# H

# Part A Probability: Non-examinable proofs

This appendix provides additional technical detail for each chapter of the Part A Probability notes. Since we have not added systematic references from within the chapters to maintain the flow there at the intended level of the course, we use the same structure here for ease of reference. Specifically, Section H.$j$ corresponds to Chapter $j$, and we mostly restate any relevant claims here. Subdivisions in Section H.3 are not aligned with subdivisions of Chapter 3, though. As with anything in the previous appendices that goes beyond the prerequisites of this course, the material of this appendix goes beyond the syllabus and is non-examinable.

## H.1 Review of Prelims Probability

We refer to Appendix D for a discussion of Prelims Probability, Appendix E for some further developments based on Prelims Analysis and Probability. As discussed in the final remark of Appendix E, this leaves three items, two of which are discussed in Appendix G using results from Part A Integration previewed in Appendix F. The final unproven detail is that the weak law of large numbers holds with only a first moment. As noted previously, this is proved in Part B Probability, Measure and Martingales using martingale methods, but is also included here by rigorously establishing and using generating function methods, which does not require martingales but uses the measure-theoretic approach to expectations of Section G.5.

## H.2 Convergence of random variables, limit theorems

*Remark* H.1. In the (review) proof of Markov's inequality in Section 2.4, we proceed slightly differently from the Prelims proof, but we still use the order properties of expectation without restricting the random variable to be either discrete or continuous or assuming the finiteness of the expectation. In the proof of the weak law of large numbers, we use the additivity of variance for independent random variables. This was all set up in Section E.2, specifically Theorems E.5 and Theorem E.13, as well as the discussion after Theorem E.8.

*Remark* H.2. The (non-examinable) proof of the strong law of large numbers in Section 2.5.1 is under the additional assumption that fourth moments are finite. For a general proof, we

refer to Part B Probability, Measure and Martingales. In our special case, we also use the general linearity of expectation of Theorem E.8, as well as Corollary E.14.

The appropriate pointer for interchanging series and expectation is to Part B Probability, Measure and Martingales. Recall the sketch provided in Sections G.4–G.5.

With this setup, Tonelli's theorem yields in the context of Section 2.5.1 that

$$\mathbb{E}\left[\sum_{n=1}^{\infty}\left(\frac{S_n}{n}-\mu\right)^4\right] = \int_{\omega\in\Omega}\int_{n\in\mathbb{N}}\left(\frac{S_n(\omega)}{n}-\mu\right)^4 \mu_1(dn)\mathbb{P}(d\omega)$$
$$= \int_{n\in\mathbb{N}}\int_{\omega\in\Omega}\left(\frac{S_n(\omega)}{n}-\mu\right)^4 \mathbb{P}(d\omega)\mu_1(dn) = \sum_{n=1}^{\infty}\mathbb{E}\left[\left(\frac{S_n}{n}-\mu\right)^4\right].$$

*Remark* H.3. The central limit theorem has been stated without proof in Chapter 2, indicating that proofs based on generating functions will be provided in Chapter 3.

## H.3 Generating functions

### H.3.1 The convergence theorem for probability generating functions

The convergence theorem for probability generating functions is stated without proof, but we do remark that this can be proved "with a bit more work" than the uniqueness theorem, which was proved in Prelims Probability, all based on the theory of power series.

**Theorem H.4.** *For $\mathbb{N}$-valued random variables $X$ and $X_n$, $n \geq 1$, with probability mass functions (pmfs) $p$ and $p_n$, $n \geq 1$, and probability generating functions (pgfs) $G$ and $G_n$, $n \geq 1$,*

$$\forall k \geq 0\ p_n(k) \to p(k) \quad \iff \quad \forall s \in [0,1]\ G_n(s) \to G(s).$$

*Proof.* "⇒": Let $s \in [0,1]$. Let $\varepsilon > 0$. Since $p\colon \mathbb{N} \to [0,1]$ is a pmf, there is $k_0$ such that

$$\sum_{k=0}^{k_0-1} p(k) \geq 1 - \frac{\varepsilon}{4}.$$

Since $p_n(k) \to p(k)$, there is $n_0$ such that

$$\forall n \geq n_0,\ \sum_{k=0}^{k_0-1} p_n(k) \geq 1 - \frac{\varepsilon}{2} \quad \text{and} \quad \forall n \geq n_0\ \forall 0 \leq k \leq k_0 - 1,\ |p_n(k) - p(k)| < \frac{\varepsilon}{4k_0}.$$

But then

$$|G_n(s) - G(s)| \leq \sum_{k=0}^{k_0-1} s^k |p_n(k) - p(k)| + \sum_{k=k_0}^{\infty} p_n(k) + \sum_{k=k_0}^{\infty} p(k) < k_0 \frac{\varepsilon}{4k_0} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon.$$

"⇐": Clearly the $s = 0$ case yields $p_n(0) \to p(0)$. But then

- either $p(0) = 1$ and $0 \leq G_n(s) - p_n(0) \leq \sum_{k=1}^{\infty} p_n(k) = 1 - p_n(0) \to 0$ implies that $p_n(k) \to 0$ for all $k \geq 1$,

- or $p(0) < 1$ and for all $n \geq n_0$, also $p_n(0) < 1$, and for all $s \in [0,1]$,

$$\sum_{k=0}^{\infty} \frac{p_n(k+1)}{1-p_n(0)} s^k = \frac{G_n(s) - p_n(0)}{1-p_n(0)} \to \frac{G(s)-p(0)}{1-p(0)} = \sum_{k=0}^{\infty} \frac{p(k+1)}{1-p(0)} s^k.$$

An induction shows that $p_n(k) \to p(k)$ for all $k \geq 0$. □

### H.3.2 Moment generating functions, characteristic functions and their uniqueness theorems

*Remark* H.5. The only reason why the proof of Theorem H.5 is an "informal proof" is that we interchange expectation and series. To formalise this, we recall the argument in Remark H.2, where a similar interchange is justified for non-negative random variables, using Tonelli's theorem. In the case here where the random variables can take positive and negative values, the product space of counting measure and probability measure is the same, but we need to establish integrability to apply Fubini's Theorem. To this end, we may compute the repeated integral in either order, by Tonelli's Theorem, and the finiteness of the mgf at $\pm t$ yields

$$\begin{aligned}\int_{\omega\in\Omega}\int_{k\in\mathbb{N}} \left|\frac{t^k}{k!}(X(\omega))^k\right| \mu_1(dk)\mathbb{P}(d\omega) &= \int_{\omega\in\Omega} e^{|tX(\omega)|}\mathbb{P}(d\omega) \\ &= \mathbb{E}[e^{|tX|}] \leq \mathbb{E}[e^{tX}+e^{-tX}] = \mathbb{E}[e^{tX}] + \mathbb{E}[e^{-tX}] < \infty.\end{aligned}$$

This establishes that the function $f(k,\omega) = (k!)^{-1}t^k(X(\omega))^k$ is integrable with respect to the product measure of $\mu_1$ and $\mathbb{P}$, and we may interchange to find, when $|t| \leq t_0$,

$$\begin{aligned}M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{k\in\mathbb{N}} \frac{t^k}{k!}X^k\right] &= \int_{\omega\in\Omega}\int_{k\in\mathbb{N}} f(k,\omega)\mu_1(dk)\mathbb{P}(d\omega) \\ = \int_{k\in\mathbb{N}}\int_{\omega\in\Omega} f(k,\omega)\mathbb{P}(d\omega)\mu_1(dk) &= \sum_{k\in\mathbb{N}} \mathbb{E}\left[\frac{t^k}{k!}X^k\right] = \sum_{k\in\mathbb{N}} \frac{t^k\mathbb{E}[X^k]}{k!}.\end{aligned}$$

Turning to moment generating functions and characteristic functions, it is instructive to combine them into an analytic transform, as follows.

**Definition H.6.** Let $X$ be a real-valued random variable. We define the *analytic transform*

$$M_X(w) = \mathbb{E}[e^{wX}] = \mathbb{E}[\cos(vX)e^{uX}] + i\mathbb{E}[\sin(vX)e^{uX}]$$

for all $w = u + iv \in \mathbb{C}$ for which these expectations exist. We denote the set of such $w$-values by $\mathcal{D}_X$.

This extends the moment generating function for $w = u \in \mathbb{R}$ and relates to the characteristic function for $w = iv \in i\mathbb{R}$ as $\phi_X(v) = M_X(iv)$.

**Lemma H.7.** *For any real-valued random variable, $\mathcal{D}_X = I + i\mathbb{R}$ for some interval $I$ that contains* 0.

*Proof.* The observation $0 \in \mathcal{D}_X$ is elementary. If $a, c \in \mathcal{D}_X \cap \mathbb{R}$ with $a \leq 0 \leq c$ and $a \leq b \leq c$, then $e^{bX} \leq e^{aX}\mathbf{1}_{\{X<0\}} + e^{cX}\mathbf{1}_{\{X\geq 0\}}$ and by order properties of expectation we conclude that $[a,c] \subset \mathcal{D}_X \cap \mathbb{R}$ and hence that $\mathcal{D}_X \cap \mathbb{R}$ is an interval $I$.

Similarly, $|\cos(vX)e^{uX}| \leq e^{vX}$ and $|\sin(vX)e^{uX}| \leq e^{vX}$, hence $I + i\mathbb{R} \subseteq \mathcal{D}_X$. Conversely, if $u + iv \in \mathcal{D}_X$, then $e^{uX} \leq (|\cos(vX)| + |\sin(vX)|)e^{uX}$, so $u \in \mathcal{D}_X$. □

**Theorem H.8.** *Consider a real-valued random variable $X$ such that $(-t_0, t_0) \subseteq \mathcal{D}_X \cap \mathbb{R}$ for some $t_0 > 0$. Then the function $M_X \colon (-t_0, t_0) + i\mathbb{R} \to \mathbb{C}$ is holomorphic.*

*Proof.* The argument in Remark H.5 applies with $t \in (-t_0, t_0)$ replaced by $z \in (-t_0, t_0) + i\mathbb{R}$. Hence, $M_X$ is a power series in $z$ and therefore holomorphic in the disk of radius $t_0$ around the origin. Similarly, for all $w \in (-t_0, t_0) + i\mathbb{R}$,

$$\mathbb{E}[e^{zX}] = \mathbb{E}\left[e^{wX} \sum_{k\in\mathbb{N}} \frac{(z-w)^k}{k!} X^k\right] = \sum_{k\in\mathbb{N}} \frac{\mathbb{E}[X^k e^{wX}]}{k!}(z-w)^k$$

has positive radius of convergence, so $M_X$ is holomorphic on $(-t_0, t_0) + i\mathbb{R}$. □

**Corollary H.9.** *The standard normal distribution has characteristic function $\phi(v) = e^{-v^2/2}$.*

*Proof.* For the moment generating function $M(u) = e^{u^2/2}$, $u \in \mathbb{R}$, we recognise the unique (by the identity theorem, Theorem F.47) holomorphic extension $M(w) = e^{w^2/2}$, $w \in \mathbb{C}$, that includes the characteristic function $\phi(v) = M(iv) = e^{(iv)^2/2} = e^{-v^2/2}$, $v \in \mathbb{R}$, as claimed. □

Let us now first consider characteristic functions. We will deduce the uniqueness theorem for characteristic functions from the following inversion formula.

**Theorem H.10** (Inversion formula for characteristic functions)**.** *Consider any real-valued random variable $X$ and its characteristic function $\phi_X$. Then for all $a, b \in \mathbb{R}$ with $a < b$, we have, as $m \to \infty$,*

$$\frac{1}{2\pi}\int_{-m}^{m} \frac{e^{-iat} - e^{-ibt}}{it}\phi_X(t)dt \to \frac{1}{2}\mathbb{P}(X = a) + \mathbb{P}(a < X < b) + \frac{1}{2}\mathbb{P}(X = b).$$

*Proof.* We write the integral on the left-hand side as a double integral against the product measure of $\mathbb{P}$ and Lebesgue measure on $[-m, m]$, in the sense of Section G.4, with integrand

$$f(\omega, t) = \frac{e^{-iat} - e^{-ibt}}{it} e^{itX(\omega)}, \quad (\omega, t) \in \Omega \times [-m, m].$$

Since this function is bounded measurable and the product measure is finite on $\Omega \times [-m, m]$, the function $f$ is integrable and Fubini's Theorem applies to give

$$\int_{[0,m]}\int_{\Omega} f(\omega, t)\mathbb{P}(d\omega)dt = \int_{\Omega}\int_{[0,m]} f(\omega, t)dt\mathbb{P}(d\omega) = \mathbb{E}\left[\int_{[0,m]} \frac{e^{it(X-a)} - e^{it(X-b)}}{it}dt\right]$$

and, after change of variables

$$\int_{[-m,0]}\int_{\Omega} f(\omega, t)\mathbb{P}(d\omega)dt = \int_{\Omega}\int_{[-m,0]} f(\omega, t)dt\mathbb{P}(d\omega) = \mathbb{E}\left[\int_{[0,m]} \frac{e^{-it(X-b)} - e^{-it(X-a)}}{it}dt\right].$$

By linearity of expectation and integration, they sum to

$$\begin{aligned}\int_{-m}^{m} \frac{e^{-iat} - e^{-ibt}}{it}\phi_X(t)dt &= 2\mathbb{E}\left[\int_0^m \frac{\sin(t(X-a)) - \sin(t(X-b))}{t}dt\right] \\ &= \mathbb{E}\left[2\int_0^{m(X-a)} \frac{\sin(u)}{u}du - 2\int_0^{m(X-b)} \frac{\sin(u)}{u}du\right],\end{aligned}$$

with the convention that $\int_0^{-r} g(u)du = -\int_{-r}^0 g(u)du$ for $r > 0$, and for even functions $g$ this further equals $-\int_0^r g(u)du$. Borrowing from Complex Analysis that $\int_0^r \frac{\sin(u)}{u} \to \pi/2$ as $r \to \infty$, we find that the term under the expectation converges to $2\pi$ on $\{a < X < b\}$, to $\pi$ on $\{X = a\}$ and $\{X = b\}$ and to 0 on $\{X < a\}$ and $\{X > b\}$. To apply the Dominated Convergence Theorem and let $m \to \infty$, we further check that the term under the expectation is bounded by $4\int_0^\pi (\sin(u)/u)du$, so that

$$\frac{1}{2\pi}\int_{-m}^{m} \frac{e^{-iat} - e^{-ibt}}{it}\phi_X(t)dt \to \frac{1}{2}\mathbb{P}(X = a) + \mathbb{P}(a < X < b) + \frac{1}{2}\mathbb{P}(X = b).$$

□

**Corollary H.11** (Uniqueness theorem for characteristic functions)**.** *If $X$ and $Y$ are random variables with the same characteristic function, then $X$ and $Y$ have the same distribution.*

*Proof.* By Theorem H.10, we have

$$\frac{1}{2}\mathbb{P}(X = a) + \mathbb{P}(a < X < b) + \frac{1}{2}\mathbb{P}(X = b) = \frac{1}{2}\mathbb{P}(Y = a) + \mathbb{P}(a < Y < b) + \frac{1}{2}\mathbb{P}(Y = b)$$

Letting $a \to -\infty$, and $b \downarrow x$, using Theorem D.46 and Lemma E.11, this yields equality of cumulative distribution functions, which completes the proof. □

As a consequence of the analyticity of Theorem H.8, we obtain a Complex Analysis proof of the uniqueness theorem for moment generating functions, which we restate here.

**Corollary H.12** (Uniqueness theorem for moment generating functions)**.** *If $X$ and $Y$ are random variables with the same moment generating function, which is finite on $[-t_0, t_0]$ for some $t_0 > 0$, then $X$ and $Y$ have the same distribution.*

*Proof.* The two holomorphic functions $M_X$ and $M_Y$ coincide on $(-t_0, t_0)$. By the identity theorem of Complex Analysis, Theorem F.47, $M_X$ and $M_Y$ coincide on the domain $(-t_0, t_0) + i\mathbb{R}$. In particular, $\phi_X = \phi_Y$, and by Corollary H.11, $X$ and $Y$ have the same distribution. □

### H.3.3 Convergence theorems for moment generating functions and characteristic functions

Turning to convergence theorems, let us first establish the Skorokhod representation theorem.

**Theorem H.13** (Skorokhod representation theorem)**.** *Let $Y$ and $X_1, X_2, \ldots$ be random variables such that $X_n$ converges to $Y$ in distribution. Then there exists a probability space and random variables $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots$ with the same distributions as $Y$ and $X_1, X_2, \ldots$, such that $\widetilde{X}_n$ converges to $\widetilde{Y}$ almost surely.*

*Proof.* Consider a probability space with a uniform random variable $U$ on $[0, 1]$. We use the construction of Corollary G.2 and define $\widetilde{Y} = Q_Y(U)$ where $Q_Y(u) = \inf\{x \in \mathbb{R} \colon F_Y(x) > u\}$ and similarly define $\widetilde{X}_n = Q_{X_n}(U)$, using the same uniform random variable $U$.

By assumption, we have $F_{X_n}(x) \to F_Y(x)$ for all $x \in \mathbb{R}$ where $F_Y$ is continuous. Now fix any $u \in [0, 1]$ such that the pre-image $F_Y^{-1}(\{u\})$ has at most one element. Then $x := Q_Y(u)$

is the unique $x \in \mathbb{R}$ such that $F_Y(x-) \le u \le F_Y(x)$. In particular, for all $\varepsilon > 0$, we have $F_Y(x+\varepsilon) > u$, there is a continuity point $x' < x + \varepsilon$ of $F_Y$ such that

$$F_{X_n}(x') \to F_Y(x') > u \quad \Rightarrow \quad F_{X_n}(x+\varepsilon) \ge F_{X_n}(x') > u \text{ for } n \text{ sufficiently large.}$$

Hence $Q_{X_n}(u) \le x + \varepsilon$ for $n$ sufficiently large. Similarly, $F_Y(x-\varepsilon) < u$ and there is a continuity point $x'' > x - \varepsilon$ of $F_Y$ such that

$$F_{X_n}(x'') \to F_Y(x'') < u \quad \Rightarrow \quad F_{X_n}(x-\varepsilon) \le F_{X_n}(x'') < u \text{ for } n \text{ sufficiently large.}$$

But then $Q_{X_n}(u) \in [x-\varepsilon, x+\varepsilon]$ for all $n$ sufficiently large, i.e. $Q_{X_n}(u) \to x = Q_Y(u)$. The set of $u$-values that we excluded is countable, so this entails that $\widetilde{X}_n = Q_{\widetilde{X}_n}(U) \to Q_Y(U) = \widetilde{Y}$ almost surely. □

The first convergence theorem we establish characterises convergence in distribution in terms of the convergence of a much wider class of expectations which includes the real and imaginary parts of characteristic functions. The equivalent condition in this theorem often serves as the definition of convergence in distribution and generalises straightforwardly to random variables taking values in $\mathbb{R}^d$ or other topological spaces.

**Theorem H.14.** *Suppose $X_1, X_2, \ldots$ and $Y$ are random variables. Then*

$$X_n \xrightarrow{d} Y \text{ as } n \to \infty \iff \forall g\colon \mathbb{R} \to \mathbb{R} \text{ bounded continuous, } \mathbb{E}[g(X_n)] \to \mathbb{E}[g(Y)] \text{ as } n \to \infty.$$

*Proof.* "⇒": If $X_n \xrightarrow{d} Y$, we may instead consider $\widetilde{X}_n \xrightarrow{\text{a.s.}} \widetilde{Y}$ where $\widetilde{X}_n \overset{d}{=} X_n$ for each $n \ge 1$ and $\widetilde{Y} \overset{d}{=} Y$, by Theorem H.13. Then for any bounded continuous $g\colon \mathbb{R} \to \mathbb{R}$, the Dominated Convergence Theorem then yields $\mathbb{E}[g(X_n)] = \mathbb{E}[g(\widetilde{X}_n)] \to \mathbb{E}[g(\widetilde{Y})] = \mathbb{E}[g(Y)]$.

"⇐": Let $x \in \mathbb{R}$ be such that $F_Y$ is continuous at $x$. We have to show that $F_{X_n}(x) \to F_Y(x)$, i.e. $\mathbb{E}[g(X_n)] \to \mathbb{E}[g(Y)]$ for the discontinuous function $g = \mathbf{1}_{(-\infty,x]}$. To this end, let $\varepsilon > 0$. By continuity of $F_Y$ at $x$, there is $\delta > 0$ such that

$$F_Y(x+\delta) - \varepsilon/2 \le F_Y(x) \le F_Y(x-\delta) + \varepsilon/2.$$

There are continuous $g^{\pm}$ with $\mathbf{1}_{(-\infty,x-\delta]} \le g^- \le \mathbf{1}_{(-\infty,x]} \le g^+ \le \mathbf{1}_{(-\infty,x+\delta]}$, to which our hypothesis applies and yields $\mathbb{E}[g^{\pm}(X_n)] \to \mathbb{E}[g^{\pm}(Y)]$. In particular,

$$\exists n_0 \ge 0\ \forall n \ge n_0,\ \mathbb{E}[g^-(Y)] - \varepsilon/2 \le \mathbb{E}[g^-(X_n)] \le F_{X_n}(x) \le \mathbb{E}[g^+(X_n)] \le \mathbb{E}[g^+(Y)] + \varepsilon/2.$$

such that, as required,

$$\begin{aligned}\forall n \ge n_0,\ F_Y(x) - \varepsilon &\le F_Y(x-\delta) - \varepsilon/2 \le \mathbb{E}[g^-(Y)] - \varepsilon/2 \\ &\le F_{X_n}(x) \le \mathbb{E}[g^+(Y)] + \varepsilon/2 \le F_Y(x+\delta) + \varepsilon/2 \le F_Y(x) + \varepsilon.\end{aligned}$$

□

We will use the following convergence criterion.

**Lemma H.15.** *A sequence of random variables $X_n$, $n \ge 1$, converges in distribution to $Y$ if and only if the following two conditions hold.*

*A. The family of distributions of $X_n$, $n \geq 1$, is* tight *in that*

$$\forall \varepsilon > 0 \ \exists c_0 \geq 0 \text{ such that } \forall n \geq 1, \ \mathbb{P}(|X_n| \geq c_0) < \varepsilon.$$

*B. All subsequences of $(X_n)$ that converge in distribution, converge in distribution to $Y$.*

*Proof.* "⇒": For condition A, let $\varepsilon > 0$. By properties of the limiting cumulative distribution function, there is $c_0$ such that $\mathbb{P}(|Y| \geq c_0) < \varepsilon/2$, without loss of generality such that $\mathbb{P}(|Y| = c_0) = 0$. By convergence in distribution,

$$\mathbb{P}(X_n \leq -c_0) \leq \mathbb{P}(Y \leq -c_0) + \varepsilon/4 \quad \text{and} \quad \mathbb{P}(X_n \leq c_0) \geq \mathbb{P}(Y \leq c_0) - \varepsilon/4.$$

Hence, $\mathbb{P}(|X_n| \geq c_0) \leq \mathbb{P}(|Y| \geq c_0) + \varepsilon/2 < \varepsilon$. Condition B is straightforward.

"⇐": Assume that $X_n$ does not converge to $X$ in distribution, then there is $x \in \mathbb{R}$ where $F_Y$ is continuous but such that $F_{X_n}(x)$ does not converge to $F_Y(x)$. We can therefore find $\varepsilon > 0$ and a subsequence $(X_{n(m)})$ along which $|F_{X_{n(m)}}(x) - F_Y(x)| > \varepsilon$ for all $m \geq 1$. To find the desired contradiction (to Condition B), it suffices to show that $(X_{n(m)})$ has a subsequence that converges in distribution.

To this end, consider an enumeration $q_i$, $i \geq 1$, of $\mathbb{Q}$. Since $[0,1]$ is compact, the Bolzano-Weierstrass Theorem, Theorem B.12, allows us to extract convergent subsequences that we can inductively refine to have cumulative distribution functions converge at $q_1, \ldots, q_k$ along the $k$th subsequence. From this $k$th subsequence, we take the $k$th term, denoted by $X_{n(m(k))}$, for each $k \geq 1$, to build a "diagonal" subsequence $(X_{n(m(k))})$ along which cumulative distribution functions $F_{X_{n(m(k))}}$ converge at all $q_i$, $i \geq 1$, to some limit $\widetilde{F}\colon \mathbb{Q} \to [0,1]$, which is increasing as a limit of increasing functions.

For convergence in distribution of $(X_{n(m(k))})$, we will not care about the value of $\widetilde{F}$ where left and right limits differ and define the right-continuous modification and extension $F(r) = \inf\{\widetilde{F}(q)\colon q \in \mathbb{Q} \cap (r, \infty)\}$, $r \in \mathbb{R}$. Now consider any $r \in \mathbb{R}$ where $F$ is continuous. By continuity, we can find three rationals $q < q_- \leq r < q_+$ such that

$$F(r) - \varepsilon < F(q) \leq F(q_-) \leq F(r) \leq F(q_+) < F(r) + \varepsilon.$$

By convergence of $F_{X_{n(m(k))}}$ to $\widetilde{F} \leq F$ at $q_\pm$ and noting that $\widetilde{F}(q_-) \geq F(q)$, we conclude that for all $k$ sufficiently large

$$F(r) - \varepsilon < F_{X_{n(m(k))}}(q_-) \leq F_{X_{n(m(k))}}(r) \leq F_{X_{n(m(k))}}(q_+) < F(r) + \varepsilon.$$

For $(X_{n(m(k))})$ to converge in distribution, we need to further ensure that the limit $F$ is a cumulative distribution function. As $F$ is increasing and right-continuous, it remains to show that $F(r)$ tends to 0 and 1 as $r \to -\infty$ and $r \to \infty$, respectively. By monotonicity and boundedness in $[0,1]$, the limits always exist in $[0,1]$, so the only way they can fail to be 0 or 1 is if they are $\delta$ or $1 - \delta$ for some $\delta > 0$. But this contradicts Condition A for $\varepsilon = \delta/2$ since for $k$ sufficiently large, Condition A entails for any $c_1 > c_0$ such that $F$ is continuous at $\pm c_1$

$$\begin{aligned} &F_{X_{n(m(k))}}(-c_1) \leq \mathbb{P}(|X_{n(m(k))}| \geq c_0) < \varepsilon = \delta/2 < \delta \\ \text{and} \quad &F_{X_{n(m(k))}}(c_1) \geq 1 - \mathbb{P}(|X_{n(m(k))}| \geq c_0) > 1 - \varepsilon = 1 - \delta/2 > 1 - \delta, \end{aligned}$$

and these inequalities are preserved in the limit as $k \to \infty$ giving $F(-c_1) \leq \delta/2 < \delta$ and $F(c_1) \geq 1 - \delta/2 > 1 - \delta$. □

**Theorem H.16** (Convergence theorem for characteristic functions)**.** *Suppose $X_1, X_2, \ldots$ and $Y$ are random variables. Then*

$$X_n \xrightarrow{d} Y \text{ as } n \to \infty \iff \forall t \in \mathbb{R},\ \phi_{X_n}(t) \to \phi_Y(t) \text{ as } n \to \infty.$$

*Proof.* "⇒": This follows from Theorem H.14, applied to real and imaginary parts.

"⇐": We apply Lemma H.15. For Condition A, note that characteristic functions are continuous, by the Dominated Convergence Theorem. In particular $\phi_Y(t) \to \phi_Y(0) = 1$ as $t \downarrow 0$. Now let $\varepsilon > 0$. Then we can find $m > 0$ such that

$$\frac{1}{2m}\int_{-m}^{m}(1-\phi_Y(t))dt < \frac{\varepsilon}{4}.$$

By hypothesis, we have $\phi_{X_n}(t) \to \phi_Y(t)$, so by dominated convergence, we have, for $n$ sufficiently large,

$$\frac{1}{2m}\int_{-m}^{m}(1-\phi_{X_n}(t))dt < \frac{\varepsilon}{2},$$

where Fubini's theorem applied to the bounded measurable function $f(t,\omega) = (1 - e^{itX_n(\omega)})$ on $[-m,m] \times \Omega$ yields

$$\frac{1}{2m}\int_{-m}^{m}(1-\phi_{X_n}(t))dt = \mathbb{E}\left[\frac{1}{2m}\int_{-m}^{m}(1-e^{itX_n})dt\right] = \mathbb{E}\left[1 - \frac{\sin(mX_n)}{mX_n}\right].$$

Hence

$$\begin{aligned}\mathbb{P}(|X_n| \geq 2/m) &\leq \mathbb{E}\left[2\left(1 - \frac{1}{|mX_n|}\right)\mathbf{1}_{\{|X_n|\geq 2/m\}}\right] \\ &\leq 2\mathbb{E}\left[1 - \frac{\sin(mX_n)}{mX_n}\right] < \varepsilon.\end{aligned}$$

For Condition B we just note that the convergence in distribution of any subsequence $(X_{n(k)})$ entails the convergence of their characteristic functions to the characteristic function of the limiting distribution, by Theorem H.14, as argued in the "⇒" direction. But the characteristic functions of $(X_{n(k)})$ converge to the characteristic function of $Y$ by hypothesis. By the uniqueness theorem, the limiting distribution of $(X_{n(k)})$ is the distribution of $Y$. □

**Theorem H.17** (Convergence theorem for moment generating functions)**.** *Suppose $X_1, X_2, \ldots$ and $Y$ are random variables whose moment generating functions $M_{X_1}, M_{X_2}, \ldots$ and $M_Y$ are all finite on $[-t_0, t_0]$ for some $t_0 > 0$. Then*

$$\forall t \in [-t_0, t_0],\ M_{X_n}(t) \to M_Y(t) \text{ as } n \to \infty \quad \Rightarrow \quad X_n \xrightarrow{d} Y \text{ as } n \to \infty.$$

*Proof.* We apply Lemma H.15. For Condition A, we obtain from Markov's inequality

$$\forall c \geq 0 \quad \mathbb{P}\big(|X_n| \geq c\big) = \mathbb{P}\big(e^{t_0|X_n|} \geq e^{t_0 c}\big) \leq e^{-t_0 c}\mathbb{E}\big[e^{t_0|X_n|}\big] \leq e^{-t_0 c}\big(M_{X_n}(t_0) + M_{X_n}(-t_0)\big).$$

Now let $\varepsilon > 0$. By hypothesis, the sequence $(M_{X_n}(t_0) + M_{X_n}(-t_0))$ converges and is hence bounded, by $C$, say, so that we can find $c = c_0$ sufficiently large to make the right-hand side smaller than $\varepsilon$ for all $n \geq 1$.

For Condition B, consider any subsequence $(X_{n(k)})$ with $X_{n(k)} \xrightarrow{d} Z$ for some random variable $Z$. Note that $g(r) = C\exp(-t_0\log(r)/t) = Cr^{-t_0/t}$ is integrable over $[1,\infty)$ for all $t \in (0,t_0)$ and the Dominated Convergence Theorem applies to $f_n(r) = \mathbb{P}(|X_n| > \log(r)/t)$, $n \geq 1$, by the display above. Then applying Proposition E.9 to $e^{t|X_{n(k)}|}$, $[1,\infty)$-valued, yields

$$\mathbb{E}[e^{t|X_{n(k)}|}] = 1 + \int_1^\infty \mathbb{P}\left(|X_{n(k)}| > \frac{\log(r)}{t}\right) dr \to 1 + \int_1^\infty \mathbb{P}\left(|Z| > \frac{\log(r)}{t}\right) dr = \mathbb{E}[e^{t|Z|}].$$

Extending $g(r) = 1$ on $[0,1)$, the same argument for one-sided tails $\mathbb{P}(X_{n(k)} > \log(r)/t) \leq \mathbb{P}(|X_{n(k)}| > \log(r)/t)$ yields $\mathbb{E}[e^{tX_{n(k)}}] \to \mathbb{E}[e^{tZ}]$ for all $t \in (0,t_0)$, and similarly for $t \in (-t_0,0)$. The case $t = 0$ is trivial. By uniqueness of limits, $\mathbb{E}[e^{tY}] = \mathbb{E}[e^{tZ}]$ for all $t \in [-t_0/2, t_0/2]$ and by the Uniqueness Theorem for moment generating functions, $Z \stackrel{d}{=} Y$, as required. □

### H.3.4 Further details relevant for the generating functions chapter

*Remark* H.18. In Section 3.2.2 we use that $(1+a/n+o(1/n))^n \to e^a$. In other words, if $a_n \to a$, then $(1+a_n/n)^n \to e^a$. In anticipation of applications for characteristic functions, let us here consider $a_n \in \mathbb{C}$, $n \geq 1$, and $a \in \mathbb{C}$. By continuity of exp and of the holomorphic log on the ball of radius 1 around 1 with $\log(1) = 0$, this is equivalent to $n\log(1+a_n/n) \to a$. Taylor's theorem gives $\log(1+z)/z \to 1$ as $z \to 0$. For $a \neq 0$, this entails $(n/a_n)\log(1+a_n/n) \to 1$, which is, as required. For $a = 0$, $\log(1+z)/z \to 1$ gives in particular that $|\log(1+z)| \leq 2|z|$ for $|z|$ sufficiently small. This entails that $|n\log(1+a_n/n)| \leq 2|a_n| \to 2|a| = 0$, as required.

*Remark* H.19. In (3.2), we claim a Taylor expansion for characteristic functions. Whereas for moment generating functions, the existence of moments followed from the existence of the moment generating function on some $[-t_0,t_0]$, $t_0 > 0$, this does not translate to characteristic functions, which always exist regardless of the existence of moments. We therefore assume that $\mathbb{E}[|X|^k] < \infty$ and then note that $|i^k X^k e^{itX}| = |X|^k$ for all $t \in \mathbb{R}$, and the Differentiability Lemma, Theorem G.20, yields that $\phi_X$ is $k$ times continuously differentiable with

$$\phi_X^{(k)}(t) = i^k\mathbb{E}[X^k e^{itX}]$$

Since clearly $\phi_X^{(k)}(0) = i^k\mathbb{E}[X^k]$ and $\phi_X(0) = 1$, Taylor's theorem (for the real and imaginary parts) yields, as $t \to 0$,

$$\phi_X(t) = 1 + it\mathbb{E}[X] + i^2t^2\frac{\mathbb{E}[X^2]}{2!} + \cdots + i^{k-1}t^{k-1}\frac{\mathbb{E}[X^{k-1}]}{(k-1)!} + o(t^{k-1}).$$

where we stress that the argument so far has only provided an expansion up to order $k-1$, not $k$. In particular, the case $k = 2$ relevant for the weak law of large numbers requires $\mathbb{E}[X^2] < \infty$ and the case $k = 3$ relevant for the central limit theorem requires $\mathbb{E}[|X|^3] < \infty$.

This can be improved, though, and we will prove the expansion up to order $k$ in (3.2) under the assumption $\mathbb{E}[|X|^k] < \infty$. Let us restate this here.

**Proposition H.20.** *Let $X$ be a random variable with characteristic function $\phi_X$ and finite $k$th moment $\mathbb{E}[|X|^k] < \infty$. Then*

$$\phi_X(t) = 1 + it\mathbb{E}[X] + i^2t^2\frac{\mathbb{E}[X^2]}{2!} + \cdots + i^kt^k\frac{\mathbb{E}[X^k]}{(k)!} + o(t^k).$$

*Proof.* Rather than applying the Taylor expansion to $\phi_X$ directly, we will use a Taylor expansion inside the expectation. Specifically, we expand $e^{iy}$ around 0, and we derive an alternative remainder term directly. We set

$$h_k(y) = e^{iy} - \sum_{m=0}^{k} \frac{(iy)^m}{m!},$$

and we note that $h_k(0) = 0$ and $h_k'(y) = ih_{k-1}(y)$ for all $k \geq 1$. We also note that $|h_0(y)| \leq 2$. We proceed inductively and obtain bounds for all $k \geq 1$

$$|h_{k-1}(y)| \leq 2\frac{|y|^{k-1}}{(k-1)!} \quad \Rightarrow \quad |h_k(y)| = \left| \int_0^y ih_{k-1}(s)ds \right| \leq 2\int_0^{|y|} \frac{|s|^{k-1}}{(k-1)!}ds = 2\frac{|y|^k}{k!}.$$

Similarly, we can start from the estimate

$$|h_0(y)| = |e^{iy} - 1| = \left| \int_0^y \frac{1}{i} e^{is} ds \right| \leq \int_0^{|y|} |e^{is}| ds = |y|$$

and show inductively that $|h_k(y)| \leq |y|^{k+1}/(k+1)!$ for all $k \geq 0$. In particular, this entails

$$\left| \phi_X(t) - \sum_{m=0}^{k} i^m t^m \frac{\mathbb{E}[X^m]}{m!} \right| \leq \mathbb{E}\Big[ \big| h_k(itX) \big| \Big] \leq t^k \frac{\mathbb{E}[\min\{t|X|^{k+1}, 2(k+1)|X|^k\}]}{(k+1)!}$$

where on the one hand the expectation is finite provided that just $\mathbb{E}[|X|^k] < \infty$, and on the other hand it tends to 0 as $t \to 0$, by the Dominated Convergence Theorem. □

*Remark* H.21. We then also say "Apart from working with complex power series instead of real power series, there are no additional complications when translating the proof from mgfs to cfs." This is now easily checked, by replacing all mgfs by cfs and $t$ by $it$, as appropriate.

**Proposition H.22.** *The Cauchy distribution with probability density function* $f(x) = \frac{1}{\pi(1+x^2)}$, $x \in \mathbb{R}$, *has characteristic function* $\phi(t) = e^{-|t|}$.

*Proof.* Let $t > 0$. Consider the holomorphic function

$$g(z) = \frac{e^{itz}}{\pi(1+z^2)} = \frac{1}{z-i}\frac{e^{itz}}{\pi(z+i)} \quad \text{with} \quad \mathrm{Res}_i(g) = \frac{e^{-t}}{2\pi i} \text{ at the simple pole } i.$$

Consider the semi-circular contour of radius $R > 1$ in the upper half-plane. Then the integral along the (anti-clockwise) semi-circle $\gamma_R$ vanishes as $R \to \infty$ by Jordan's Lemma, applied to the meromorphic function $f(z) = \frac{1}{\pi(1+z^2)}$ that satisfies $f(z) \to 0$ as $|z| \to \infty$. The only singularity inside the semi-circular contour is at $z = i$. By the Residue Theorem,

$$e^{-t} = 2\pi i \,\mathrm{Res}_i(g) = \int_{-R}^{R} g(x)dx + \int_{\gamma_R} g(z)dz \to \int_{-\infty}^{\infty} e^{itx} f(x)dx.$$

By symmetry, $\phi(-t) = \phi(t) = e^{-t}$. Since also $\phi(0) = 1$, this gives altogether $\phi(t) = e^{-|t|}$ for all $t \in \mathbb{R}$. □

## H.4 Joint distribution for continuous random variables

In Section 4.1 we express probabilities as bivariate integrals over bivariate Borel sets for bivariate jointly continuous random variables. More formally, the Borel $\sigma$-algebra is defined to be the smallest $\sigma$-algebra that contains all rectangles of the form $I_1 \times \cdots \times I_n$ where $I_1, \ldots, I_n \subseteq \mathbb{R}$ are intervals. Recall that all intervals can be expressed via complements, countable unions and countable intersections in terms of intervals of the form $(-\infty, b]$. An analogous statement holds for rectangles in $\mathbb{R}^n$, which can be expressed in terms of rectangles of the form $(-\infty, b_1] \times \cdots \times (-\infty, b_n]$. Specifically, for $A = (a_1, b_1] \times \cdots \times (a_n, b_n]$ for $-\infty \leq a_j < b_j < \infty$, we can achieve this by taking differences noting that

$$\begin{aligned}
&(a_1, b_1] \times \cdots (a_{j-1}, b_{j-1}] \times (a_j, b_j] \times (-\infty, b_{j+1}] \times \cdots (-\infty, b_n] \\
&= \Big( (a_1, b_1] \times \cdots (a_{j-1}, b_{j-1}] \times (-\infty, b_j] \times (-\infty, b_{j+1}] \times \cdots (-\infty, b_n] \Big) \\
&\qquad \setminus \Big( (a_1, b_1] \times \cdots (a_{j-1}, b_{j-1}] \times (-\infty, a_j] \times (-\infty, b_{j+1}] \times \cdots (-\infty, b_n] \Big).
\end{aligned}$$

**Theorem H.23.** *Let $X = (X_1, \ldots, X_n)$ be a jointly continuous random vector with joint p.d.f. $f_X$. Then for all Borel sets $A \in \mathcal{M}_{\mathrm{Bor}}(\mathbb{R}^n)$, the set $\{X \in A\} = \{\omega \in \Omega \colon X(\omega) \in A\}$ is an event (in $\mathcal{F}$) and we have*

$$\mathbb{P}(X \in A) = \int_A f_X(x) dx,$$

*where this integral is against Lebesgue measure on $\mathbb{R}^n$.*

*Proof.* We adapt the proof of Theorem G.7, which covers the case $n = 1$. For $n \geq 2$, the claim holds for $A = (-\infty, x_1] \times \cdots \times (-\infty, x_n]$ by definition. Since the Borel $\sigma$-algebra of $\mathbb{R}^n$ is generated by the collection of subsets of this form, the remainder of the argument of Theorem G.7 can be read verbatim as an argument in $\mathbb{R}^n$. □

*Remark* H.24. The Probability part of the proof of Theorem 4.1 is complete. As mentioned before in a different context, what has not been proved rigorously is the change of variables formula for two- or higher-dimensional integrals. A version of the two-dimensional case was stated but not proved in Theorem F.26 within Part A Integration.

Let us state here the version provided as a non-examinable 15-page document in the Part A Integration course. This is sufficient for our purposes.

**Theorem H.25.** *Consider two open sets $D, R \subseteq \mathbb{R}^n$ and a bijective transformation $T \colon D \to R$ that is continuously differentiable with continuously differentiable inverse. Then a function $f \colon R \to \mathbb{R}$ is Lebesgue integrable over $R$ if and only if $(f \circ T)|\mathrm{det} J_T|$ is Lebesgue integrable over $D$. In that case*

$$\int_R f = \int_D (f \circ T)|\mathrm{det} J_T|.$$

*Remark* H.26. The heuristic derivation of conditional densities in (4.3) is claimed to be justified if $f_{X,Y}$ is "sufficiently smooth." Let us explore conditions under which this holds. In fact, it suffices to have the following limits

$$\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \int_{u=x}^{x+\varepsilon} f_X(u) du = f_X(x) \text{ and } \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \int_{u=x}^{x+\varepsilon} \int_{v=-\infty}^{y} f_{X,Y}(u, v) dv du = \int_{v=-\infty}^{y} f_{X,Y}(x, v) dv.$$

The first one holds if $f_X$ is right-continuous at $x$. In the same way, the second one follows if we establish that $u \mapsto \int_{v=-\infty}^{y} f_{X,Y}(u,v)dv$ is continuous at $u = x$. This holds by the Dominated Convergence Theorem if we assume that $u \mapsto f_{X,Y}(u,v)$ is right-continuous at $x$ for all $v$ and $f_{X,Y}(u,v) \le g(v)$ for some function $g$ that is integrable on $(-\infty, y]$ for all $y$. If we strengthen the integrability of $g$ to integrability on $\mathbb{R}$, this also includes the right-continuity of $f_X$ at $x$.

*Remark* H.27. The purpose of Section 4.4.3 is to avoid getting bogged down in lengthy transformation formula arguments for bivariate normal distributions. We claim that having

$$Y = \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1) + \sqrt{1-\rho^2}\sigma_2 Z_2 + \mu_2 =: g_X(Z_2)$$

for independent $X$ and $Z_2$ entails that the conditional distributions of $Y = g_X(Z_2)$ given $X = x$ is just the distribution of $g_x(Z_2)$. You may feel that this makes intuitive sense since when $X = x$, all randomness of $Z_2$ still remains. More formally, we can easily check that by a quick argument involving the transformation formula for probability density functions,

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X,Z_2}(x, g_x^{-1}(y))|(g_x^{-1})'(y)|}{f_X(x)} = f_{Z_2}(g_x^{-1}(y))|(g_x^{-1})'(y)| = f_{g_x(Z_2)}(y).$$

## H.5 Markov chains: Introduction

**Lemma H.28.** *Consider a countable set $I$, an initial distribution $\lambda$ and a transition matrix $P = (p_{ij})_{i,j\in I}$. Then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we can define a Markov chain $(Y_n)_{n\ge 0}$ with initial distribution $\lambda$ and transition matrix $P$.*

*Proof.* Without loss of generality, $I$ is countably infinite (the finite case being analogous). Fix bijections $f\colon I \to \mathbb{N}$ and write $g = f^{-1}$. Then, $p_{i,g(n)}$, $n \in \mathbb{N}$, $p_{i,x} = 0$, $x \notin \mathbb{N}$, is a probability mass function on $\mathbb{R}$, for each $i \in I$. Consider the associated cumulative distribution functions $F_i\colon \mathbb{R} \to [0,1]$, $i \in I$ and let $Q_i(u) = \inf\{x \in \mathbb{R}\colon F_i(x) > u\}$, $u \in [0,1]$.

Let $U$ be uniformly distributed on $[0,1]$. Recall from Corollary G.2 that $Q_i(U)$ has cumulative distribution function $F_i$, and hence $\mathbb{P}(g(Q_i(U)) = j) = \mathbb{P}(Q_i(U) = f(j)) = p_{ij}$ for all $i, j \in I$. Similarly, the cumulative distribution function $F_\lambda$ associated with $\lambda_{g(n)}$, $n \in \mathbb{N}$, and $Q_\lambda(u) = \inf\{x \in \mathbb{R}\colon F_\lambda(x) > u\}$, $u \in [0,1]$, gives access to a random variable $g(Q_\lambda(U))$ with the initial distribution.

Now consider the setting of Corollary G.4, where $X_{k-1} := U_k$, $k \ge 1$, is a sequence of independent uniform random variables. We claim that the process $Y_0 = g(Q_\lambda(X_0))$, $Y_n = g(Q_{Y_{n-1}}(X_n))$, $n \ge 1$, is a Markov chain with initial distribution $\lambda$ and transition matrix $P$. To see this, note that for all $n \ge 0$ and all $i_0, \dots, i_n \in I$

$$\begin{aligned}
&\{Y_0 = i_0, Y_1 = i_1, \dots, Y_n = i_n\}\\
&= \{g(Q_\lambda(X_0)) = i_0, g(Q_{i_0}(X_1)) = i_1, \dots, g(Q_{i_{n-1}}(X_n)) = i_n\}\\
&= \{F_\lambda(f(i_0)) < X_0 \le F_\lambda(f(i_0)+1),\\
&\qquad F_{i_0}(f(i_1)) < X_1 \le F_{i_0}(f(i_1)+1), \dots, F_{i_{n-1}}(f(i_n)) < X_n \le F_{i_{n-1}}(f(i_n)+1)\}
\end{aligned}$$

and hence by independence of $X_0, \dots, X_n$,

$$\mathbb{P}(Y_0 = i_0, Y_1 = i_1, \dots, Y_n = i_n)$$

$$= \mathbb{P}(g(Q_\lambda(X_0)) = i_0)\mathbb{P}(g(Q_{i_0}(X_1)) = i_1) \cdots \mathbb{P}(g(Q_{i_{n-1}}(X_n)) = i_n)$$
$$= \lambda_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n},$$

as required for a Markov chain with initial distribution $\lambda$ and transition matrix $P$. □

*Remark* H.29. One may question why we only considered the joint distributions of finitely many random variables, not of the entire sequence. We are interested in events of the form

$$E = \{W_n \neq i \text{ for all } n \geq 0\} \quad \text{or} \quad E' = \{W \text{ visits } i \text{ infinitely often}\}$$

that depend on infinitely many random variables. While $E$ is easily well-approximated by events only depending on finitely many random variables, this may at first not seem to be the case for $E'$ since $E'$ poses no restriction on any finite number of $W_n$. However, this can still be done:

$$E' = \bigcap_{k \geq 1} \bigcup_{1 \leq n_1 < \cdots < n_k} \{W_{n_j} = i, 1 \leq j \leq k\}$$

so the probabilities of such events of interest to us are determined by the distributions of any finite family of random variables. This can be formalised further in Kolmogorov's consistency theorem which in our context essentially says that (a consistent system of) distributions of any finite family of random variables gives rise to a unique distribution of the sequence of random variables, and this means that probabilities of all events that depend on the sequence are indeed determined by this unique extension. We will not make precise what the meaning of this statement is nor how to prove anything like this.

*Remark* H.30. We use the term "transition matrix" also when the state space is countably infinite. This means that the entries of products of transition matrices are (convergent!) series. Most prominently, in the setting of Theorem 5.2, we use notation $P^n$ for the $n$th power of $P$. In general, such "matrix" products require certain assumptions, but in our context all entries and the row sums are bounded by 1, so even before the identification with $n$-step transition probabilities, the convergence of the series is guaranteed by simple comparison tests.

*Remark* H.31. Sections 5.7 ad 5.8 involve successive hitting times. The fact that the process after the hitting time is independent of the process before the hitting time (and that the process after the hitting time is another Markov chain) can be proved directly using the Markov property as discussed in Seciton 5.2, or it is an instance of the Strong Markov property that generalises the (simple) Markov property from fixed times to certain random times.

**Definition H.32.** Consider a Markov chain $W$ with state space $S$. A *stopping time* is a random variable $N \colon \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$, for which the event $\{N = n\}$ depends only on $W_0, \ldots, W_n$, for all $n \geq 0$.

**Lemma H.33.** *For any $B \subseteq S$, the first hitting time $H^B = \inf\{n \geq 0 \colon W_n \in B\}$ is a stopping time*

This is an (elementary) optional exercise in Part A Probability.

**Theorem H.34** (Strong Markov property)**.** *Let $W$ be a Markov chain and $N$ a stopping time. Then conditionally given $\{N < \infty, W_N = i\}$, the post-$N$ process $(W_{N+k}, k \geq 0)$ is a Markov chain starting from $i$ that is conditionally independent of $(W_j, 0 \leq j \leq N-1)$.*

This is a (harder) optional exercise in Part A Probability. Let us explore the meaning of this statement here.

*Remark* H.35. In the exploration of the (simple) Markov property (the case $N = n$ deterministic in Theorem H.34) in Section 5.2, we expressed the conditional independence in several ways relating joint (conditional) distributions. As noted in Remark H.29, it is natural and indeed sufficient for our purposes and much more generally to consider distributions of finitely many random variables. In the statement of the strong Markov property, this remark applies to both the post-$N$ process and the pre-$N$ process, which has a random length.

Specifically, if $\mathbb{P}(N = \infty) = 0$, the random process $(W_j, 0 \le j \le N-1)$ takes values in the countable set $\bigcup_{n\ge 0} S^n$, except on an event $\{N = \infty\}$ of probability 0, so its distribution is naturally expressed as the probability mass function

$$\mathbb{P}((W_j, 0 \le j \le N-1) = (i_j, 0 \le j \le n-1)), \qquad (i_j, 0 \le j \le n-1) \in S^n, \quad n \ge 0.$$

More generally, we can use events of the form

$$E = \bigcup_{0\le n<\infty} \left\{W_0 \in A_0^{(n)}, \dots, W_{n-1} \in A_{n-1}^{(n)}\right\} \cap \{N = n\}.$$

If $\mathbb{P}(N = \infty) > 0$, the statement of the Markov is still conditional given $N < \infty$ and the above probability mass function expresses the relevant part of the distribution of the pre-$N$ process $(W_j, 0 \le j \le N-1)$, on $\{N < \infty\}$, while for the sake of also capturing the distribution on $\{N = \infty\}$, we remark that we can continue to rely on the observation in Remark H.29 that the (consistent) specification of the joint distributions of finitely many random variables also identifies probabilities of events of interest that depend on the entire sequence. For applications beyond the strong Markov property, we may therefore complement the above by events of the form

$$E = \{W_0 \in A_0^{(\infty)}, \dots, W_{k-1} \in A_{k-1}^{(\infty)}, N = \infty\}.$$

or combine the above events more generally into events of the form

$$E = \{W_0 \in A_0, \dots, W_{k-1} \in A_{k-1}, N \ge k\}.$$

Let us note three corollaries of the strong Markov property that clarify steps of arguments respectively in the remark offering an alternative approach at the end of the gambler's ruin example in Section 5.7, in the setup of recurrence and transience at the beginning of Section 5.8, and when we revisit the gambler's ruin in Section 5.10.1.

**Corollary H.36.** *Let $X$ be a general birth-and-death chain with non-trivial transition probabilities*

$$p_{i,i+1} = p_i, \quad i \ge 0, \qquad p_{i,i-1} = 1 - p_i, \quad i \ge 1, \qquad p_{0,0} = 1 - p_0$$

*for some $p_i \in [0,1]$, $i \ge 0$. Let $r_i = \mathbb{P}_i(\text{hit } i-1)$, $i \ge 1$. Then $\mathbb{P}_i(\text{hit } 0) = r_i \cdots i_1$, $i \ge 1$.*

*Proof.* For $i = 1$, this is trivial. Inductively assuming the claim for some $i \ge 1$, we note that with probability 1, birth-and-death processes starting from $i+1$ must visit $i$ before visiting 0, i.e. $H^{\{i\}} \le H^{\{0\}}$. We apply the strong Markov property at the stopping time $H^{\{i\}}$ to

find that conditionally given $\{H^{\{i\}} < \infty\} = \{H^{\{i\}} < \infty, X_{H^{\{i\}}} = i\}$, the post-$H^{\{i\}}$ process $(X_{H^{\{i\}}+k}, k \geq 0)$ is distributed like $X$ starting from $i$ and

$$\begin{aligned}\mathbb{P}_{i+1}(\text{hit } 0) = \mathbb{P}_{i+1}(H^{\{0\}} < \infty) &= \mathbb{P}_{i+1}(H^{\{i\}} < \infty)\mathbb{P}_{i+1}(H^{\{0\}} < \infty | H^{\{i\}} < \infty) \\ &= \mathbb{P}_{i+1}(\text{hit } i)\mathbb{P}_i(H^{\{0\}} < \infty) = r_{i+1}r_i \cdots r_1.\end{aligned}$$

□

**Corollary H.37.** *Suppose that $(X_n, n \geq 0)$ is a Markov chain on a state space $S$ and that $i \in S$ is a state that is transient in the sense that*

$$\mathbb{P}_i(X_n = i \text{ for some } n \geq 1) = p < 1.$$

*Then the total number of visits to $i$ has geometric distribution with parameter $1 - p$.*

*Proof.* Let $R_0^{\{i\}} = 0$ and denote by $R_m^{\{i\}} = \inf\{n \geq R_{m-1}^{\{i\}} + 1 \colon X_n = i\}$, $m \geq 1$, the successive return times, with the conventions that $\inf \emptyset = \infty$ and that $R_m^{\{i\}} = \infty$ if $R_{m-1}^{\{i\}} = \infty$. Then the number $G$ of returns to $i$ satisfies

$$\mathbb{P}_i(G = 0) = \mathbb{P}_i(R_1^{\{i\}} = \infty) = \mathbb{P}_i(X_n \neq i \text{ for all } n \geq 1) = 1 - p \quad \text{and} \quad \mathbb{P}_i(G \geq 1) = p.$$

The strong Markov property at $R_m^{\{i\}}$, which is easily seen to be a stopping time, yields that conditionally given $\{R_m^{\{i\}} < \infty\} = \{R_m^{\{i\}} < \infty, X_{R_m^{\{i\}}} = i\}$, the post-$R_m^{\{i\}}$ process $(X_{R_m^{\{i\}}+k}, k \geq 0)$ has the same distribution as $X$, starting from $i$. Note that the $(m+1)^{\text{st}}$ return time of $X$ is finite if and only if the first return time of the post-$R_m^{\{i\}}$ process is finite. Inductively, assuming that $\mathbb{P}(G \geq m) = p^m$ for some $m \geq 1$, this yields

$$\begin{aligned}\mathbb{P}_i(G \geq m+1) = \mathbb{P}_i(R_{m+1}^{\{i\}} < \infty) &= \mathbb{P}_i(R_m^{\{i\}} < \infty)\mathbb{P}_i(R_{m+1}^{\{i\}} < \infty | R_m^{\{i\}} < \infty) \\ &= \mathbb{P}_i(R_m^{\{i\}} < \infty)\mathbb{P}_i(R_1^{\{i\}} < \infty) = p^{m+1}.\end{aligned}$$

□

**Corollary H.38.** *In the setting of Corollary H.36, let $d_i = \mathbb{E}_i[H^{\{i-1\}}]$, $i \geq 1$. Then $\mathbb{E}_i[H^{\{0\}}] = d_1 + \cdots + d_i$, $i \geq 1$.*

To deduce this carefully from the strong Markov property is an optional exercise in Part A Probability.

*Remark* H.39. In the proof of Theorem 5.8, interchanging expectation and series is justified by Tonelli's theorem.

**Proposition H.40.** *(a) Let $C$ be a recurrent communicating class. Either all states in $C$ are positive recurrent, or all are null recurrent (so we may refer to the whole class as positive recurrent or null recurrent).*

*(b) Every finite recurrent class is positive recurrent.*

*(c) If $C$ is positive recurrent, then $\mathbb{E}_j[\inf\{n \geq 1 \colon X_n = i\}] < \infty$ for all $i, j \in C$.*

This is an optional exercise in Part A Probability.

## H.6 Markov chains: stationary distributions and convergence to equilibrium

*Remark* H.41. In the proof of the ergodic theorem for Markov chains, there are three points of detail that we would like to clarify.

1. In the recurrent case, we claim that the times between successive visits of a state $i$ are i.i.d.. To prove this, we proceed as in the proof of Corollary H.37, where now $\{R_m^{(i)} < \infty\} = \{R_m^{(i)} < \infty, X_{R_m^{(i)}} = i\}$ has probability 1. Hence, the strong Markov property yields that $(X_{R_m^{(i)}+k}, k \geq 0)$ is (unconditionally!) independent of $(X_j, 0 \leq j \leq R_m^{(i)} - 1)$ and hence of $(R_1^{(i)}, R_2^{(i)} - R_1^{(i)}, \ldots, R_m^{(i)} - R_{m-1}^{(i)})$. Furthermore, $(X_{R_m^{(i)}+k}, k \geq 0)$ is a Markov chain starting from $i$ whose first return time to $i$ is $R_{m+1}^{(i)} - R_m^{(i)}$, which therefore has a distribution that does not depend on $m$. By induction, $R_1^{(i)}, R_2^{(i)} - R_1^{(i)}, \ldots, R_m^{(i)} - R_{m-1}^{(i)}, R_{m+1}^{(i)} - R_m^{(i)}$ are independent.

2. We use the strong law of large numbers for these i.i.d. random variables, which may or may not have finite fourth moment, indeed even for random variables with infinite expectation. In order to have a complete proof of the ergodic theorem, we therefore need to establish a sufficiently general strong law of large numbers. For finite mean, we leave this to Part B Probability, Measure and Martingales. For infinite mean, we explore this in the following theorem.

3. We claim that the asymptotics $T_k/k \to m_i \in (0, \infty]$ for the time $T_k$ of the $k$th visit to $i$ implies the asymptotics $V_i(n)/n \to 1/m_i \in [0, \infty)$ for the number $V_i(n)$ of visits to $i$ by time $n$. To prove this formally, we first note that recurrence implies that $V_i(n) \to \infty$ almost surely. Furthermore, we note that $T_{V_i(n)} \leq n \leq T_{V_i(n)+1}$, and on the event $\{V_i(n) \to \infty\}$, this yields
$$\frac{T_{V_i(n)}}{V_i(n)} \leq \frac{n}{V_i(n)} \leq \frac{T_{V_i(n)+1}}{V_i(n)+1} \frac{V_i(n)+1}{V_i(n)}.$$
By algebra of limits we conclude that on the event $\{V_i(n) \to \infty\} \cap \{T_k/k \to m_i\}$ of probability 1, we have sandwiched $n/V_i(n)$ between two sequences that converge to $m_i$. Hence, $V_i(n)/n$ converges to $1/m_i$ almost surely.

**Theorem H.42** (Strong law of large numbers for non-negative random variables with infinite mean)**.** *Let $X_n$, $n \geq 1$, be i.i.d. non-negative random variables with $\mu := \mathbb{E}[X_1] = \infty$. Let $S_n = X_1 + \cdots + X_n$, $n \geq 1$. Then*
$$\frac{S_n}{n} \to \mu = \infty \qquad \textit{almost surely, as } n \to \infty.$$

*Proof.* For any $K \in [0, \infty)$, consider the sequence $X_n^{(K)} := \min\{X_n, K\}$, $n \geq 1$, of i.i.d. random variables with finite mean $\mu_K := \mathbb{E}[X_1^{(K)}] < \infty$ and $\mathbb{E}\big[\big(X_1^{(K)}\big)^4\big] \leq K^4 < \infty$. Let $S_n^{(K)} = X_1^{(K)} + \cdots + X_n^{(K)}$, $n \geq 1$. By the strong law of large numbers for random variables with finite fourth moment,
$$\frac{S_n}{n} \geq \frac{S_n^{(K)}}{n} \to \mu_K \qquad \text{almost surely, as } n \to \infty.$$

Now note that $X_1^{(K)}$, $K \geq 1$, is an increasing sequence of random variables with limit $X_1$. The monotone convergence theorem yields that $\mu_K = \mathbb{E}[X_1^{(K)}] \to \mathbb{E}[X_1] = \mu = \infty$. Now fix $M \geq 1$. Then $\mu_K \geq M + 1$ for all $K$ sufficiently large and $S_n/n \geq M$ for all $n$ sufficiently large, except possibly on a set $A_M$ of probability 0. But then $S_n/n \to \infty$ except possibly on $\bigcup_{M \geq 1} A_M$, which also has probability zero by countable subadditivity. □

*Remark* H.43. In the informal proof of Lemma 6.11, we use the ergodic theorem to obtain the asymptotics for the number of visits to $i$ and $j$, but we also rely on the observation that a proportion $p_{i,j}$ of the visits to $i$ is followed by a transition into $j$. There are several ways to make this rigorous. Maybe the quickest is to make the second statement precise using the strong law of large numbers and then to carefully combine the two almost sure limits. In the following, we work out an alternative proof that is more explicit about the dependence structure induced by the repeated trials of transitioning to $j$ after visiting $i$.

Indeed, our aim is to apply the ergodic theorem directly. We consider $S = \{(i,j) \in I^2 \colon p_{i,j} > 0\} \subset I^2$ and the $S$-valued process $W_n = (X_n, X_{n+1})$, $n \geq 1$. It is easy to see that this is an irreducible Markov chain with (non-zero) transition probabilities, for $(i,j), (j,k) \in S$, given by

$$q_{(i,j),(j,k)} = \mathbb{P}(W_{n+1} = (j,k) | W_n = (i,j)) = \mathbb{P}(X_{n+2} = k | X_{n+1} = j, X_n = i) = p_{j,k}.$$

Then the expected return time $m_{(i,j)}$ by $W$ from $(i,j)$ back to itself can be split at the successive visits of $X$ to $i$, each of which has probability $p_{i,j}$ of being followed by a transition to $j$, independently of previous visits to $i$, by the strong Markov property. Hence, the number $G$ of visits to $i$ before the first transition from $i$ to $j$ is geometrically distributed with success probability $p_{i,j}$. The first visit to $i$ takes a number of steps with mean $\mathbb{E}_j[H^{\{i\}}] = \mathbb{E}_i\big[R_1^{\{i\}} | X_1 = j\big] - 1$, the unsuccessful trials take independent identically distibuted numbers $N_m$ of steps with mean $\mathbb{E}_i\big[R_1^{\{i\}} | X_1 \neq j\big]$, and the successful trial adds 1, so that

$$\begin{aligned}
m_{(i,j)} &= \mathbb{E}_i\big[R_1^{\{i\}} | X_1 = j\big] - 1 + \mathbb{E}\left[\sum_{m=1}^{G-1} N_m\right] + 1 \\
&= \mathbb{E}_i\big[R_1^{\{i\}} | X_1 = j\big] + \mathbb{E}[G-1]\mathbb{E}[N_1] \\
&= \mathbb{E}_i\big[R_1^{\{i\}} | X_1 = j\big] + \frac{1 - p_{i,j}}{p_{i,j}} \mathbb{E}_i\big[R^{\{i\}} | X_1 \neq j\big] \\
&= \frac{1}{p_{i,j}} \Big( \mathbb{E}_i\big[R_1^{\{i\}} | X_1 = j\big] \mathbb{P}_i(X_1 = j) + \mathbb{E}_i\big[R_1^{\{i\}} | X_1 \neq j\big] \mathbb{P}_i(X_1 \neq j) \Big) \\
&= \frac{1}{p_{i,j}} \mathbb{E}_i\big[R_1^{\{i\}}\big] = \frac{m_i}{p_{i,j}}.
\end{aligned}$$

This allows us to relate proportions $V_j(n)$ of $X$ and $V_{(i,j)}(n-1)$ of $W$ as

$$\frac{1}{m_j} \leftarrow \frac{V_j(n)}{n} = \frac{\mathbf{1}\{X_0 = j\}}{n} + \sum_{i \in I} \frac{V_{(i,j)}(n-1)}{n-1} \frac{n-1}{n} \to \sum_{i \in I} \frac{p_{i,j}}{m_i},$$

certainly when $I$ is finite, and for counably infinite $I$ by the argument presented at the end of the proof of Lemma 6.11.

*Remark* H.44. In the proof of Lemma 6.12, we indicate an argument based on convergence in probability to see that for all $\epsilon > 0$ and $n$ sufficiently large

$$\mathbb{E}\left[\frac{V_n(i)}{n}\right] = \mathbb{E}\left[\frac{V_n(i)}{n}\mathbf{1}\left\{\left|\frac{V_n(i)}{n} - \frac{1}{m_i}\right| > \varepsilon\right\}\right] + \mathbb{E}\left[\frac{V_n(i)}{n}\mathbf{1}\left\{\left|\frac{V_n(i)}{n} - \frac{1}{m_i}\right| \leq \varepsilon\right\}\right]$$
$$\begin{cases} \leq \mathbb{P}\left(\left|\frac{V_n(i)}{n} - \frac{1}{m_i}\right| > \varepsilon\right) + \frac{1}{m_i} + \varepsilon \leq \frac{1}{m_i} + 2\varepsilon \\ \geq 0 + \left(\frac{1}{m_i} - \varepsilon\right)\mathbb{P}\left(\left|\frac{V_n(i)}{n} - \frac{1}{m_i}\right| \leq \varepsilon\right) \geq \frac{1}{m_i} - (1 + \frac{1}{m_i})\varepsilon, \end{cases}$$

since $0 \leq V_i(n)/n \leq 1$, and hence

$$\frac{\mathbb{E}[V_n(i)]}{n} \to \frac{1}{m_i} \qquad \text{as } n \to \infty.$$

Alternatively, this follows straight from the dominated convergence theorem.

Recall that the proof of the ergodic theorem was based on the strong law of large numbers, whose proof is only completed in Part B Probability, Measure and Martingales, so the advantage of the above argument is that the strong law of large numbers, while not yet fully proved, is a key result for both theory and applications, while the dominated convergence theorem is a key result in a measure-theoretic approach to Probability.

*Remark* H.45. Lemma 6.12 shows that the existence of a stationary distribution on a class $C \subseteq I$ implies that $m_i < \infty$ for some $i \in C$, i.e. that some $i \in C$ is positive recurrent. Since positive recurrence is a class property, the same holds for all $i \in C$. Conversely, in the setting of the ergodic theorem, we note that $\sum_{i \in I} V_i(n)/n = 1$ for all $n \geq 1$, and as the terms under the sum converge to $1/m_i$, we obtain that $\sum_{i \in I}(1/m_i) = 1$ for finite $I$, by algebra of limits. For countably infinite $I$, this is more delicate. One way to extend to this case is to note that $\sum_{i \in J}(1/m_i) \leq 1$ for all finite $J \subset I$. But then, in the positive recurrent case, we have $0 < M := \sum_{i \in I}(1/m_i) \leq 1$ and $\pi_i = 1/(m_i M)$, $i \in I$, is a stationary distribution, which by Lemma 6.12 is actually $\pi_i = 1/m_i$, so $M = 1$.

**Proposition H.46.** *Let $X$ be an irreducible Markov chain with transition matrix $P$ and state space $S$. Then $i \in S$ is aperiodic if and only if $p_{i,i}^{(n)} > 0$ for all sufficiently large $n$.*

This is an optional exercise in Part A Probability, which is useful to formalise a step in the rather informal proof of the Markov chain convergence theorem given in the main body of the lecture notes. Another optional exercise is to make all informal steps in that proof formal.

## H.7 Poisson processes

*Remark* H.47. Poisson processes are families of uncountably many random variables. However, they do not require any bigger probability spaces because these uncountably many random variables can all be expressed as functions of a countable family of random variables. Specifically, the definition based on i.i.d. exponentially distributed inter-arrival times does precisely this. As we have demonstrated previously, there is a probability on which these

i.i.d. exponentially distributed random variables $Y_n\colon \Omega \to [0,\infty)$, $n \geq 1$, can be defined. The construction

$$N_t(\omega) = \#\{n \geq 1\colon Y_1(\omega) + \cdots + Y_n(\omega) \leq t\} = \sum_{n\geq 1} \mathbf{1}\{Y_1(\omega) + \cdots + Y_n(\omega) \leq t\}, \qquad t \geq 0,$$

then specifies all $N_t\colon \Omega \to \{0,1,2,\ldots\} \cup \{\infty\}$, $t \geq 0$, and these are random variables since sums and limits (series) of random variables are also random variables.

*Remark* H.48. The main body of the lecture notes has an informal proof that the increments in the setting of the definition based on independent inter-arrival times are independent. This is informal, because it involves conditioning infinitely many random variables on finitely many continuously distributed random variables. Rather than formalising this, we give a technically simpler alternative proof based on conditioning on the event $\{N_s = k\}$ of positive probability. We claim that conditionally given $\{N_s = k\}$, the post-$s$ process $\widetilde{N}_u = N_{s+u} - N_s$, $u \geq 0$, is a Poisson process of rate $\lambda$ in the sense that its inter-arrival times $\widetilde{Y}_n$, $n \geq 1$, which on $\{N_s = k\}$ satisfy

$$\widetilde{Y}_1 = Y_{k+1} - (s - T_k), \quad \widetilde{Y}_2 = Y_{k+2}, \widetilde{Y}_3 = Y_{k+3}, \ldots$$

are i.i.d. Exp($\lambda$). Since $\{N_s = k\} = \{T_k \leq s, T_k + Y_{k+1} > s\}$ and also $\{\widetilde{Y}_1 \leq z\}$ only depend on $Y_1, \ldots, Y_{k+1}$, the intersection of these events is independent of the random variables $\widetilde{Y}_n$, $n \geq 2$, and it suffices to show that the conditional distribution of $\widetilde{Y}_1$ given $\{N_s = k\}$ is also Exp($\lambda$). To this end, we calculate $\mathbb{P}(\widetilde{Y}_1 > z | N_s = k)$ as

$$\mathbb{P}(Y_{k+1} - (s - T_k) > z | T_k \leq s, T_k + Y_{k+1} > s) = \frac{\mathbb{P}(Y_{k+1} > (s - T_k) + z, T_k \leq s)}{\mathbb{P}(Y_{k+1} > (s - T_k) \quad , T_k \leq s)}.$$

Numerator and denominator are of the same form, and written as an integral, this gives

$$\mathbb{P}(Y_{k+1} > (s - T_k) + z, T_k \leq s) = \int_0^s \int_{s-t+z}^\infty f_{T_k}(t) f_{Y_{k+1}}(y) dy dt = e^{-\lambda s} \frac{\lambda^k}{(k-1)!} \frac{s^k}{k} e^{-\lambda z}.$$

After cancellation, the preceding equation further equals $e^{-\lambda z}$, as required to identify Exp($\lambda$) as the conditional distribution of $\widetilde{Y}_1$ given $\{N_s = k\}$. Since this conditional distribution does not depend on $k$, it is also the unconditional distributio, by the law of total probability

$$\mathbb{P}(\widetilde{Y}_1 > z) = \sum_{k=0}^\infty \mathbb{P}(\widetilde{Y}_1 > z | N_s = k) \mathbb{P}(N_s = k) = \sum_{k=0}^\infty e^{-\lambda z} \mathbb{P}(N_s = k) = e^{-\lambda z},$$

and so $\mathbb{P}(N_s = k, \widetilde{Y}_1 > z) = \mathbb{P}(N_s = k)\mathbb{P}(\widetilde{Y}_1 > z)$. Hence, $N_s$ and $\widetilde{Y}_1$ are independent and $\widetilde{N}$ is a Poisson process of rate $\lambda$ that is independent of $N_s$. An induction establishes independent Poisson increments.

*Remark* H.49. Our proof of the claim that a Poisson process in the sense of independent Poisson increments has independent Exp($\lambda$) inter-arrival times is not a formal proof either. Indeed, there is a non-trivial question about whether $Y_1 = \inf\{t > 0\colon N_t > 0\}$ is even a random variable in the sense that sets like $\{Y_1 < z\} = \{\omega \in \Omega\colon N_t > 0 \text{ for some } t \in (0,z)\} = \bigcup_{t\in(0,z)} \{N_t > 0\}$ are in the $\sigma$-algebra $\mathcal{F}$ of the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which $N_t$, $t \geq 0$, are defined. The subtlety is that $\{Y_1 < z\}$ depends on uncountably many random variables, and uncountable unions of events need not be events in $\mathcal{F}$.

If we further stipulate that $(N_t, t \geq 0)$ is a counting process with values in $\{0, 1, 2, \ldots\}$ that is increasing in the sense that $N_s \leq N_t$, the answer to this question is "yes" as it suffices to restrict the union to $t \in \mathbb{Q}$, and we then similarly find that $T_k$, $k \geq 1$, and hence $Y_k = T_k - T_{k-1}$, $k \geq 2$, are also random variables. Clearly, the existence of such a process has been proved by the construction from independent $\mathrm{Exp}(\lambda)$ inter-arrival times, but the further question that is being asked here is whether the independent Poisson increments uniquely identify the distribution of the inter-arrival times. In the setting of (increasing) counting processes, the answer is "yes," because we can re-express events such as

$$\{T_1 \leq s_1, T_2 \leq s_2, \ldots, T_k \leq s_k\} = \{N_{s_1} \geq 1, N_{s_2} \geq 2, \ldots, N_{s_k} \geq k\}.$$

so the joint distribution of any finite number of arrival times, or equivalently inter-arrival times, is uniquely identified by the joint distributions of $(N_t, t \geq 0)$ at finitely many times, or equivalently the increments between those times.

**Theorem H.50.** *Definition 7.4 via infinitesimal increments defines a rate-$\lambda$ Poisson process.*

*Proof.* In the setting where $(N_t, t \geq 0)$ is a counting process that satisfies the Poisson increments definition, $(N_t, t \geq 0)$ already satisfies conditions (i) and (ii) of Definition 7.4. For (iii), just note that the Poisson increments yield that the distribution of $N(s, s+h]$ does not depend on $s \geq 0$ and that

$$\begin{aligned}
&\mathbb{P}(N(s, s+h] = 0) = e^{-\lambda h} = 1 - \lambda h + o(h) \\
&\mathbb{P}(N(s, s+h] = 1) = \lambda h e^{-\lambda h} = \lambda h + o(h) \\
&\mathbb{P}(N(s, s+h] \geq 2) = 1 - \mathbb{P}(N(s, s+h] \leq 1) = 1 - e^{-\lambda h} - \lambda h e^{-\lambda h} = o(h).
\end{aligned}$$

It remains to prove that (i)–(iii) suffice to uniquely characterise the Poisson process of rate $\lambda$. Indeed, as the distribution of $N(s, s+h]$ does not depend on $s$, it suffices to identify the distribution of $N_t$ for all $t > 0$. We find from the i.i.d. increments that for all $k \in \{0, 1, 2, \ldots\}$,

$$\begin{aligned}
\mathbb{P}(N_t = k) &= \mathbb{P}\left(\sum_{i=1}^{n} N\left(\frac{i-1}{n}t, \frac{i}{n}t\right] = k\right) \\
&\geq \mathbb{P}\left(N\left(\frac{i-1}{n}t, \frac{i}{n}t\right] = 1 \text{ for } k \text{ values of } i \text{ and } N\left(\frac{i-1}{n}t, \frac{i}{n}t\right] = 0 \text{ otherwise}\right) \\
&= \binom{n}{k}\left(\lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right)^k \left(1 - \lambda\frac{t}{n} + o\left(\frac{t}{n}\right)\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{(\lambda t + o(1))^k}{k!}\left(1 - \frac{\lambda t + o(1)}{n}\right)^{n-k} \\
&\to \frac{(\lambda t)^k}{k!}e^{-\lambda t}.
\end{aligned}$$

This is a lower bound since we ignored the possibility that any of the $n$ intervals contributes 2 or more to $N_t$. We obtain an upper bound by adding to the above a term that includes all of these possibilities (and even drops the requirement to sum to $k$). Specifically, we add

$$\mathbb{P}\left(\bigcup_{i=1}^{n}\left\{N\left(\frac{i-1}{n}t, \frac{i}{n}t\right] \geq 2\right\}\right) \leq \sum_{i=1}^{n}\mathbb{P}\left(N\left(\frac{i-1}{n}t, \frac{i}{n}t\right] \geq 2\right) = n\, o\left(\frac{t}{n}\right) \to 0,$$

and a sandwich argument completes the proof. □

*Remark* H.51. In the statement of Theorem 7.6 on thinning, we mark points of a Poisson process independently with probability $p$. Formally, we can use an independent sequence of Bernoulli variables $I_k$, $k \geq 1$, such that $I_k$ is associated with the point at $T_k = Y_1 + \cdots + Y_k$, on a probability space with a sequence of independent uniform variables $U_m$, $m \geq 1$, where we use the even subsequence to construct $Y_k = -(\log(U_{2k}))/\lambda \sim \text{Exp}(\lambda)$ and the odd subsequence to construct the $I_k = \mathbf{1}\{U_{2k-1} \leq p\} \sim \text{Bernoulli}(p)$, $k \geq 1$.

This notation for Bernoulli variables works well to show (i)–(iii) in the Poisson increments definition. Specifically, for (ii), the events

$$E_j := \{N(t_{j-1}, t_j] = n_j, I_{n_1+\cdots+n_{j-1}+1} = i_{j,1}, \ldots, I_{n_1+\cdots+n_j} = i_{j,n_j}\}, \quad 1 \leq j \leq k,$$

are independent for all $n_j \in \{0, 1, 2, \ldots\}$ and $i_{j,1}, \ldots, i_{j,n_j} \in \{0, 1\}$, $1 \leq j \leq k$. Summing their probabilities over the $n_j$ and $i_{j,r}$, for fixed sums $i_{j,1} + \cdots + i_{j,n_j} = m_j$, $1 \leq j \leq k$, we conclude the independence of events $\{M(t_{j-1}, t_j] = m_j\}$, $1 \leq j \leq k$, as required.

In Remark 7.7 we claim that the process of unmarked points is an independent Poisson process. Indeed, the independence of events $\{M(t_{j-1}, t_j] = m_j, L(t_{j-1}, t_j] = \ell_j\}$ can be read as above, just also fixing $n_j = m_j + \ell_j$, $1 \leq j \leq k$. We then revisit the calculation for (iii) in the proof of Theorem 7.6 and find

$$\begin{aligned}
\mathbb{P}(M(0,t] = m, L(0,t] = \ell) &= \mathbb{P}\left(\sum_{r=1}^{N(0,t]} I_r = m \,\middle|\, N(0,t] = m + \ell\right) \mathbb{P}(N(0,t] = m + \ell) \\
&= \mathbb{P}\left(\sum_{r=1}^{m+\ell} I_r = m \,\middle|\, N(0,t] = m + \ell\right) \mathbb{P}(N(0,t] = m + \ell) \\
&= \mathbb{P}\left(\sum_{r=1}^{m+\ell} I_r = m\right) \mathbb{P}(N(0,t] = m + \ell) \\
&= \binom{m+\ell}{m} p^m (1-p)^\ell \frac{e^{-\lambda} \lambda^{m+\ell}}{(m+\ell)!} \\
&= \left(\frac{e^{-\lambda p}(p\lambda)^m}{m!}\right) \left(\frac{e^{-(1-p)\lambda}((1-p)\lambda)^\ell}{\ell!}\right).
\end{aligned}$$

**Proposition H.52.** *Let $B_r^{(n)}$, $r \geq 1$, be independent Bernoulli variables with success probability $\lambda/n$, for each $n \geq 1$. Define*

- *the discrete-time counting process $X_m^{(n)} = B_1^{(n)} + \cdots + B_m^{(n)}$, $m \geq 0$,*
- *the discrete-time arrival times $R_k^{(n)} = \inf\{m \geq 0 \colon X_m^{(n)} \geq k\}$, $k \geq 1$, and also $R_0^{(n)} = 0$,*
- *the discrete-time inter-arrival times $G_j^{(n)} = R_j^{(n)} - R_{j-1}^{(n)}$.*

*Then the increments $X_{m_j}^{(n)} - X_{m_{j-1}}^{(n)}$, $1 \leq j \leq k$, are independent* Binomial$(m_j - m_{j-1}, \lambda/n)$ *for all $0 = m_0 \leq m_1 \leq \cdots \leq m_k$ and $k \geq 1$. Also, the discrete-time inter-arrival times form a sequence of independent* Geometric$(\lambda/n)$ *random variables.*

*Proof.* This is elementary. □

**Corollary H.53.** *In the setting of Proposition H.52, consider continuous-time counting processes $N_t^{(n)} := X_{\lfloor nt \rfloor}^{(n)}$, $t \geq 0$, that incorporate $n$ discrete time steps per unit time, with*

*arrival times* $T_k^{(n)} := \inf\{t \geq 0 \colon N_t^{(n)} \geq k\} = R_k^{(n)}/n$, $k \geq 1$, *and inter-arrival times* $Y_j^{(n)} := T_j^{(n)} - T_{j-1}^{(n)} = G_j^{(n)}/n$, $j \geq 1$. *Then we have, as* $n \to \infty$,

$$\mathbb{P}(N^{(n)}(0,t_1] = m_1, \ldots, N^{(n)}(t_{k-1},t_k] = m_k) \to \mathbb{P}(N(0,t_1] = m_1, \ldots, N(t_{k-1},t_k] = m_k)$$

*and*

$$\mathbb{P}(Y_1^{(n)} \leq y_1, \ldots, Y_k^{(n)} \leq y_k) \to \mathbb{P}(Y_1 \leq y_1, \ldots, Y_k \leq y_k),$$

*where* $N$ *is a Poisson process of rate* $\lambda$ *with inter-arrival times* $Y_j$, $j \geq 1$.

*Proof.* Just note that the independence observations in Proposition H.52 and the independence properties of Poisson increments and inter-arrival times allow us to split the joint probabilities into products of probabilities, which each converge by the convergence theorems of Binomial to Poisson and Geometric to Exponential as noted in Section 7.4. □

We can view these as simple instances of multivariate convergence in distribution. Specifically, they can be seen as "finite-dimensional convergence" of the processes $N^{(n)}$ to the limiting process $N$. Other instances of multivariate convergence in distribution can be obtained by combining two applications of the central limit theorem to pairs of random variables that are independent within each pair, which comes up in Question 8 on Problem Sheet 2. It takes more development of this notion to deduce the convergence of probabilities that relate the two random variables in ways that do not factorise. While this was applied in an informal way in the problem sheet question, we refer to the Part C course on Limit Theorems and Large Deviations in Probability for relevant theory in the much higher generality of weak convergence of probability measures on metric spaces, which also allows to strengthen the convergence in Corollary H.53 to a convergence of distributions on a the space of right-continuous functions with left limits equipped with the Skorokhod topology.

# I

# Prelims Algebra: Relevant material

This appendix contains a selection of statements of definitions and theorems from the Prelims Linear Algebra courses that are relevant for Part A Probability. Most importantly, what we need are eigenvalues and eigenvectors of square matrices. Relevant for some examples are also some properties of permutations (borrowed from the Groups and Groups Actions course) and the notion of diagonalisability of a square matrix. Again for the sake of dealing with certain examples, we add material on recurrence relations that was essentially included in the Prelims Probability course. Specifically, we state results on solutions to recurrence relations for first and second order with constant coefficients, and we include a proof in the case of second order results, extrapolating slightly from the special cases covered in Prelims Probability.

## I.1 Linear Algebra I

**Definition I.1.** Let $m, n$ be positive integers. An $m \times n$ *matrix* is an array of real numbers arranged into $m$ rows and $n$ colums. *Row vectors* in $\mathbb{R}^n$ are $1 \times n$ matrices and *column vectors* in $\mathbb{R}^n_{\text{col}}$ are $n \times 1$ matrices.

**Definition I.2** (Matrix multiplication)**.** If $A = (a_{ij})$ is an $m \times n$ matrix and $B = (b_{jk})$ an $n \times p$ matrix, then the *product* $C = AB$ is the $m \times p$ matrix with entries

$$c_{ik} = \sum_{j=1}^{n} a_{ij} b_{jk}, \quad 1 \leq i \leq m \text{ and } 1 \leq k \leq p.$$

When $m = n = p$, we write $A^2$ for the product $AA$ and inductively $A^{q+1} = A^q A$ for $q \geq 2$. We also define $A^0 = I_n$, where $I_n = (\delta_{ij})$ is the identity matrix with entries $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. Note that $A^q A^r = A^{q+r}$ for all $q, r \geq 0$.

**Definition I.3.** We say that vectors $v_1, \ldots, v_m \in \mathbb{R}^n$ are *linearly independent* if the only solution to the equation

$$\alpha_1 v_1 + \cdots + \alpha_m v_m = 0, \qquad \text{where } \alpha_1, \ldots, \alpha_m \in \mathbb{R},$$

is $\alpha_1 = \cdots = \alpha_m = 0$.

## I.2 Linear Algebra II

Let $M_n(\mathbb{R})$ be the set of $n \times n$ matrices with real entries. For $A \in M_n(\mathbb{R})$ we write $A = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ where $\mathbf{a}_i \in \mathbb{R}^n_{\text{col}}$, $1 \le i \le n$, are the columns of $A$.

**Definition I.4.** A function $D\colon M_n(\mathbb{R}) \to \mathbb{R}$ is *determinantal* if it is

(a) multilinear in the columns:

$$D[\ldots, c\mathbf{a}_i + \mathbf{b}_i, \ldots] = cD[\ldots, \mathbf{a}_i, \ldots] + D[\ldots, \mathbf{b}_i, \ldots],$$

(b) alternating: $D[\ldots, \mathbf{a}_i, \mathbf{a}_{i+1}, \ldots] = 0$ if $\mathbf{a}_i = \mathbf{a}_{i+1}$,

(c) and $D(I_n) = 1$.

**Proposition I.5.** *Let $D\colon M_n(\mathbb{R}) \to \mathbb{R}$ be a determinantal map. Then*

*(a)* $D[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots] = -D[\ldots, \mathbf{a}_j, \ldots, \mathbf{a}_i, \ldots]$,

*(b) and* $D[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots] = 0$ *when* $\mathbf{a}_i = \mathbf{a}_j$, $i \neq j$.

**Definition I.6.** Let $S$ be a set. A bijection $S \to S$ is called a *permutation* of $S$ and the set of permutations of $S$ is denoted $\mathrm{Sym}(S)$. If $n$ is a positive integer, then we write $S_n$ for $\mathrm{Sym}(\{1, \ldots, n\})$. An element $\sigma \in S_n$ which switches two elements $1 \le i < j \le n$ and fixes the others is called a *transposition*.

**Theorem I.7.** *Let $S$ be a set.*

*(a) Then* $\mathrm{Sym}(S)$ *forms a group under composition. It is called the* symmetric group *of $S$.*

*(b) The cardinality of $S_n$ is $n!$.*

**Definition I.8.** A permutation is said to be *odd* (resp. *even*) if it can be written as a composition of an odd (resp. even) number of transpositions. We define the *sign* of $\sigma \in S_n$ as $\mathrm{sign}(\sigma) = 1$ if $\sigma$ is even and $\mathrm{sign}(\sigma) = -1$ if $\sigma$ is odd.

**Theorem I.9.** *(a) Every permutation can be written as a composition of transpositions.*

*(b) Every permutation is either even or odd, but not both.*

**Theorem I.10.** *For each $n \in \mathbb{N}$ there exists a unique determinantal function $D\colon M_n(\mathbb{R}) \to \mathbb{R}$ and it is explicitly given by*

$$D[\mathbf{a}_1, \ldots, \mathbf{a}_n] = \sum_{\sigma \in S_n} \mathrm{sign}(\sigma) a_{\sigma(1),1} \cdots a_{\sigma(n),n}.$$

**Definition I.11.** For an $A \in M_n(\mathbb{R})$, a column vector $v \in \mathbb{R}^n_{\text{col}}$ is a *right eigenvector* if $v \neq 0$ and $Av = cv$ for some $c \in \mathbb{R}$. We call $c \in \mathbb{R}$ an *eigenvalue* of $A$ if $Av = cv$ for some nonzero $v \in \mathbb{R}^n_{\text{col}}$. We call a row vector $\lambda \in \mathbb{R}^n$ a *left eigenvector* if $\lambda \neq 0$ and $\lambda A = c\lambda$.

**Definition I.12.** For $A \in M_n(\mathbb{R})$ the *characteristic polynomial* of $A$ is defined as $\chi_A(x) = \det(A - xI_n)$.

**Theorem I.13.** *Let $A \in M_n(\mathbb{R})$. Then $c$ is an eigenvalue of $A$ if and only if $c$ is a root of the characteristic polynomial $\chi_A(x)$ of $A$.*

**Theorem I.14.** *Let $c_1, \ldots, c_m$, for $m \leq n$, be the distinct eigenvalues of $A$ and $v_1, \ldots, v_m$ corresponding eigenvectors. Then $v_1, \ldots, v_m$ are linearly independent.*

**Definition I.15.** A matrix $A \in M_n(\mathbb{R})$ is *diagonalisable* if there is an invertible matrix $U$ such that $B := U^{-1}AU$ is a diagonal matrix.

**Proposition I.16.** *If $A \in M_n(\mathbb{R})$ is diagonalisable with diagonal matrix $B = U^{-1}AU$, then the diagonal entries of $B$ are the eigenvalues, and the colums in $U$ corresponding eigenvectors.*

**Corollary I.17.** *If $A \in M_n(\mathbb{R})$ has $n$ distinct eigenvalues, then $A$ is diagonalisable.*

**Corollary I.18.** *If $A \in M_n(\mathbb{R})$ is diagonalisable with diagonal matrix $B = U^{-1}AU$, then $A^n = UB^nU^{-1}$, where $B^n$ is the diagonal matrix whose entries are the $n$th powers of the corresponding entries of $B$, for all $n \geq 0$.*

## I.3 Recurence relations

**Definition I.19.** Let $k \geq 1$. A *$k$th order linear recurrence relation* (or *difference equation* ) has the form

$$\sum_{j=0}^{k} a_j u_{n+j} = f(n) \tag{I.1}$$

with $a_0 \neq 0$ and $a_k \neq 0$, where $a_0, \ldots, a_k$ are constants independent of $n$. A *solution* to (I.1) is a sequence $(u_n)_{n\geq 0}$ satisfying (I.1) for all $n \geq 0$. The equation (I.1) is called *homogeneous* if $f(n) = 0$ for all $n \geq 0$ and *inhomogeneous* otherwise.

**Theorem I.20.** *The* general solution $(u_n)$ *of* (I.1) *can be written as $u_n = v_n + w_n$ where $(v_n)$ is a particular solution to* (I.1) *and $(w_n)$ solves the associated homogeneous equation*

$$\sum_{j=0}^{k} a_j w_{n+j} = 0. \tag{I.2}$$

**Corollary I.21.** *For $k = 1$, the general solution to $u_{n+1} = au_n + c$ is given by*

$$u_n = Aa^n + \frac{c}{1-a} \quad \text{if } a \neq 1 \qquad \text{and} \qquad u_n = A + nc \quad \text{if } a = 1,$$

*for $n \geq 0$ and an arbitrary constant $A \in \mathbb{R}$.*

**Proposition I.22.** *For $k = 2$, the general solution to*

$$u_{n+1} = au_n + bu_{n-1} + c, \qquad n \geq 1,$$

*is given in terms of the solutions $\lambda_1$ and $\lambda_2$ of the auxiliary equation $\lambda^2 = a\lambda + b$, by*

$$\begin{aligned}
u_n &= A_1\lambda_1^n + A_2\lambda_2^n + \frac{c}{1-a-b} && \text{if } \lambda_1 \neq \lambda_2 \text{ and } a+b \neq 1,\\
u_n &= A_1\lambda_1^n + A_2\lambda_2^n + \frac{c}{1+b}n && \text{if } \lambda_1 \neq \lambda_2 \text{ and } a+b = 1,\\
u_n &= (A_1 + A_2 n)\lambda^n + \frac{c}{(\lambda-1)^2} && \text{if } \lambda_1 = \lambda_2 = \lambda \neq 1,\\
u_n &= A_1 + A_2 n + \frac{c}{2}n^2 && \text{if } \lambda_1 = \lambda_2 = 1.
\end{aligned}$$

*Proof.* Factorise the *auxiliary equation* $\lambda^2 = a\lambda + b$ into $\lambda^2 - a\lambda - b = (\lambda - \lambda_1)(\lambda - \lambda_2) = 0$ identifying two roots $\lambda_1, \lambda_2 \in \mathbb{C}$, not necessarily distinct. Then $\lambda_1 + \lambda_2 = a$ and $b = \lambda_1(\lambda_1 - a)$.

Now consider any (real) solution $(x_n)$ to the homogeneous equation $x_{n+1} = ax_n + bx_{n-1}$. Then $y_n = x_n - \lambda_1 x_{n-1}$ satisfies for $n \geq 1$

$$y_{n+1} = x_{n+1} - \lambda_1 x_n = (a - \lambda_1)x_n + bx_{n-1} = (a - \lambda_1)(x_n - \lambda_1 x_{n-1}) = \lambda_2 y_n.$$

Inductively, $y_{n+1} = B\lambda_2^n$ for some $B \in \mathbb{R}$. But then $(x_n)$ solves the first order equation

$$x_{n+1} = \lambda_1 x_n + B\lambda_2^n. \tag{I.3}$$

The general solution to the associated homogeneous equation $z_{n+1} = \lambda_1 z_n$ is $z_n = A_1\lambda_1^n$.

*Case 1:* $\lambda_1 \neq \lambda_2$. A particular solution of (I.3) is $v_n = y_{n+1}/(\lambda_2 - \lambda_1)$ since for $n \geq 1$

$$v_{n+1} = \frac{B}{\lambda_2 - \lambda_1}\lambda_2^{n+1} = \frac{B}{\lambda_2 - \lambda_1}(\lambda_1 + \lambda_2 - \lambda_1)\lambda_2^n = \lambda_1 v_n + B\lambda_2^n.$$

Let $A_2 = B/(\lambda_2 - \lambda_1)$. By Theorem I.20,

$$x_n = z_n + v_n = A_1\lambda_1^n + A_2\lambda_2^n$$

is the general solution of the inhomogeneous first-order equation (I.3), for each fixed $B = A_2(\lambda_2 - \lambda_1)$, and of the homogeneous second order equation $x_{n+1} = ax_n + bx_{n-1}$, for arbitrary constants $A_1, A_2 \in \mathbb{R}$.

We can find a constant particular solution $t_n = s$ of the inhomogeneous second order equation if $a + b \neq 1$ as then

$$at_n + bt_{n-1} + c = (a + b)s + c = s = t_{n+1} \quad \iff \quad s = \frac{c}{1 - a - b}.$$

If $a + b = 1$, then $\lambda_1 = 1$ and $\lambda_2 = -b \neq 1$, so there is a particular solution $t_n = rn$ as then

$$at_n + bt_{n-1} + c = arn + b(rn - r) + c = (a + b)rn + c - br = r(n + 1) \quad \iff \quad r = \frac{c}{1 + b}.$$

*Case 2:* $\lambda_1 = \lambda_2 = \lambda$. Then $a = 2\lambda$ and $b = -\lambda^2$. If $\lambda = 0$, this all degenerates to $u_0 = A_1$, $u_1 = A_2$ and $u_n = c$, $n \geq 2$. Now suppose that $\lambda \neq 0$. A particular solution of (I.3) is $v_n = ny_{n+1}/\lambda$ since then

$$av_n + bv_{n-1} = 2\lambda n y_{n+1}/\lambda - \lambda^2(n - 1)y_n/\lambda = B\lambda^n(2n - (n - 1)) = (n + 1)y_{n+2}/\lambda = v_{n+1}.$$

By Theorem I.20, the general solution of (I.3) can be written as $x_n = z_n + v_n = (A_1 + nA_2)\lambda^n$ for fixed $B = \lambda A_2$, and similarly for the general solution of $x_{n+1} = ax_n + bx_{n-1}$, now for arbitrary constants $A_1, A_2 \in \mathbb{R}$.

As in Case 1, a constant particular solution $t_n = c/(\lambda - 1)^2$ exists bar one exception $\lambda = 1$. When $\lambda = 1$, $t_n = cn^2/2$ is easily seen to be a particular solution. □