \right| \mathcal{E}) = \frac{1}{n^k} \sum_{1 \leq i_1, \ldots, i_k \leq n} P(A_{i_1} \cap \cdots \cap A_{i_k} | \mathcal{E})
\]

\[
= O\left( \frac{k^2}{n} \right) + \frac{1}{n^k} \sum_{\text{distinct}} P(A_{i_1} \cap \cdots \cap A_{i_k} | \mathcal{E})
\]

(13.17) Here we noted that the number of terms when two indices coincide is bounded by \(n^{k-1}{k \choose 2}\). The key point now is that exchangeability implies

\[
P(A_{i_1} \cap \cdots \cap A_{i_k} | \mathcal{E}) = P(A_1 \cap \cdots \cap A_k | \mathcal{E})
\]

(13.18) and so

\[
E\left( \prod_{j=1}^k \left( \frac{1}{n} \sum_{i=1}^n 1_{A_i^j} \right) \right| \mathcal{E}) = O\left( \frac{k^2}{n} \right) + \frac{1}{n^k} \frac{n!}{(n-k)!} P(A_1 \cap \cdots \cap A_k | \mathcal{E})
\]

(13.19) But \(\sum_{i=1}^n 1_{A_i^j}\) is measurable with respect to \(\mathcal{E}\) and \(P(A_i^j | \mathcal{E}) = P(A_i^j | \mathcal{E})\). Hence,

\[
\prod_{j=1}^k P(A_i^j | \mathcal{E}) = O\left( \frac{k^2}{n} \right) + \frac{1}{n^k} \frac{n!}{(n-k)!} P(A_1 \cap \cdots \cap A_k | \mathcal{E})
\]

(13.20) Taking \(n \to \infty\) for \(k\) fixed, we get (13.16). The second clause in the statement now follows by the properties of the conditional distribution. \(\square\)

Notice that the above proof would make \((X_1, \ldots, X_k | \mathcal{E})\) indistinguishable from i.i.d. as long as \(k = \Theta(\sqrt{n})\) as \(n \to \infty\). An interesting question is what happens when \((X_1, X_2, \ldots)\) are already i.i.d. For this case we get:

**Theorem 13.4 (Hewitt-Savage Zero-One Law)** If \(X_1, X_2, \ldots\) are i.i.d. then \(P\) is trivial on \(\mathcal{E}\) — which means that \(P(A) \in \{0, 1\}\) for all \(A \in \mathcal{E}\).

**Proof.** The proof parallels the proof of Kolmogorov’s Zero-One Law. Suppose \(A \in \mathcal{E}\). Then, given \(\varepsilon > 0\), there is \(n \geq 1\) and \(A_n \in \sigma(X_1, \ldots, X_n)\) such that \(\|1_A - 1_{A_n}\|_1 < \varepsilon\). Consider now the permutation \(\pi\) that swaps \((1, \ldots, n)\) for \((n+1, \ldots, 2n)\) and keeps all other indices intact. Then also \(\|1_A - 1_{A_n} \circ \pi\| < \varepsilon\). But \(A_n\) is independent of \(\pi(A_n)\) and so we get, taking \(\varepsilon \downarrow 0\), that \(A\) is independent of itself. Hence \(P(A) = P(A)^2\) implying \(P(A) \in \{0, 1\}\). \(\square\)
14. $L^2$-Martingales and Some Inequalities

We will wrap up the subject of martingales by discussing martingales in $L^2$ and some useful inequalities. An $L^2$-martingale is a martingale $\{M_n, \mathcal{F}_n\}$ such that $M_n \in L^2$ for all $n$. Then we have:

**Lemma 14.1** Consider an $L^2$-martingale $\{M_n, \mathcal{F}_n\}$. Then $\{M_n - M_{n-1} : n \geq 0\}$ are uncorrelated and thus orthogonal in $L^2$-sense. (Here $M_{-1} := 0$.) In particular,

$$E(M_n^2) = E(M_0^2) + \sum_{k=1}^n E((M_k - M_{k-1})^2)$$  \hspace{1cm} (14.1)

**Proof.** The uncorrelatedness is a trivial consequence of square integrability and the fact that

$$E(M_{n+1} - M_n | \mathcal{F}_n) = 0$$  \hspace{1cm} (14.2)

as implied by the martingale property. The second property arises from writing $M_n = M_0 + \sum_{k=1}^n (M_k - M_{k-1})$, squaring this and noting that the expectation of all “cross” terms vanishes. \(\square\)

We can use this example to demonstrate the general representation of any process as a sum of a martingale and a predictable process. This is referred to as:

**Theorem 14.2** (Doob-Meyer decomposition) Let $\{\mathcal{F}_n\}$ be a filtration and $\{X_n\}$ an adapted process with $X_n \in L^1$ for every $n$. Then there is a martingale $\{M_n, \mathcal{F}_n\}$ and a predictable process $\{A_n\}$ such that $X_n = M_n + A_n$ for each $n$. Moreover, if $\{X_n, \mathcal{F}_n\}$ is a submartingale, resp., supermartingale then $\{A_n\}$ is increasing, resp., decreasing.

**Proof.** This is really a significant theorem for continuous-time processes; in the discrete-time setting this is just a rather trivial decomposition. Define $\{A_n\}$ by $A_0 := 0$ and, inductively,

$$A_{n+1} := A_n + E(X_{n+1} - X_n | \mathcal{F}_n)$$  \hspace{1cm} (14.3)

Obviously, this makes $\{A_n\}$ predictable. On the other hand, setting $M_n := X_n - A_n$ yields an adapted process with

$$E(M_{n+1} - M_n | \mathcal{F}_n) = E(X_{n+1} - X_n | \mathcal{F}_n) - (A_{n+1} - A_n) = 0$$  \hspace{1cm} (14.4)

and so $\{M_n, \mathcal{F}_n\}$ is a martingale. (The condition $M_n \in L^1$ is checked directly.) Finally, if $\{X_n\}$ is a submartingale, then $A_{n+1} \geq A_n$ (and similarly for $\{X_n\}$ being a supermartingale.) \(\square\)

To see the connection to $L^2$ martingales, note that if $\{M_n, \mathcal{F}_n\}$ is a martingale in $L^2$, then (by Jensen’s inequality) $\{M_n^2, \mathcal{F}_n\}$ is a submartingale. The Doob-Meyer decomposition advises us to define $\{A_n\}$ by

$$A_{n+1} - A_n := E(M_{n+1}^2 | \mathcal{F}_n) - M_n^2 = E((M_{n+1} - M_n)^2 | \mathcal{F}_n)$$  \hspace{1cm} (14.5)

to get that

$$M_n^2 - \sum_{k=1}^n E((M_k - M_{k-1})^2 | \mathcal{F}_{k-1})$$  \hspace{1cm} (14.6)

is a martingale. Taking expectation we then get (14.1). We then immediately get:
Lemma 14.3 Suppose that \( \{M_n, \mathcal{F}_n\} \) is an \( L^2 \)-martingale with \( \sum_{k \geq 1} E((M_k - M_{k-1})^2) < \infty \). Then \( M_n \rightarrow M_\infty \) almost surely and in \( L^2 \).

Proof. The inequality (14.1) shows that \( \sup_{n \geq 1} \|M_n\|_2 < \infty \). In particular, \( \{M_n\} \) is UI and so \( M_n \rightarrow M_\infty \) almost surely and in \( L^1 \). The identity

\[
E \left( (M_n - M_m)^2 \right) = \sum_{k=m+1}^{n} E \left( (M_k - M_{k-1})^2 \right) \tag{14.7}
\]

then shows that \( M_n \rightarrow M_\infty \) in \( L^2 \) as well. \( \square \)

For \( L^2 \)-martingales, we next note the following important inequality:

Theorem 14.4 (Doob’s \( L^2 \)-inequality) Let \( \{M_n, \mathcal{F}_n\} \) be an \( L^2 \)-martingale (or a non-negative \( L^2 \)-submartingale) with \( M_0 = 0 \). Then for each \( \lambda > 0 \),

\[
P \left( \max_{1 \leq k \leq n} M_k > \lambda \right) \leq \frac{1}{\lambda^2} E \left( M_n^2 1_{\{\max_{1 \leq k \leq n} M_k > \lambda\}} \right) \tag{14.8}
\]

Proof. The proof follows the same argument as Komogorov’s inequality for sums of independent zero-mean random variables. In fact, the inequality generalizes Komogorov’s inequality from such sums to martingales.

Consider the events

\[
A_k := \{M_k > \lambda > \max_{1 \leq \ell < k} M_\ell\}, \quad 1 \leq k \leq n. \tag{14.9}
\]

These are disjoint events and their union is exactly the event under consideration. Using that \( M_k/\lambda > 1 \) on \( A_k \), we then get

\[
P \left( \max_{1 \leq k \leq n} M_k > \lambda \right) = \sum_{k=1}^{n} P(A_k) \leq \frac{1}{\lambda^2} \sum_{k=1}^{n} E(M_k^2 1_{A_k}) \tag{14.10}
\]

Next we note that

\[
E(M_n^2 1_{A_k}) = E \left( (M_k + (M_n - M_k))^2 1_{A_k} \right) = E \left( M_k^2 1_{A_k} \right) + E \left( (M_n - M_k)^2 1_{A_k} \right) + 2E \left( M_k (M_n - M_k) 1_{A_k} \right) \tag{14.11}
\]

Using conditioning on \( \mathcal{F}_k \) and noting that both \( M_k \) and \( 1_{A_k} \) are \( \mathcal{F}_k \)-measurable while \( M_k E(M_n - M_k | \mathcal{F}_k) = 0 \) for \( \{M_n\} \) martingale (and \( \geq 0 \) for \( \{M_n\} \) a non-negative submartingale), we conclude

\[
E(M_n^2 1_{A_k}) \geq E(M_k^2 1_{A_k}), \quad k = 1, \ldots, n. \tag{14.12}
\]

Plugging this in (14.10) we thus get

\[
P \left( \max_{1 \leq k \leq n} M_k > \lambda \right) \leq \frac{1}{\lambda^2} \sum_{k=1}^{n} E(M_k^2 1_{A_k}) = E \left( M_n^2 1_{\bigcup_{k=1}^{n} A_k} \right) \tag{14.13}
\]

This is the statement we were to prove. \( \square \)

These inequalities generalize to \( L^p \)-martingales for all \( p \geq 1 \); for details see the homework assignment. The concentration estimate gets even better if the martingale has bounded increments, and is thus in \( L^\infty \):
Theorem 14.5 (Azuma-Hoeffding inequality) Let \( \{M_n, \mathcal{F}_n\} \) be a submartingale with \( M_0 = 0 \) such that, for some non-random sequence \( \{c_k\} \) of non-negative numbers,
\[
|M_k - M_{k-1}| \leq c_k, \quad k = 1, \ldots, n. \tag{14.14}
\]
Then
\[
P(M_n > \lambda) \leq \exp\left\{ -\frac{1}{2} \sum_{k=1}^{n} \frac{\lambda^2}{c_k^2} \right\}. \tag{14.15}
\]

Proof. The proof is based on the following inequality
\[
e^{ty} \leq \cosh(tc) + \frac{y}{c} \sinh(tc), \quad |y| \leq c, t \in \mathbb{R}. \tag{14.16}
\]
This follows from the fact that the right-hand side can be written as
\[
cosh(tc) + \frac{y}{c} \sinh(tc) = e^{tc + \frac{y}{2c}} + e^{-tc - \frac{y}{2c}}, \tag{14.17}
\]
Under the condition \(|y| \leq c\), both \( \frac{c+y}{2c} \) and \( \frac{c-y}{2c} \) are non-negative and they add up to one. So, by the convexity of the exponential, the right-hand side is at most the exponential of
\[
tc + \frac{y}{2c} - tc - \frac{y}{2c} = ty, \tag{14.18}
\]
and so we get (14.16).

To see how (14.16) implies the desired claim, consider the random variable \( e^{tM_n} \) for any \( t \geq 0 \). Since \( M_n \) is bounded, this is integrable and so
\[
E(e^{tM_n}) = E(e^{tM_{k-1}E(e^{t(M_k-M_{k-1})}|\mathcal{F}_k)}) \tag{14.19}
\]
But (14.16) ensures that
\[
E(e^{t(M_k-M_{k-1})}|\mathcal{F}_k) \leq \cosh(tc_k) + \frac{1}{c_k} \sinh(tc_k)E(M_k - M_{k-1}|\mathcal{F}_k) \tag{14.20}
\]
and this is equal to \( \cosh(tc_k) \) for martingales and \( \geq \cosh(tc_k) \) for submartingales since \( t \geq 0 \). Hence, using induction,
\[
E(e^{tM_k}) \leq \prod_{k=1}^{n} \cosh(tc_k), \quad t \geq 0. \tag{14.21}
\]
But \( x \mapsto \log \cosh(x) \) is symmetric, zero at \( x = 0 \) and with a decreasing third derivative and so \( \log \cosh(x) \leq \frac{1}{2} x^2 \). It follows that
\[
E(e^{tM_k}) \leq \exp\left\{ \frac{t^2}{2} \sum_{k=1}^{n} c_k^2 \right\}, \quad t \geq 0. \tag{14.22}
\]
Now use the exponential Chebyshev inequality to write
\[
P(M_n > \lambda) \leq e^{-\lambda t} E(e^{tM_k}), \quad t \geq 0. \tag{14.23}
\]
Plugging the above bound and minimizing over \( t \) shows that the optimal value to use is \( t := \lambda / \left( \sum_{k=1}^{n} \frac{c_k^2}{c_k^2} \right)^{-1} \). Doing so yields the desired inequality. \( \square \)
Numerous generalizations exists that permit relaxing the strict boundedness requirement. However, none is such that it would simply permit to drop the boundedness altogether and replace $\sum_{k=1}^n c_k^2$ by the expectation of $M_n^2$ — which is something one might naturally hope for there.

For submartingales, one gets the bound on the upper tail of $M_n$. For martingales, the symmetry yields two-sided control:

$$P\left(\left| M_n \right| > \lambda \right) \leq 2 \exp\left\{ -\frac{1}{2} \frac{\lambda^2}{\sum_{k=1}^n c_k^2} \right\}. \quad (14.24)$$

In particular, a typical value of $M_n$ will be at most order $\sqrt{\sum_{k=1}^n c_k^2}$. Bounds of this form belong to the area of *concentration of measure*. 
15. Markov Chains: Definitions and Examples

Along with the discussion of martingales, we have introduced the concept of a discrete-time stochastic process. In this chapter we will study a particular class of such stochastic processes called Markov chains. Informally, a Markov chain is a discrete-time stochastic process for which, given the present state, the future and the past are independent. The formal definition is as follows:

**Definition 15.1** Consider a probability space \((\Omega, \mathcal{F}, \mathbb{P})\) with a filtration \(\{\mathcal{F}_n\}_{n \geq 0}\) and a standard Borel space \((S, \mathcal{B}(S))\). Let \((X_n)_{n \geq 0}\) be an \(S\)-valued stochastic process adapted to \(\{\mathcal{F}_n\}\). We call this process a Markov chain with state space \(S\) if for every \(B \in \mathcal{B}(S)\) and every \(n \geq 0\),

\[
\mathbb{P}(X_{n+1} \in B | \mathcal{F}_n) = \mathbb{P}(X_{n+1} \in B | X_n), \quad \mathbb{P}\text{-a.s.} \quad (15.1)
\]

Here and henceforth, \(\mathbb{P}(\cdot | X_n)\) abbreviates \(\mathbb{P}(\cdot | \sigma(X_n))\). (The conditional distributions exist by Theorem 7.1.)

The relation (15.1) expresses the fact that, in order to know the future, we only need to know the present state. In practical situations, in order to prove that (15.1) is true, we typically calculate \(\mathbb{P}(X_{n+1} \in B | \mathcal{F}_n)\) explicitly and show that it depends only on \(X_n\). Since \(\sigma(X_n) \subseteq \mathcal{F}_n\), eq. (15.1) then follows by the “smaller always wins” principle for conditional expectations.

If we turn the above definition around, we can view (15.1) as a prescription that defines the Markov chain. Indeed, suppose we know the distribution of \(X_0\). Then (15.1) allows us to calculate the joint distribution of \((X_0, X_1)\). Similarly, if we know the distribution of \((X_0, \ldots, X_n)\) the above lets us calculate \((X_0, \ldots, X_{n+1})\). The object on the right-hand side of (15.1) then falls into the following category:

**Definition 15.2** A function \(p : S \times \mathcal{B}(S) \to [0, 1]\) is called a transition probability if

1. \(B \mapsto p(x, B)\) is a probability measure on \(\mathcal{B}(S)\) for all \(x \in S\).
2. \(x \mapsto p(x, B)\) is \(\mathcal{B}(S)\)-measurable for all \(B \in \mathcal{B}(S)\).

Our first item of concern is the existence of a Markov chain with given transition probabilities and initial distribution:

**Theorem 15.3** Let \((p_n)_{n \geq 0}\) be a sequence of transition probabilities and let \(\mu\) be a probability measure on \(\mathcal{B}(S)\). Then there exists a unique probability measure \(\mathbb{P}^\mu\) on the measurable space \((S^{\mathbb{N}_0}, \mathcal{B}(S^{\mathbb{N}_0}))\), where \(\mathbb{N}_0 = \mathbb{N} \cup \{0\}\), such that \(\omega \mapsto X_n(\omega) = \omega_n\) is a Markov chain with respect to the filtration \(\mathcal{F}_n = \sigma(X_0, \ldots, X_n)\) and such that for all \(B \in \mathcal{B}(S)\),

\[
P^\mu(X_0 \in B) = \mu(B) \quad (15.2)
\]

and

\[
P^\mu(X_{n+1} \in B | X_n) = p_n(X_n, B), \quad \mathbb{P}^\mu\text{-a.s.} \quad (15.3)
\]

for all \(n \geq 0\). In other words, \((X_n)\) defined by the coordinate maps on \((S^{\mathbb{N}_0}, \mathcal{B}(S^{\mathbb{N}_0}), \mathbb{P}^\mu)\) is a Markov chain with transition probabilities \((p_n)\) and initial distribution \(\mu\).

**Proof.** This result is a direct consequence of the Kolmogorov Extension Theorem. Indeed, recall that \(\mathcal{F}\) is the least \(\mathcal{B}(S^{\mathbb{N}_0})\)-algebra containing all sets \(A_0 \times \cdots \times A_n \times S \times \ldots\) where \(A_j \in \mathcal{B}(S)\).
We will define a consistent family of probability measures on the measurable space \((S^n, \mathcal{B}(S^n))\) by putting
\[ P_{(n)}^\mu(A_0 \times \cdots \times A_n) := \int_{A_0} \mu(dx_0) \int_{A_1} p_0(x_0, dx_1) \cdots \int_{A_n} p_{n-1}(x_{n-1}, dx_n), \]
and extending this to \(\mathcal{B}(S^n)\). It is easy to see that these measures are consistent because if \(A \in \mathcal{B}(S^n)\) then
\[ P_{(n+1)}^\mu(A \times S) = P_{(n)}^\mu(A). \]
By Kolmogorov’s Extension Theorem, there exists a unique probability measure \(P^\mu\) on the infinite product space \((S^{\infty}, \mathcal{B}(S^{\infty}))\) such that \(P^\mu(A \times S \times \ldots) = P_{(n)}^\mu(A)\) for every \(n \geq 0\) and every set \(A \in \mathcal{B}(S^n)\).

It remains to show that the coordinate maps \(X_n(\omega) := \omega_n\) define a Markov chain such that (15.2–15.3) hold true. The proof of (15.2) is easy:
\[ P^\mu(X_0 \in B) = P^\mu(B \times S \times \ldots) = P_{(0)}^\mu(B) = \mu(B). \]
In order to prove (15.3), we claim that for all \(B \in \mathcal{B}(S)\) and all \(A \in \mathcal{F}_n\),
\[ \mathbb{E}(1_{\{X_{n+1} \in B\}} 1_A) = \mathbb{E}(1_{X_n \in B}) 1_A. \]
To this end we first note that, by interpreting both sides as probability measures on \(\mathcal{B}(S^{\infty})\), it suffices to prove this just for \(A \in \mathcal{F}_n\) of the form \(A = A_0 \times \cdots \times A_n \times S \times \ldots\). But for such \(A\) we have
\[ \mathbb{E}(1_{\{X_{n+1} \in B\}} 1_A) = P_{(n+1)}^\mu(A_0 \times \cdots \times A_n \times B) \]
which by inspection of (15.4) — indeed, it suffices to integrate the last coordinate — simply equals \(\mathbb{E}(1_A p_n(X_n, B))\) as desired. — Once (15.7) is proved, it remains to note that \(p_n(X_n, B)\) is \(\mathcal{F}_n\)-measurable. Hence \(p_n(X_n, B)\) is a version of \(P^\mu(X_{n+1} \in B | \mathcal{F}_n)\). But \(\sigma(X_n) \subset \mathcal{F}_n\) and since \(p_n(X_n, B)\) is \(\sigma(X_n)\)-measurable, the “smaller always wins” principle implies that (15.1) holds. Thus \((X_n)\) is a Markov chain satisfying (15.2–15.3) as we desired to prove.

While the existence argument was carried out in a “continuum” setup, the remainder of this chapter will be specialized to the case when the \(X_n\)’s can take only a countable set of values with positive probability. Denoting the (countable) state space by \(S\), the transition probabilities \(p_n(x, dy)\) will become functions \(p_n: S \times S \to [0, 1]\) with the property
\[ \sum_{y \in S} p_n(x, y) = 1 \]
for all \(x \in S\). (Clearly, \(p_n(x, y)\) is an abbreviation of \(p_n(x, \{y\})\).) We will call such \(p_n\) a stochastic matrix. Similarly, the initial distribution \(\mu\) will become a function \(\mu: S \to [0, 1]\) with the property
\[ \sum_{x \in S} \mu(x) = 1. \]
determined by two objects: the initial distribution $\mu$ and the transition matrix $p(x,y)$, satisfying (15.10) and (15.9) respectively. For the rest of this chapter we will focus on time-homogeneous Markov chains.

Let us make a remark on general notation. As used before, the object $P^\mu$ denotes the Markov chain with initial distribution $\mu$. If $\mu$ is the point mass at state $x \in S$, then we denote the resulting measure by $P^x$. We now proceed by a list of examples of countable-state time-homogeneous Markov chains.

**Example 15.4 (SRW starting at $x$)** Let $S = \mathbb{Z}^d$ and, using $x \sim y$ to denote that $x$ and $y$ are nearest neighbors on $\mathbb{Z}^d$, let

$$p(x,y) = \begin{cases} \frac{1}{2d}, & \text{if } x \sim y, \\ 0, & \text{otherwise.} \end{cases}$$

(15.11)

Consider the measure $P^x$ generated from the initial distribution $\mu(y) = \delta_x(y)$ using the above transition matrix. As is easy to check, the resulting Markov chain is the simple random walk started at $x$.

**Example 15.5 (Galton-Watson branching process)** Consider i.i.d. random variables $(\xi_n)_{n \geq 1}$ taking values in $\mathbb{N} \cup \{0\}$ and define the stochastic matrix $p(n,m)$ by

$$p(n,m) = \mathbb{P}(\xi_1 + \cdots + \xi_n = m), \quad n,m \geq 0.$$  

(15.12)

(Here we use the interpretation $p(0,m) = 0$ unless $m = 0$.) Let $S = \mathbb{N} \cup \{0\}$ and let $P^1$ be the corresponding Markov chain started at $x = 1$. As is easy to verify, the result is exactly the Galton-Watson branching process discussed in the context of martingales.

**Example 15.6 (Ehrenfest chain)** In his study of convergence to equilibrium in thermodynamics, Ehrenfest introduced the following simple model: Consider two boxes with altogether $m$ labeled balls. At each time step we pick one ball at random and move it to the other box.

To formulate the problem more precisely, we will only keep track of the number of balls in one of the boxes. Thus, the set of possible values — i.e., the state space — is simply $S = \{0,1,\ldots,m\}$. To calculate the transition probabilities, we note that the number of balls always changes only by one. The resulting probabilities are thus

$$p(n,n-1) = \frac{n}{m} \quad \text{and} \quad p(n,n+1) = \frac{m-n}{m}.$$  

(15.13)

Note that this automatically gives zero probability to the situation when there is no balls left or where all balls are in one box. The initial distribution can be whatever is of interest.

**Example 15.7 (Birth-death chain)** A generalization of Ehrenfest chain is the situation where we think of a population evolving in discrete time steps. At each time only one of three things can happen: Either an individual is born or dies or nothing happens. The state space of such chain will be $S = \mathbb{N} \cup \{0\}$. The transition probabilities will then be determined by three sequences $(\alpha_x)$, $(\beta_x)$ and $(\gamma_x)$ via

$$p(x,x+1) = \alpha_x, \quad p(x,x-1) = \beta_x, \quad p(x,x) = \gamma_x.$$  

(15.14)

Clearly, we need that $\alpha_x + \beta_x + \gamma_x = 1$ for all $z \in S$ and, since the number of individuals is always non-negative, we also require that $\beta_0 = 0$. 


Example 15.8 (Stochastic matrix) An abstract example of a Markov chain arises whenever we are given a square stochastic matrix. For instance,

\[
p = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{5} & \frac{4}{5} \end{pmatrix}
\]

(15.15)
defines a Markov chain on \(S = \{0, 1\}\) with the matrix elements corresponding to the values of \(p(x, y)\).

Example 15.9 (Random walk on a graph) Consider a graph \(G = (V, E)\), where \(V\) is the countable set — the vertex set — and \(E \subseteq V \times V\) is a binary relation — the edge set of \(G\). We will assume that \(G\) is unoriented, i.e., \(E\) is symmetric, and that there are no self-loops, i.e., \(E\) is antireflexive. Let \(d(x)\) denote the degree of vertex \(x\) which is simply the number of \(y \in V\) with \((x, y) \in E\) — i.e., the number of neighbors of \(x\). Suppose that there are no isolated vertices, which means that \(d(x) > 0\) for all \(x\).

We define a random walk on \(V\) as a Markov chain on \(S = V\) with transition probabilities

\[
p(x, y) = \frac{a_{xy}}{d(x)},
\]

(15.16)
where \((a_{xy})_{x, y \in V}\) is the adjacency matrix,

\[
a_{xy} = \begin{cases} 1, & \text{if } (x, y) \in E, \\ 0, & \text{otherwise}. \end{cases}
\]

(15.17)
Since \(\sum_{y \in V} a_{xy} = d(x)\) we see \(p(x, y)\) is a stochastic matrix. (Here we assumed without saying that \(G\) is locally finite, i.e., \(d(x) < \infty\) for all \(x \in V\).)

Note that this extends the “usual” simple random walk on \(\mathbb{Z}^d\) — which we defined in terms of sums of i.i.d. random variables — to any locally finite graph. The initial distribution is typically concentrated at one point; namely, the starting point of the walk, see Example 15.4.

Having amassed a bunch of examples, we begin investigating some general properties of countable-state time-homogeneous Markov chains. (To keep the statements of theorems and lemmas concise, we will not state these as our assumptions any more.)
16. STATIONARY AND REVERSIBLE MEASURES

The first question we will try to explore is that of stationarity. To motivate the forthcoming definitions, consider a Markov chain $(X_n)_{n \geq 0}$ with transition probability $p$ and $X_0$ distributed according to $\mu_0$. As is easy to check, the law of $X_1$ is then described by measure $\mu_1$ which is computed by

$$\mu_1(y) = \sum_{x \in S} \mu_0(x)p(x,y). \quad (16.1)$$

We are interested in the situations when the distribution of $X_1$ is the same as that of $X_0$. This leads us to this, somewhat more general, definition:

**Definition 16.1** A (positive) measure $\nu$ on the state space $S$ is called stationary if

$$\nu(y) = \sum_{x \in S} \nu(x)p(x,y), \quad y \in S. \quad (16.2)$$

If $\nu$ has total mass one we call it stationary distribution.

**Remark 16.2** While we allow ourselves to consider measures $\mu$ on $S$ that are not normalized, we will always assume that the measure assigns finite mass to every element of $S$.

Clearly, once the laws of $X_0$ and $X_1$ are the same, then all $X_n$'s have the same law (provided the chain is time-homogeneous). Let us find stationary measures for the Ehrenfest and birth-death chains:

**Lemma 16.3** Consider the Ehrenfest chain with state space $S = \{0, 1, \ldots, m\}$ and let the transition matrix be as in (15.13). Then

$$\nu(k) = \binom{m}{k} \frac{m}{m}^{m-k} \frac{1}{m}^k, \quad k = 0, \ldots, m, \quad (16.3)$$

is a stationary distribution.

**Proof.** We have to show that $\nu$ satisfies (16.2). First we note that

$$\sum_{k \in S} \nu(k)p(k,l) = \nu(l-1)p(l-1,l) + \nu(l+1)p(l+1,l). \quad (16.4)$$

Then we calculate

$$\text{rhs of (16.4)} = 2^{-m}\left[ \binom{m}{l-1} \frac{m-l+1}{m} + \binom{m}{l+1} \frac{l+1}{m} \right]$$

$$= 2^{-m} \binom{m}{l} \left[ \frac{l}{m-l+1} \frac{m-l+1}{m} + \frac{m-l+1}{m} \frac{l+1}{m} \right]. \quad (16.5)$$

The proof is finished by noting that, after a cancellation, the bracket is simply one. □

Concerning the birth-death chain, we state the following:
Lemma 16.4  Consider the birth-death chain on $\mathbb{N} \cup \{0\}$ characterized by sequences $(\alpha_n)$, $(\beta_n)$ and $(\gamma_n)$, cf (15.14). Suppose that $\beta_n > 0$ for all $n \geq 1$. Then $\nu(0) = 1$ and
\[
\nu(n) = \prod_{k=1}^{n} \frac{\alpha_{k-1}}{\beta_k}, \quad n \geq 1,
\] defines a stationary measure of the chain.

In order to prove this lemma, we will introduce the following interesting concept:

Definition 16.5  A measure $\nu$ of a Markov chain on $S$ is called reversible if
\[
\nu(x)p(x,y) = \nu(y)p(y,x)
\]
holds for all $x, y \in S$.

As we will show later, reversibility owes its name to the fact that, if we run the chain backwards (whatever this means for now) starting from $\nu$, we would get the same Markov chain. A simple consequence of reversibility is stationarity:

Lemma 16.6  A reversible measure is automatically stationary.

Proof. We just have to sum both sides of (16.7) over $y$. Since $p(x,y)$ is stochastic, the left hand side produces $\nu(x)$ while the right-hand side gives $\sum_{y \in S} \nu(y)p(y,x)$.

Equipped with these observations, the proof of Lemma 16.4 is a piece of cake:

Proof of Lemma 16.4. We claim that $\nu$ in (16.6) is reversible. To prove this we observe that (16.6) implies for all $k \geq 0$
\[
\nu(k+1) = \nu(k) \frac{\alpha_k}{\beta_{k+1}} = \nu(k) \frac{p(k,k+1)}{p(k+1,k)}.
\] (16.8)
This shows that (16.7) holds for $x$ and $y$ differing by one. For $x = y$ (16.7) holds trivially and in the remaining cases $p(x,y) = 0$, so (16.7) is proved in general. Hence $\nu$ is reversible and thus stationary.

Remark 16.7  Recall that stationary measures need not be finite and thus the existence of a stationary measure does not imply the existence of a stationary distribution. (The distinction is really whether finite or infinite because a finite stationary measure can always be normalized.) For the Ehrenfest chain we immediately produced a stationary distribution. However, for the birth-death chain the question whether $\nu$ is finite or infinite depends sensitively on the asymptotic properties of the ratio $\alpha_{k+1}/\beta_k$. We will return to this later in this chapter.

Next we will address the underlying meaning of the reversible measure. We will do this by showing that reversing the “flow of time” we obtain another Markov chain, which in the reversible situation will be the same as the original chain.

Theorem 16.8 (Reversed chain)  Consider a Markov chain $(X_n)_{n \geq 0}$ started from a stationary initial distribution $\mu$ and transition matrix $p(x,y)$. Fix $N$ large and let
\[
Y_n = X_{N-n}, \quad n = 1, \ldots, N.
\] (16.9)
Then \((Y_n)_{n=0}^N\) is a (time-homogeneous) Markov chain — called the reversed chain — with initial distribution \(\mu\) and transition matrix \(q(x,y)\) defined by

\[
q(x,y) = \frac{\mu(y)}{\mu(x)} p(y,x), \quad \mu(x) > 0,
\]

(16.10)

(The values \(q(x,y)\) for \(x\) such that \(\mu(x) = 0\) are immaterial since such \(x\) will never be visited starting from the initial distribution \(\mu\).)

**Proof.** Fix a collection of values \(y_1, \ldots, y_N \in S\) and consider the probability \(\mathbb{P}(Y_n = y_n, n = 0, \ldots, N)\), where \(Y_n\) are defined from \(X_n\) as in (16.9). Since \((X_n)\) is a Markov chain, we have

\[
\mathbb{P}(Y_n = y_n, n = 0, \ldots, N) = \mu(y_N) p(y_N, y_{N-1}) \cdots p(y_1, y_0).
\]

A simple argument now shows that either this probability vanishes or \(\mu(y_k) > 0\) for all \(k = 0, \ldots, N\). Focussing on the latter cases we can rewrite this as follows:

\[
\mathbb{P}(Y_n = y_n, n = 0, \ldots, N) = \frac{\mu(y_N)}{\mu(y_{N-1})} p(y_N, y_{N-1}) \cdots \frac{\mu(y_1)}{\mu(y_0)} p(y_1, y_0) \mu(y_0)
\]

\[
= \frac{\mu(y_0)}{\mu(y_{N-1})} q(y_0, y_1) \cdots q(y_{N-1}, y_N).
\]

Hence \((Y_n)\) is a Markov chain with initial distribution \(\mu\) and transition matrix \(q(x,y)\). \(\square\)

Clearly, if \(\mu\) is a reversible measure then \(q(x,y) = p(x,y)\), i.e., the dual chain \((Y_n)\) has the same law as \((X_n)\). This allows us to extend any stationary Markov chain to negative infinity — into a two-sided chain \((X_n)_{n \in \mathbb{Z}}\). As an additional exercise we will apply these concepts to the random walk on a locally finite graph.

**Lemma 16.9** Consider a locally finite unoriented graph \(G = (V,E)\) and let \(d(x)\) denote the degree of vertex \(x\). Suppose that there are no isolated vertices, i.e., \(d(x) > 0\) for every \(x\). Then the measure

\[
v(x) = d(x), \quad x \in V,
\]

(16.13)

is reversible and hence stationary for the random walk on \(G\).

**Proof.** We have \(p(x,y) = a_{xy}/d(x)\), where \(a_{xy}\) is the adjacency matrix. Since \(G\) is unoriented, the adjacency matrix is symmetric. This allows us to calculate

\[
v(x)p(x,y) = d(x) \frac{a_{xy}}{d(x)} = a_{xy} = d(y) \frac{a_{yx}}{d(y)} = v(y)p(y,x).
\]

(16.14)

Thus \(v\) is reversible and hence stationary. \(\square\)

Clearly, \(v\) is finite if and only if \(E\) is finite which by the fact that no vertex is isolated implies that \(V\) is finite. However, the stationary measure may not be unique. Indeed, if \(G\) has two separate components, even the restriction of \(v\) to one of the component would be stationary. We proceed by analyzing the question of uniqueness (and existence) of stationary measures.
17. Existence/Uniqueness of Stationary Measures

As alluded to in the example of the random walk on a general graph, a simple obstruction to uniqueness is when there are parts of the state space $S$ for which the transition from one to another happens with zero probability. We introduce a proper name for this situation:

**Definition 17.1** We call the transition matrix $p$ (or the Markov chain itself) irreducible if for each $x, y \in S$ there is a number $n = n(x, y)$ such that

$$p^n(x, y) = \sum_{y_1, \ldots, y_{n-1} \in S} p(x, y_1)p(y_1, y_2)\ldots p(y_{n-1}, y) > 0. \quad (17.1)$$

The object $p^n(x, y)$ — not to be confused with $p_n(x, y)$ — is the $(x, y)$-th entry of the $n$-th power of the transition matrix $p$. As is easy to check, $p^n(x, y)$ simply equal the probability that the Markov chain started at $x$ is at $y$ at time $n$, i.e., $P^x(X_n = y) = p^n(x, y)$.

Irreducibility can be characterized in terms of a stopping time.

**Lemma 17.2** Consider a Markov chain $(X_n)_{n \geq 0}$ on $S$ and let $T_y = \inf\{n \geq 1: X_n = y\}$ be the first time the chain visits $y$ (note that we do not count the initial state $X_0$ in this definition). Then the chain is irreducible if and only if for all $x, y,$

$$P^x(T_y < \infty) > 0. \quad (17.2)$$

**Proof.** This is a trivial consequence of the fact that $P^x(X_n = y) = p^n(x, y)$ and that $P^x(T < \infty) = \sum_{n \geq 1} p^n(x, y)$.

However, irreducibility alone is not sufficient to guarantee the existence and uniqueness of a stationary measure. The principal concept here is *recurrence*, which we have already encountered in the context of random walks:

**Definition 17.3** A state $x \in S$ is called recurrent if $P^x(T_x < \infty) = 1$. A Markov chain is recurrent if every state is recurrent.

**Theorem 17.4** (Existence of stationary measure) Consider an irreducible Markov chain $(X_n)_{n \geq 0}$ and let $x \in S$ be a recurrent state. Then

$$\nu_x(y) = \mathbb{E}_x \left( \sum_{n=1}^{T_x} 1_{\{X_n = y\}} \right) = \sum_{n \geq 1} P^x(X_n = y; T_x \geq n) \quad (17.3)$$

is finite for all $y \in S$ and defines a stationary measure on $S$. Moreover, any other stationary measure is a multiple of $\nu_x$.

The crux of the proof and the principal reason why we need recurrence is because of the following observation: If $T_x < \infty$ almost surely, we can also write $\nu_x$ as follows:

$$\nu_x(y) = \mathbb{E}_x \left( \sum_{n=0}^{T_x-1} 1_{\{X_n = y\}} \right) = \sum_{n \geq 1} P^x(X_n = y; T_x \geq n). \quad (17.4)$$

The first equality comes from the fact that if $y \neq x$ then $X_n \neq y$ for $n = 0$ and $n = T_x$ anyway, while if $y = x$ then the sum in the first expectation in (17.3) and (17.4) equals one in both cases. The second equality in (17.4) follows by a convenient relabeling.
Proof of existence. Let \( x \in S \) be a recurrent state. Then

\[
P^x(X_n = y, T_x \geq n) = \sum_{z \in S} P^x(X_{n-1} = z, X_n = y, T_x \geq n)
\]

\[
= \sum_{z \in S} \mathbb{E}_x \left( P^x(X_n = y) | \mathcal{F}_{n-1} \right) \mathbb{1}_{\{X_{n-1} = z\}} \mathbb{1}_{\{T \geq n\}}
\]

\[
= \sum_{z \in S} p(z,y) P^x(X_{n-1} = z, T_x \geq n),
\]

(17.5)

where we used that \( \{X_{n-1} = z\} \) and \( \{T_x \geq n\} \) are both \( \mathcal{F}_{n-1} \)-measurable to derive the second line. The third line is a consequence of the fact that on \( \{X_{n-1} = z\} \) we have that \( P^x(X_n = y | \mathcal{F}_{n-1}) = P^y(X_n = y | X_{n-1}) = p(z,y) \).

Summing the above over \( n \geq 1 \), applying discrete Fubini (everything is positive) and invoking (17.3) on the left and (17.4) on the right-hand side gives us \( v_x(y) = \sum_{z \in S} v_x(z)p(z,y) \). It remains to show that \( v_x(y) < \infty \) for all \( y \in S \). First note that \( v_x(x) = 1 \) by definition. Next we note that we actually have

\[
v_x(x) = \sum_{z \in S} v_x(z)p^n(z,x) \geq v_x(y)p^n(y,x)
\]

(17.6)

for all \( n \geq 1 \) and all \( y \in S \). Thus, \( v_x(y) < \infty \) whenever \( p^n(y,x) > 0 \). By irreducibility this will happen for some \( n \) for every \( y \in S \) and so \( v_x(y) < \infty \) for all \( y \in S \). Hence \( v_x \) is a stationary measure.

The proof of uniqueness provides some motivation for how \( v_x \) was constructed:

Proof of uniqueness. Suppose \( x \) is a recurrent state and let \( v_x \) be the stationary measure in (17.3). Let \( \mu \) be another stationary measure (we require that \( \mu(y) < \infty \) for all \( y \in S \) even though, as we will see, it is enough to assume that \( \mu(x) < \infty \)). Stationarity of \( \mu \) can also be written as

\[
\mu(y) = \mu(x)p(x,y) + \sum_{z \neq x} \mu(z)p(z,y).
\]

(17.7)

Plugging this for \( \mu(z) \) in the second term and iterating gives us

\[
\mu(y) = \mu(x) \left[ p(x,y) + \sum_{z \neq x} p(x,z)p(z,y) + \cdots \sum_{z_1, \ldots, z_n \neq x} p(x,z_1) \cdots p(z_n,y) \right] + \sum_{z_0, \ldots, z_n \neq x} \mu(z_0)p(z_0, z_1) \cdots p(z_n,y).
\]

(17.8)

We would like to pass to the limit and conclude that the last term tends to zero. However, a direct proof of this appears unlikely and so we proceed by using inequalities. Noting that the \( k \)-th term in the bracket equals \( P^x(X_k = y, T_x \geq k) \), we have

\[
\mu(y) \geq \mu(x) \sum_{k=1}^{n+1} P^x(X_k = y, T_x \geq k) \xrightarrow{n \to \infty} \mu(x)v_x(y).
\]

(17.9)

In particular, we have \( \mu(y) \geq \mu(x)v_x(y) \) for all \( y \in S \).

Our goal is to show that equality holds. Suppose that for some \( x, y \) we have \( \mu(y) > \mu(x)v_x(y) \).

By irreducibility, there exists \( n \geq 1 \) such that \( p^n(y,x) > 0 \) and so

\[
\mu(x) = \sum_{z \in S} \mu(z)p^n(z,x) > \mu(x) \sum_{z \in S} v_x(z)p(z,x) = \mu(x)v_x(x) = \mu(x),
\]

(17.10)
a contradiction. So \( \mu(y) = \mu(x)\nu_x(y) \), i.e., \( \mu \) is a rescaled version of \( \nu_x \).

We finish by noting that there are criteria for existence of stationary measures for transient chains (proved by Harris (Proceedings AMS 1957) and Veech (Proceedings AMS 1963). In particular, are irreducible transient chains without stationary measures and also those with stationary measures. (For the latter, note that the random walk on any locally-finite graph has a reversible, and thus stationary, measure.)
18. Strong Markov Property and Density of Returns

In the proof of existence and uniqueness of the stationary measure we have barely touched upon the recurrence property. Before we delve deeper into that subject let us state and prove an interesting consequence of the definition of Markov chain.

We will suppose that our Markov chain is defined on its canonical measurable space \((\Omega, \mathcal{F})\), where \(\Omega := S^\mathbb{N}\) and \(\mathcal{F}\) is the product \(\sigma\)-algebra. The \(X_n\)'s are represented by the coordinate maps \(X_n(\omega) := \omega_n\). This setting permits the consideration of the shift operator \(\theta\) which acts on sequences \(\omega\) by

\[
(\theta \omega)_n = \omega_{n+1}.
\]

For any \(n \geq 1\) we define \(\theta^n\) to be the \(n\)-fold composition of \(\theta\), i.e., \((\theta^n \omega)_k = \omega_{k+n}\). If \(N\) is a stopping time of the filtration \(\mathcal{F}_n = \sigma(X_0, \ldots, X_n)\), then \(\theta^N\) is defined to be \(\theta^n\) on \(\{N = n\}\). On \(\{N = \infty\}\) we leave \(\theta^N\) undefined.

**Theorem 18.1** (Strong Markov property) Consider a Markov chain \((X_n)_{n \geq 0}\) with initial distribution \(\mu\) and let \(N\) be a stopping time of \(\{\mathcal{F}_n\}\). Suppose that \(P^\mu(N < \infty) > 0\) and let \(\theta^N\) be as defined above. Then for all \(B \in \mathcal{F}\),

\[
P^\mu(1_B \circ \theta^N | \mathcal{F}_N) = P^\mu_X(B) \quad P^\mu\text{-a.s. on } \{N < \infty\}.
\]

Here \(\mathcal{F}_N = \{A \in \mathcal{F} : A \cap \{N = n\} \in \mathcal{F}_n, n \geq 0\}\) and \(X_N\) is defined to be \(X_n\) on \(\{N = n\}\).

This property is called “strong” because it is a strengthening of the Markov property

\[
P^\mu(1_B \circ \theta^N | \mathcal{F}_n) = P^\mu_X(B)
\]

to random \(n\). In our case the proof of the Markov property and the strong Markov property amount more or less to the same. In particular, no additional assumptions are needed. This is not true for continuous-time where strong Markov property may fail in the absence of (rather natural) continuity conditions.

**Proof of Theorem 18.1.** Let \(A \in \mathcal{F}_N\) be such that \(A \subset \{N < \infty\}\). First we will partition according to the values of \(N\):

\[
\mathbb{E}_\mu \left(1_A(1_B \circ \theta^N)\right) = \sum_{n \geq 0} \mathbb{E}_\mu \left(1_A \cap \{N = n\} (1_B \circ \theta^n)\right).
\]

(18.4)

(Note that, by our assumptions about \(A\), we do not have to include the value \(N = \infty\). Once we are on \(\{N = n\}\) we can replace \(\theta^N\) by \(\theta^n\).) Now \(A \cap \{N = n\} \in \mathcal{F}_n\) while \(1_B \circ \theta^n\) is \(\sigma(X_n, X_{n+1}, \ldots)\)-measurable. This allows us to condition on \(\mathcal{F}_n\) and use (15.3):

\[
\mathbb{E}_\mu \left(1_A \cap \{N = n\} (1_B \circ \theta^n)\right) = \mathbb{E}_\mu \left(1_A \cap \{N = n\} \mathbb{E}_\mu(1_B \circ \theta^n | \mathcal{F}_n)\right) = \mathbb{E}_\mu(1_A \cap \{N = n\} P^\mu_X(B)) = \mathbb{E}_\mu\left(1_A \cap \{N = n\} P^\mu_X(B)\right).
\]

Plugging this back to (18.4) and summing over \(n\) we deduce

\[
\mathbb{E}_\mu \left(1_A(1_B \circ \theta^N)\right) = \mathbb{E}_\mu \left(1_A P^\mu_X(B)\right)
\]

for all \(A \in \mathcal{F}_N\) with \(A \subset \{N < \infty\}\). Since \(P^\mu_X(B)\) is \(\mathcal{F}_N\) measurable, a standard argument implies that \(P^\mu(1_B \circ \theta^N | \mathcal{F}_N)\) equals \(P^\mu_X(B)\) almost surely on \(\{N < \infty\}\).  

We proceed by listing some applications of the Strong Markov property. Consider the stopping time $T_x$ defined by

$$T_x = \inf\{n \geq 1 : X_n = x\}. \quad (18.7)$$

Here we deliberately omit $n = 0$ from the sum, so that even in $P_x$ we may have $T_x = \infty$ almost surely. Let $T^n_x$ denote the $n$-th iteration of $T_x$ by which we mean $T_x \circ \theta^{T_{n-1}}_x$. One of the principal conclusions of the strong Markov property is the following observation:

**Lemma 18.2** Consider a Markov chain with state space $S$, let $x, y \in S$ and let $T_y$ be as defined above. Consider arbitrary events $A_0, \ldots, A_{n-1} \in \mathcal{F}_{T_y}$ with $A_j \subset \{T_y < \infty\}$ and let $A_n \in \mathcal{F}$. Then for the Markov chain started at $x$, the events

$$\theta^{T_j}_y(A_j), \quad j = 0, \ldots, n, \quad (18.8)$$

are independent and $P_x(\theta^{T_j}_y(A_j)) = P_y(A_j)$ for all $j = 1, \ldots, n$. (Here $T_0^y = 0$.)

**Proof.** Suppose $n = 1$. By $A_0 \subset \{T_y < \infty\}$, the strong Markov property, and the fact that $X_{T_y} = y$ on $\{T_y < \infty\}$,

$$P_x(A_0 \cap \theta^{T_y}_y(A_1)) = \mathbb{E}_x(1_{A_0}P^{T_y}_y(A_1)) = P_x(A_0)P_y(A_1). \quad (18.9)$$

The general case follows by induction. \hfill \Box

Another interesting conclusion is:

**Corollary 18.3** Consider a Markov chain with state space $S$ and let $T_x$ be as defined above. Then for all $x, y \in S$,

$$P_x(T^n_y < \infty) = P_x(T_y < \infty)P_y(T_y < \infty)^{n-1}. \quad (18.10)$$

**Proof.** Let $A_j = \{T_j < \infty\}$ in Lemma 18.2 and apply $\{T^n_y < \infty\} = \bigcap_{j \leq n-1} \theta^{T_j}_y(A_j)$. \hfill \Box

As for the random walk this statement allows us to characterize recurrence of $x$ in terms of the expected number of visits of the chain back to $x$.

**Corollary 18.4** Let $N(x) := \sum_{n \geq 1} 1(X_n = x)$. Then

$$\mathbb{E}_x(N(y)) = \frac{P_x(T_y < \infty)}{1 - P_y(T_y < \infty)}. \quad (18.11)$$

Here the right-hand side should be interpreted as zero if the numerator vanishes and as infinity if the numerator is positive and the denominator vanishes.

**Proof.** This is a consequence of the fact that $\mathbb{E}_x(N(y)) = \sum_{n \geq 1} P_x(T^n_y < \infty)$ and the formula (18.10). \hfill \Box

**Corollary 18.5** A state $x$ is recurrent if and only if $\mathbb{E}_x(N(x)) = \infty$. In particular, for an irreducible Markov chain either all states are recurrent or none of them are. Finally, an irreducible finite-state Markov chain is recurrent.

**Proof.** A state $x$ is recurrent iff $P_x(T_x < \infty) = 1$ which by (18.11) is true iff $\mathbb{E}_x(N(x)) = \infty$. To show the second claim, suppose that $x$ is recurrent and let us show that so is any $y \in S$. To that
end, let \( k \) and \( l \) be numbers such that \( p^k(y,x) > 0 \) and \( p^l(x,y) > 0 \) — these numbers exist by irreducibility. Then

\[
p^{n+k+l}(y,y) \geq p^k(y,x)p^n(x,x)p^l(x,y),
\]

which implies

\[
\mathbb{E}_y(N(y)) = \sum_{m \geq 1} p^m(y,y) \geq \sum_{n \geq 1} p^k(y,x)p^n(x,x)p^l(x,y) = p^k(y,x)p^l(x,y)\mathbb{E}_x(N(x)).
\]

But \( p^k(y,x)p^l(x,y) > 0 \) and so \( \mathbb{E}_x(N(x)) = \infty \) implies \( \mathbb{E}_y(N(y)) = \infty \). Hence, all states of an irreducible Markov chain are recurrent if one of them are. Finally, if \( S \) is finite, the trivial relation \( \sum_{x \in S} N(x) = \infty \) implies \( \mathbb{E}_z(N(x)) = \infty \) for all \( x, z \in S \). Then (18.11) yields \( P^x(T_x < \infty) = 1 \), i.e., every state is recurrent.

In the previous section we concluded that, for irreducible Markov chains, recurrence is a class property, i.e., a property that either holds for all states or none. We have also shown that, once the chain is recurrent (on top of being irreducible), there exists a stationary measure. Next we will give conditions under which the stationary measure has finite mass which means it can be normalized to produce a stationary distribution. To that end we introduce the following definitions:

**Definition 18.6** A state \( x \in S \) of a Markov chain is said to be
- transient if \( P^x(T_x < \infty) < 1 \).
- null recurrent if \( P^x(T_x < \infty) = 1 \) but \( \mathbb{E}_x T_x = \infty \).
- positive recurrent if \( \mathbb{E}_x T_x < \infty \).

We will justify the terminology later. Our goal is to show that a stationary distribution exists if and only if every state of an (irreducible) chain is positive recurrent. The principal result is formulated as follows:

**Theorem 18.7** Consider a Markov chain with state space \( S \). If there exists a stationary measure \( \mu \) with \( 0 < \mu(S) < \infty \) then every \( x \) with \( \mu(x) > 0 \) is recurrent. If the chain is irreducible, then

\[
\mu(x) = \frac{\mu(S)}{\mathbb{E}_x T_x}
\]

for all \( x \in S \). In particular, \( \mathbb{E}_x T_x < \infty \) for all \( x \), i.e., every state is positive recurrent.

**Proof.** Let \( x \) be such that \( \mu(x) > 0 \). Since \( \mu \) is stationary, \( \mu(x) = \sum_{z \in S} \mu(z)p^n(z,x) \) for all \( n \geq 1 \). Therefore

\[
\infty = \sum_{n \geq 1} \mu(x) = \sum_{z \in S} \mu(z) \sum_{n \geq 1} p^n(z,x) = \sum_{z \in S} \mu(z)\mathbb{E}_z(N(x)).
\]

But (18.11) implies \( \mathbb{E}_z(N(x)) \leq [1 - P^x(T_x < \infty)]^{-1} \) and so

\[
\infty \leq \frac{\mu(S)}{1 - P^x(T_x < \infty)}.
\]

Since \( \mu(S) < \infty \), we must have \( P^x(T_x < \infty) = 1 \), i.e., \( x \) is recurrent.
In order to prove the second part of the claim, note that irreducibility implies that \( \mu(x) > 0 \) for all \( x \) and so all states are recurrent. From Theorem 17.4 we have \( \mu(y) = \mu(x)v_y(y) \) which by \( \sum_y v_y(y) = v_x(S) = \mathbb{E}_xT_x \) yields

\[
\mu(S) = \mu(x) \sum_{y \in S} v_y(y) = \mu(x)\mathbb{E}_xT_x, \tag{18.17}
\]

implying (18.14). Since \( \mu(S) < \infty \) we must have \( \mathbb{E}_xT_x < \infty \). □

We summarize the interesting part of this result as a corollary:

**Corollary 18.8** For an irreducible Markov chain on a state space \( S \), the following are equivalent:

1. Some state is positive recurrent.
2. There exists a stationary measure \( \mu \) with \( \mu(S) < \infty \).
3. Every state is positive recurrent.

**Proof.** (1)⇒(2): Let \( x \) be positive recurrent. Then \( v_x \) is a stationary measure with \( v_x(S) = \mathbb{E}_xT_x < \infty \). (2)⇒(3): This is the content of Theorem 18.7. (3)⇒(1): Trivial. □

We finish this section by providing a justification for the terminology of “positive and null recurrent” states/Markov chains (both are class properties):

**Theorem 18.9** (Density of returns) Consider a Markov chain with state space \( S \). Let \( N_n(y) = \sum_{m=1}^n 1\{X_m = y\} \) be the number of visits to \( y \) before time \( n \). If \( y \) is recurrent, then for all \( x \in S \),

\[
\lim_{n \to \infty} N_n(y) = \frac{1}{\mathbb{E}_yT_y}1\{T_y < \infty\}, \quad \mathbb{P}^x\text{-a.s.} \tag{18.18}
\]

**Proof.** Let us first consider the case \( x = y \). Then recurrence implies \( 1\{T_y < \infty\} = 1 \) almost surely. Define the sequence of times \( \tau_n = T_y^n - T_y^{n-1} \) where \( \tau_1 = T_y \). By the Strong Markov Property, \( (\tau_n) \) are i.i.d. with the same distribution as \( T_y \). In terms of the \( \tau_n \)'s, we have

\[
N_n(y) = \sup\{k \geq 0: \tau_1 + \cdots + \tau_k \leq n\}, \quad (18.19)
\]

i.e., \( N_n(y) \) is a renewal sequence. The Renewal Theorem then gives us

\[
\lim_{n \to \infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}_y\tau_1} = \frac{1}{\mathbb{E}_yT_y}, \quad \mathbb{P}^x\text{-a.s.} \tag{18.20}
\]

Now we will look at the cases \( x \neq y \). If \( P^x(T_y = \infty) = 1 \) then \( N_n(y) = 0 \) almost surely for all \( n \) and there is nothing to prove. We can thus assume that \( P^x(T_y < \infty) > 0 \) and decompose according to the values of \( T_y \). We will use the Markov property which tells us that for any \( A \in \mathcal{F} \), we have \( P^x(\theta^n(A)|T_y = m) = P^y(A) \). We will apply this to the event

\[
A = \left\{ \lim_{n \to \infty} \frac{N_n(y)}{n} = \frac{1}{\mathbb{E}_yT_y} \right\}. \tag{18.21}
\]
Indeed, this event occurs almost surely in $P^y$ and so we have $P^x(\theta^m(A) | T_y = m) = 1$. But on $\theta^m(A) \cap \{T_y = m\}$ we have

$$\lim_{n \to \infty} \frac{N_{n+m}(y) - N_m(y)}{n} = \frac{1}{E_y T_y}$$

which implies that $N_n(y)/n \to 1/E_y T_y$. Therefore, $P^x(A | T_y = m) = 1$ for all $m$ with $P^x(T_y = m) > 0$. It follows that $A$ occurs almost surely in $P^x(\cdot | T_y < \infty)$. Hence, the limit in (18.18) equals $1/E_y T_y$ almost surely on $\{T_y < \infty\}$ and zero almost surely on $\{T_y = \infty\}$. This proves the claim. □

The upshot of this theorem is that a state is positive recurrent if it is visited at a positive density of times and null recurrent if it is visited infinitely often, but the density of visits is zero. Notwithstanding, in all cases all $N(y)$’s grow at roughly the same rate:

**Theorem 18.10** (Ratio limit theorem)  **Consider an irreducible Markov chain and let $x$ be a recurrent state. Let $\nu_x$ be the measure from (17.3). Then for all $x, y, z \in S$,

$$\lim_{n \to \infty} \frac{N_n(y)}{N_n(z)} = \frac{\nu_x(y)}{\nu_x(z)}, \quad P^x\text{-almost surely.}$$

The proof of this claim is part of the homework assignment for this part.
19. Convergence to Equilibrium

Markov chains are often run on a computer in order to sample from a complicated distribution on a large state space. The idea is to define a Markov chain for which the desired distribution is stationary and then wait long enough for the chain to “equilibrate.” The last aspect of Markov chains we wish to examine is the convergence to equilibrium.

When run on a computer, only one state of the Markov chain is stored at each time — this is why Markov chains are relatively easy to implement — and so we are asking about the convergence of the distribution \( P^\mu(X_n \in \cdot) \). Noting that this is captured by the quantities \( p_n(x, \cdot) \), we will thus study the convergence of \( p_n(x, \cdot) \) as \( n \to \infty \).

For irreducible Markov chains, we can generally guarantee convergence is in Cesaro sense:

**Lemma 19.1** Consider an irreducible Markov chain on \( S \). Then for all \( x, y \in S \), the limit

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} p_m(x, y) = \mu(y), \tag{19.1}
\]

exists and defines a stationary measure (possibly \( \mu \equiv 0 \)).

**Proof.** Let \( N_n(y) = \sum_{1 \leq k \leq n} 1_{\{X_k = y\}} \) and note that

\[
\sum_{m=1}^{n} p_m(x, y) = \mathbb{E}_x(N_n(y)). \tag{19.2}
\]

By Theorem 18.9 we know that \( P^x \)-almost surely, \( N_n(y)/n \) converges to the reciprocal value of \( \mathbb{E}_y T_y \) in recurrent cases. In transient cases this ratio converges to zero by direct observation; the a.s. limit thus always exist and vanishes unless the chain is positive recurrent. Since \( N_n(y)/n \leq 1 \), the Bounded Convergence Theorem applies the same is true even under expectation.

Suppose now that the chain is positive recurrent, fix \( x \) and let \( \varphi_n(x, y) = \frac{1}{n} \mathbb{E}_x(N_n(y)) \). Then the Markov property tells us

\[
\varphi_{n+1}(x, y) = \frac{p(x, y)}{n+1} + \frac{n}{n+1} \sum_{z \in S} \varphi_n(x, z) p(z, y). \tag{19.3}
\]

By our previous reasoning, \( \varphi_n(x, y) \to \varphi(y) := (\mathbb{E}_y T_y)^{-1} \) as \( n \to \infty \). Reducing the sum over \( z \) to a finite subset of \( S \), passing to the limit, and then increasing the size of this subset to comprise all of \( S \) we derive the inequality

\[
\varphi(y) \geq \sum_{z \in S} \varphi(z) p(z, y), \quad y \in S. \tag{19.4}
\]

But summing over \( y \in S \) we get equal terms on both sides so if this inequality were strict for at least one \( y \), we would have a contradiction. Hence \( \varphi \) is a stationary measure. \( \square \)

**Remark 19.2** Note that the previous proof yields the following surprising, and quite unintuitive, identity:

\[
\frac{1}{\mathbb{E}_x T_x} = \sum_{y \in S} \frac{1}{\mathbb{E}_y T_y} p(y, x), \tag{19.5}
\]

for all irreducible Markov chains (which is only interesting in the positive-recurrent case).
Proceeding similarly, we conclude that $p^n(x,y)$ themselves may not converge. For instance, if we have a chain that “hops” between two states, $p^n(x,y)$ will oscillate between zero and one as $n$ changes. The obstruction is clearly related to periodicity — if there was a slightest probability to not “hop,” the chain would soon get out of sync and the equilibrium would be reached.

In order to classify the periodic situation, let $I_\varepsilon = \{ n \geq 1 : p^n(x,x) > 0 \}$. By standard arguments, $I_\varepsilon$ is an additive semigroup (a set closed under addition). This allows us to define a number $d_\varepsilon$ as the largest integer that divides all $n \in I_\varepsilon$. (Since 1 divides all integers, such number indeed exists.) We call $d_\varepsilon$ the period of $x$.

**Lemma 19.3** If the Markov chain is irreducible, then $d_\varepsilon$ is the same for all $x$.

**Proof.** See (5.3) Lemma on page 309 of the textbook. 

**Definition 19.4** A Markov chain is called aperiodic if $d_\varepsilon = 1$ for all $x$.

**Lemma 19.5** An aperiodic Markov chain with state space $S$ satisfies the following: For all $x,y \in S$ there exists $n_0 = n_0(x,y)$ such that $p^n(x,y) > 0$ for all $n \geq n_0$.

**Proof.** First we note that, thanks to irreducibility, it suffices to prove this for $x = y$. Indeed, if $k \geq 1$ is such that $p^k(x,y) > 0$, then $p^{n+k}(x,y) \geq p^n(x,x)p^k(x,y)$ and so if $p^n(x,x) > 0$ for $n \geq n_0(x,x)$, then $p^n(x,y) > 0$ for $n \geq k + n_0(x,x)$.

Let us set $x = y$. By aperiodicity, there exists a $k \geq 1$ such that $p^k(x,x) > 0$ and $p^{k+1}(x,x) > 0$. Since $p^{2k+m+1}(x,x) \geq p^{k+1}(x,x)p^{k+m}(x,x)$, we thus have $p^{2k}(x,x), p^{2k+1}(x,x), p^{2k+2}(x,x) > 0$. Proceeding similarly, we conclude that

$$p^{2^2 + j}(x,x) > 0, \quad j = 0, \ldots, k-1$$

(19.6)

But that implies

$$p^{2k+mk+j}(x,x) \geq p^{2^2+j}(x,x)\left[p^k(x,x)\right]^m > 0, \quad m \geq 0, \quad j = 0, \ldots, k-1.$$ 

(19.7)

Every integer $n \geq k^2$ can be written in the form $n = k^2 + mk + j$ for such $m$ and $j$ and so the result holds with $n_0(x,x) \leq k^2$. \hfill \Box

Our goals it to prove the following result:

**Theorem 19.6** Consider an irreducible, aperiodic Markov chain on state space $S$. Suppose there exists a stationary distribution $\pi$. Then for all $x \in S$,

$$p^n(x,y) \xrightarrow{n \to \infty} \pi(y), \quad y \in S.$$ 

(19.8)

The proof of this theorem will be based on a general technique, called **coupling**. The idea is as follows: We will run one Markov chain started at $x$ and the other started at a $z$ which was itself chosen at random from distribution $\pi$. As long as the chains stay away from each other, we keep generating them independently. The first moment they collide, we glue them and from that time on move both of them synchronously.

The upshot is that, if we observe only the chain started at $x$, we see a chain started at $x$ while if we observe the chain started at $z$, we observe only a chain started at $z$. But the latter was started
from stationary distribution and so it will be stationary at each time. It follows that, provided the chains glued, also the one started from \( x \) will eventually be stationary.

To make this precise, we will have to define both chains on the same probability space. We will generalize the initial distributions to any two measures \( \mu \) and \( \nu \) on \( S \). Let us therefore consider a Markov chain on \( S \times S \) with transition probabilities

\[
p((x_1, x_2), (y_1, y_2)) = \begin{cases} 
p(x_1, y_1)p(x_2, y_2) & \text{if } x_1 \neq x_2, \\
p(x_1, y_1) & \text{if } x_1 = x_2 \text{ and } y_1 = y_2, \\
0, & \text{otherwise},
\end{cases}
\]  

(19.9)

and initial distribution \( \mu \otimes \nu \). We will use \( P^{\mu \otimes \nu} \) to denote the corresponding probability measure — called the coupling measure — and \((X_n^{(1)}, X_n^{(2)})\) to denote the coupled process.

First we will verify that each of the marginals is the original Markov chain:

\begin{lemma}
Let \((X_n^{(1)}, X_n^{(2)})\) denote the coupled process in measure \( P^{\mu \otimes \nu} \). Then \((X_n^{(1)})\) is the original Markov chain on \( S \) with initial distribution \( \mu \), while \((X_n^{(2)})\) is the original Markov chain on \( S \) with initial distribution \( \nu \).
\end{lemma}

\begin{proof}
Let \( A = \{X_k^{(1)} = x_k, k = 0, \ldots, n\} \). Abusing the notation slightly, we want to show that \( P^{\mu \otimes \nu}(A) = P^{\mu}(A) \). Since \( A \) fixes only the \( X_k^{(1)} \)'s, we can calculate the probability of \( A \) by summing over the possible values of \( X_k^{(2)} \):

\[
P^{\mu \otimes \nu}(A) = \sum_{(y_k)} \mu(x_0)\nu(y_0) \prod_{k=0}^{n-1} p((x_k, y_k), (x_{k+1}, y_{k+1})).
\]  

(19.10)

Next we note that

\[
\sum_{y' \in S} p((x, y), (x', y')) = \begin{cases} 
\sum_{y' \in S} p(x, x')p(y, y') & \text{if } x \neq y, \\
p(x, x') & \text{if } x = y.
\end{cases}
\]  

(19.11)

In both cases the sum equals \( p(x, x') \) which we note is independent of \( y \). Therefore, the sums in (19.10) can be performed one by one with the result

\[
P^{\mu \otimes \nu}(A) = \mu(x_0) \prod_{k=0}^{n-1} p(x_k, x_{k+1}),
\]  

(19.12)

which is exactly \( P^{\mu}(A) \). The second marginal is handled analogously.
\end{proof}

Our next item of interest is the time when the chains first collide:

\begin{lemma}
Let \( T = \inf\{n \geq 0 : X_n^{(1)} = X_n^{(2)}\} \). Under the conditions of Theorem 19.6,

\[
P^{\mu \otimes \nu}(T < \infty) = 1
\]  

(19.13)

for any pair of initial distributions \( \mu \) and \( \nu \).
\end{lemma}

\begin{proof}
We will consider an uncoupled chain on \( S \times S \) where both original Markov chains move independently forever. This chain has the transition probability

\[
g((x_1, y_1), (x_2, y_2)) = p(x_1, y_1)p(x_2, y_2).
\]  

(19.14)

\end{proof}
As a moment’s though reveals, the time \( T \) has the same distribution in both coupled and uncoupled chains. Therefore, we just need to prove the lemma for the uncoupled chain.

First, let us note that the uncoupled chain is irreducible (this is where aperiodicity is needed). Indeed, by Lemma 19.5 aperiodicity implies that \( p^n(x_1, y_1) > 0 \) and \( p^n(x_2, y_2) > 0 \) for \( n \) sufficiently large and so we also have \( q^n((x_1, y_1), (x_2, y_2)) > 0 \) for \( n \) sufficiently large. Second, we observe that the uncoupled chain is recurrent. Indeed, \( \tilde{\pi}(x, y) = \pi(x)\pi(y) \) is a stationary distribution and, using irreducibility, every state of the chain is thus recurrent. But then, for any \( x \in S \), the first hitting time of \( (x, x) \) is finite almost surely which implies the same for \( T \), which is the first hitting time of the diagonal in \( S \times S \).

The principal idea behind coupling now reduces to the following lemma:

**Lemma 19.9 (Coupling inequality)** Consider the coupled Markov chain with initial distribution \( \mu \otimes \nu \) and let \( T = \inf\{n \geq 0: X_n^{(1)} = X_n^{(2)}\} \). Let \( \mu_n(\cdot) = P^{\mu \otimes \nu}(X_n^{(1)} \in \cdot) \) and \( \nu_n(\cdot) = P^{\mu \otimes \nu}(X_n^{(2)} \in \cdot) \) be the marginals at time \( n \). Then

\[
\|\mu_n - \nu_n\| \leq P^{\mu \otimes \nu}(T > n),
\]

where \( \|\mu_n - \nu_n\| = \sup_{A \subseteq S} |\mu_n(A) - \nu_n(A)| \) is the variational distance of \( \mu_n \) and \( \nu_n \).

**Proof.** Let \( S^+ = \{x \in S: \mu_n(x) > \nu_n(x)\} \). The proof is based on the fact that
\[
\|\mu_n - \nu_n\| = \mu_n(S^+) - \nu_n(S^+).
\]

This makes it reasonable to evaluate the difference
\[
\mu_n(S^+) - \nu_n(S^+) = P^{\mu \otimes \nu}(X_n^{(1)} \in S^+) - P^{\mu \otimes \nu}(X_n^{(2)} \in S^+)
\]
\[
= \mathbb{E}_{\mu \otimes \nu}(1_{X_n^{(1)} \in S^+} - 1_{X_n^{(2)} \in S^+})
\]
\[
= \mathbb{E}_{\mu \otimes \nu}(1_{T > n}(1_{X_n^{(1)} \in S^+} - 1_{X_n^{(2)} \in S^+})).
\]

Here we have noted that if \( T \leq n \) then either both \( \{X_n^{(1)} \in S^+\} \) and \( \{X_n^{(2)} \in S^+\} \) occur or both don’t. Estimating the difference of the two indicators by one, we thus get \( \mu_n(S^+) - \nu_n(S^+) \leq P^{\mu \otimes \nu}(T > n) \). Plugging this into (19.16), the desired estimate follows.

Now we are ready to prove the convergence to equilibrium:

**Proof of Theorem 19.6.** Consider two Markov chains, one started from \( \mu \) and the other from \( \nu \). By Lemmas 19.7 and 19.9, the variational distance between the distributions \( \mu_n \) and \( \nu_n \) of \( X_n \) in these two chains is bounded by \( P^{\mu \otimes \nu}(T > n) \). But Lemma 19.8 implies that \( P^{\mu \otimes \nu}(T > n) \) tends to zero as \( n \to \infty \) and \( \|\mu_n - \nu_n\| \to 0 \).

To get (19.8) we now let \( \mu = \delta \) and \( \nu = \pi \). Then \( \mu_n(\cdot) = p^n(\cdot, \cdot) \) while \( \nu_n = \pi \) for all \( n \). Hence we have \( \|p^n(\cdot, \cdot) - \pi\| \to 0 \) which means that \( p^n(\cdot, \cdot) \to \pi \) in the variational norm. This implies (19.8).

The method of proof is quite general and can be adapted to other circumstances. See Lindvall’s book “Lectures on the coupling method.” We observe that Lemmas 19.9 and 19.7 allow us to estimate the time it takes for the two marginals to get closer than prescribed. On the basis of the
proof of Lemma 19.7, the coupling time can be studied in terms of the uncoupled process, which is slightly easier to handle.

The above technique also provides some estimate how fast the chain converges to its equilibrium. Indeed, the coupling time provides a bound on the mixing time

$$\tau_{\text{mix}}(\varepsilon) = \sup_{x \in S} \inf \{ n \geq 0 : \| p^n(x, \cdot) - \nu \| \leq \varepsilon \}.$$ \hfill (19.18)

Much of the research devoted to Markov chain these days is devoted to deriving sharp bounds on the mixing time in various specific examples. Sophisticated techniques based on coupling and spectral analysis have been developed; unfortunately, we don’t have time to cover these in this course. Instead, we will discuss in detail a simple, yet fairly representative, Markov chain.

**Example 19.10** (Random walk on hypercube) Consider a $d$-dimensional $2 \times \cdots \times 2$ hypercube which we represent at $S = \{-1, 1\}^d$. The vertices of the hypercube are $d$-tuples of $\pm 1$’s, $x = (x_1, \ldots, x_d)$. We will consider a random walk on $S$ which, formally, is a Markov chain with transition probabilities

$$p(x, y) = \begin{cases} \frac{1}{2d}, & \text{if } (\exists i)(x_i = -y_i \& (\forall j \neq i)x_j = y_j), \\ \frac{1}{2}, & \text{if } x = y, \\ 0, & \text{otherwise}. \end{cases} \hfill (19.19)$$

The reason for the middle line is to make the chain aperiodic (which is a necessary condition for convergence $p^n(x, \cdot) \to \nu$).

To analyze the behavior of this Markov chain, we will consider a natural coupling between two copies of this chain. Explicitly, consider a Markov chain on $S \times S$ with the following random “move”: Given a state $(x, y)$ of this chain, pick $i = 1, \ldots, d$ at random and, if $x_i = y_i$, perform the above single-chain move in both coordinates simultaneously while if $x_i \neq y_i$, move them independently. To see how these chains gradually couple, we introduce the Hamming distance on $S$ which is defined by

$$d(x, y) = |\{i : 1 \leq i \leq d, x_i \neq y_i\}|.$$ \hfill (19.20)

We also introduce the stopping times

$$T_j = \inf \{ n \geq 0 : d(X_n, Y_n) \leq d(X_0, Y_0) - j \}.$$ \hfill (19.21)

**Lemma 19.11** Consider the about duplicated Markov chain whose component are each started from a fixed point, let $d_0 = d(X_0, Y_0)$ be the Hamming distance of the initial states and let $\tau_j = T_{j+1} - T_j, \ j = 0, \ldots, d_0$. Then the random times $(\tau_j)$ are independent and $\tau_j$ has geometric distribution with parameter $\frac{d_0 - j}{2d^2}$.

**Proof.** The independence is a consequence of the strong Markov property. To find the distribution of $\tau_j$, it clearly suffices to focus on the case $j = 0$. Now the Hamming distance will decrease by one only if the following occurs: An index $i$ is chosen among one of the components where the two states differ — this happens with probability $\frac{d_0 - j}{d}$ — and one of the chain moves while the other does not — which happens with probability $\frac{1}{2}$. Thus, the Hamming distance will decrease with probability $\frac{d_0 - j}{2d}$ in each step; it thus takes a geometric time with this parameter for the move to actually occur. \hfill $\square$
Next we recall the Coupon collector problem: A coupon collector collects coupons that arrive one per unit time. There are \( r \) distinct flavors of the coupons; each coupon is sampled independently at random from the set of \( r \) flavors. The goal is to find when the collector has seen coupons of all \( r \) flavors for the first time. If \( Z_n \in \{1, \ldots, r\} \) marks the flavor of the \( n \)-th coupon that arrived, we define
\[
\hat{T}^{(r)}_j = \inf\{ n \geq 0 : |Z_1, \ldots, Z_n| \geq j \}
\] (19.22)
to be the time when the collector has for the first time seen \( j \) distinct flavors of coupons. It is easily checked that \((\hat{T}^{(r)}_{j+1} - \hat{T}^{(r)}_j)\) are independent with \( \hat{T}^{(r)}_{j+1} - \hat{T}^{(r)}_j \) geometric with parameter \( \frac{j}{r} \).

The WLLN-type analysis also gives that
\[
\frac{\hat{T}^{(r)}_d}{r \log r} \to 1, \quad \text{in probability.} \tag{19.23}
\]

Going back to our Markov chain, the random times \((\tau_j)_{j \leq d_0}\) have the same law as \((\hat{T}^{(2d)}_{d_0+j} - \hat{T}^{(2d)}_{d_0})_{j \leq d_0}\). In particular \(T_{d_0}\) has the same law as \(\hat{T}^{(2d)}_{2d_0} - \hat{T}^{(2d)}_{d_0}\). But \(\hat{T}^{(2d)}_{d_0}/d\) converges in probability to a finite constant and so we conclude that \(T_{d_0} \sim 2d \log(2d)\). Since this provides an upper bound on the mixing time, we have
\[
\tau_{\text{mix}}(\epsilon) = O(d \log d). \tag{19.24}
\]

This actually turns out to be correct to within a multiplicative constant. Indeed, it is known that for \(\epsilon > 0\) and \(d \gg 1\), \(\sup_{x \in S} \|p^n_x(x, \cdot) - \nu\|\) sweeps very rapidly from nearly 1 to nearly 0 as \(n\) increases from \((\frac{1}{2} - \epsilon)d \log d\) to \((\frac{1}{2} + \epsilon)d \log d\). This feature, discovered by Persi Diaconis and referred to as cutoff phenomenon is one of the most exciting recent developments in the area of Markov chains.