Lecture Notes on Probability

By

Omer Tamuz

California Institute of Technology

2018

Contents

Disclaimer

This a not a textbook. These are lecture notes.

1 Why we need measure theory

1.1 Probability measures on countable spaces

We usually think of a probability measure μ on a *countable* set of outcomes Ω as an assignment to each $\omega \in \Omega$ of a number between 0 and 1, with the property that these numbers sum to 1. In this course we will think about it as a function $\mu\colon 2^\Omega \to [0,1]$ (assigning to each subset of Ω a number in $[0,1]$) with the following two properties:

1. *Unit mass.* $\mu(\Omega) = 1$.
2. *sigma-additivity.* if $(A_1, A_2, \ldots)$ is a sequence of disjoint sets then

$$\mu(\cup_n A_n) = \sum_n \mu(A_n).$$

For example, let $\Omega = \mathbb{N} = \{1,2,3,\ldots\}$, and define μ by $\mu(\{n\}) = 2^{-n}$. This is the distribution of the number of tosses of a fair coin until the first heads. If E is the set of even numbers, then

$$\mu(E) = \sum_{n=1}^{\infty} 2^{-2n} = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} = \frac{1}{3}.$$

As another example, when $\Omega = \{0,1\}^n$, the i.i.d. fair coin toss measure can be defined by letting, for each $k \le n$ and $a \in \{0,1\}^k$

$$\mu(\{\omega : \omega_1 = a_1, \ldots, \omega_k = a_k\}) = 2^{-k}. \tag{1.1}$$

This measure has a rich symmetry structure. For $k \le n$, let $\sigma_k\colon \Omega \to \Omega$ be the map that "flips" the k^{th} entry:

$$\sigma_k(\omega_1, \ldots, \omega_n) = (\omega_1, \ldots, \omega_{k-1}, 1-\omega_k, \omega_{k+1}, \ldots, \omega_n).$$

Denote by $\sigma_k(A)$ the set $\{\sigma_k(\omega) : \omega \in A\}$. Then

$$\mu(\sigma_k(A)) = \mu(A) \text{ for all } k \text{ and } A \subseteq \Omega. \tag{1.2}$$

That is, if we flip the k^{th} entry in each element of a set of outcomes A, then we do not change its probability. In fact, it is easy to show that (1.2) implies (1.1), and thus can be taken to be the definition of μ.

1.2 Probability measures on uncountable spaces

We would like to define the same object for a countable number of coin tosses, which makes for an uncountable set of outcomes. That is, when $\Omega = \{0,1\}^{\mathbb{N}}$, we would like to define a map $\mu\colon 2^\Omega \to [0,1]$ that satisfies unit mass and countable additivity, and additionally satisfies (1.2). Here, $\sigma_k\colon \Omega \to \Omega$ flips the k^{th} entry in the infinite sequence $(\omega_1, \omega_2, \ldots) \in \Omega$.

It turns out that no such measure exists. The next section explains why.

1.3 Vitali sets

Say that $\theta \in \Omega$ is equivalent to $\omega \in \Omega$ if there is some N such that $\omega_n = \theta_n$ for all $n \geq N$. That is, if ω and θ agree in all but finitely many entries. Equivalently, θ is equivalent to ω if there is a finite set $\bar{k} = \{k_1, k_2, \ldots, k_n\} \subset \mathbb{N}$ such that $\theta = \sigma_{k_n} \circ \sigma_{k_{n-1}} \circ \cdots \circ \sigma_{k_1}(\omega)$. For example, θ is equivalent to $(0,0,0,\ldots)$ iff $\theta_k = 1$ for only finitely many k.

Note that if θ is equivalent to ω then ω is equivalent to θ. Note also that if ω is equivalent to θ, and if θ is equivalent to to ζ, then ω is equivalent to ζ. So our equivalence is indeed an equivalence relation. Therefore, if we denote equivalence classes by

$$[\omega] = \{\theta \text{ that is equivalent to } \omega\}$$

then $[\omega] = [\theta]$ iff ω and θ are equivalent, and the collection of equivalence classes forms a partition of Ω. Denote by K the (countable) set of finite subsets $\bar{k} = \{k_1, \ldots, k_n\} \subset \mathbb{N}$, and denote $\sigma_{\bar{k}} = \sigma_{k_n} \circ \cdots \circ \sigma_{k_1}$. Then

$$[\omega] = \{\sigma_{\bar{k}}(\omega) : \bar{k} \in K\},$$

and so each equivalence class is countable.

Let V be a set of representatives of these equivalence classes. That is, V contains for each equivalence class $[\omega]$ a single element of $[\omega]$. Then $\sigma_{\bar{k}}(V) \neq \sigma_{\bar{\ell}}(V)$ whenever $\bar{k} \neq \bar{\ell}$, and

$$\bigcup_{\bar{k} \in K} \sigma_{\bar{k}}(V) = \Omega.$$

Suppose by way of contradiction that there is a μ that has all of our desired properties. Then, by sigma-additivity and unit mass,

$$\mu\left(\cup_{\bar{k} \in K} \sigma_{\bar{k}}(V)\right) = \mu(\Omega) = 1.$$

Since $\mu(\sigma_{\bar{k}}(V)) = \mu(V)$, then the r.h.s. of the above equation is zero if $\mu(V) = 0$, and infinite if $\mu(V) > 0$. We have thus reached a contradiction, and no such μ exists.

2 Measure theory

2.1 π-systems

Given a set Ω, a *π-system* on Ω is a collection $\mathscr{P}$ of subsets of Ω such that if $A, B \in \mathscr{P}$ then $A \cap B \in \mathscr{P}$.

Example 2.1. *Let $\Omega = \mathbb{R}$, and let*

$$\mathscr{P} = \{(-\infty, x] : x \in \mathbb{R}\}.$$

This is a π-system because $(-\infty, x] \cap (-\infty, y] = (-\infty, \min\{x, y\}]$.

Example 2.2. *Let $\Omega = \{0,1\}^{\mathbb{N}}$, and let $\mathscr{P}$ be the collection of sets $\{A_S\}$ indexed by finite $S \subset \mathbb{N}$ where*

$$A_S = \{\omega \in \Omega : \omega_k = 1 \text{ for all } k \in S\}.$$

This is a π-system because $A_S \cap A_T = A_{S \cup T}$.

Example 2.3. *Let X be a topological space. Then the set of closed sets in X is a π-system.*

2.2 Algebras

An *algebra of subsets* of Ω is a π-system $\mathscr{A}$ on Ω with the following additional properties:

1. $\Omega \in \mathscr{A}$.
2. If $A \in \mathscr{A}$ then its complement $A^{\mathrm{c}} \in \mathscr{A}$.

It is easy to see that if $\mathscr{A}$ is an algebra of subsets of Ω then

1. $\emptyset \in \mathscr{A}$.
2. If $A, B \in \mathscr{A}$ then $A \cup B \in \mathscr{A}$.

Example 2.4. *Let Ω be any set. Then the collection of subsets of Ω is an algebra.*

Example 2.5. *Let $\Omega = \{0,1\}^{\mathbb{N}}$, and let $\mathscr{A}_{\text{clopen}}$ be the algebra of clopen sets. That is, $\mathscr{A}_{\text{clopen}}$ is the collection of finite unions of sets A_x indexed by finite $x \in \{0,1\}^n$, where*

$$A_x = \{\omega \in \Omega : \omega_k = x_k \text{ for all } k \le n\}.$$

Exercise 2.6. *Show that $\mathscr{A}_{\text{clopen}}$ is the collection of finite* disjoint *unions of sets of the form A_x.*

Example 2.7. *Let $\Omega = \mathbb{N}$, and let $\mathscr{A}_\infty$ be the collection of sets A such that either A is finite, or else A^{c} is finite.*

Exercise 2.8. *Prove that $\mathscr{A}_{\text{clopen}}$ and $\mathscr{A}_\infty$ are algebras.*

Given an algebra $\mathscr{A}$, a *finitely additive probability measure* is a function $\mu\colon \mathscr{A} \to [0,1]$ with the following properties:

1. $\mu(\Omega) = 1$.
2. μ is *additive*. That is, if A_1, A_2 are disjoint (i.e., $A_1 \cap A_2 = \emptyset$) then
$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2).$$

Exercise 2.9. *Show that $\mu(\emptyset) = 0$.*

Exercise 2.10. *Define a finitely additive measure on the algebra $\mathscr{A}_\infty$ from Example 2.7.*

2.3 Sigma-algebras

An algebra $\mathscr{F}$ of subsets of Ω is a *sigma-algebra* if for any sequence $(A_1, A_2, \ldots)$ of elements of $\mathscr{F}$ it holds that $\cup_n A_n \in \mathscr{F}$. It follows that $\cap_n A_n \in \mathscr{F}$.

Exercise 2.11. 1. *Let I be a set, and let $\{\mathscr{F}_i\}_{i\in I}$ be a collection of sigma-algebras of subsets of Ω. Show that $\cap_{i\in I}\mathscr{F}_i$ is a sigma-algebra.*

2. *Let $\mathscr{C}$ be a collection of subsets of Ω. Then there exists a unique minimal (under inclusion) sigma-algebra $\mathscr{F} \supseteq \mathscr{C}$. $\mathscr{F}$ is called the sigma-algebra generated by $\mathscr{C}$, which we write as $\mathscr{F} = \sigma(\mathscr{C})$.*

Exercise 2.12. *Prove that $\mathscr{A}_\infty$ (Example 2.7) is not a sigma-algebra.*

Given a topological space, the *Borel sigma-algebra* $\mathscr{B}$ is the sigma-algebra generated by the open sets. Hence it is also generated by any basis of the topology.

A *measurable space* is a pair $(\Omega, \mathscr{F})$, where $\mathscr{F}$ is a sigma-algebra of subsets of Ω. A *probability measure* on $(\Omega, \mathscr{F})$ is a function $\mu\colon \mathscr{F} \to [0,1]$ with the following properties:

1. $\mu(\Omega) = 1$.
2. μ is *sigma additive*. That is, if $(A_1, A_2, \ldots)$ is a sequence of disjoint sets (i.e., $A_n \cap A_m = \emptyset$ for all $n \neq m$) then
$$\mu(\cup_n A_n) = \sum_n \mu(A_n).$$

3 Constructing probability measures

3.1 The Hahn-Kolmogorov Theorem

Theorem 3.1 (Hahn-Kolmogorov Theorem)**.** *Let $\mathscr{C}$ be a collection of subsets of Ω, and let $\mathscr{F} = \sigma(\mathscr{C})$. Let $\mu_0 : \mathscr{C} \to [0,1]$ be a countably additive map with $\mu(\Omega) = 1$. We say that a probability measure $\mu : \mathscr{F} \to [0,1]$ extends μ_0 if $\mu(A) = \mu_0(A)$ for all $A \in \mathscr{C}$.*

1. *If $\mathscr{C}$ is a π-system then there exists* at most one *probability measure μ that extends μ_0.*
2. *If $\mathscr{C}$ is an algebra then there exists* exactly one *probability measure μ that extends μ_0.*

Example 3.2. *Let $\mathscr{A} = \mathscr{A}_{\text{clopen}}$ be the algebra defined in Example 2.5. Then there is a unique map $\mu_0 : \mathscr{A} \to [0,1]$ that is additive and satisfies*

$$\mu_0(A_x) = 2^{-|x|}.$$

Furthermore, this map is countably additive.

Hence μ_0 has a unique extension $\mu : \mathscr{B} \to [0,1]$ (where $\mathscr{B} = \sigma(\mathscr{A})$ is the Borel sigma-algebra on $\{0,1\}^{\mathbb{N}}$, equipped with the product topology).

The probability measure μ is sometimes called the *Bernoulli measure* on $\{0,1\}^{\mathbb{N}}$.

Exercise 3.3. *Prove that $\mu_0 : \mathscr{A}_{\text{clopen}} \to [0,1]$ is countably additive.*

Example 3.4. *Let $\mathscr{P}$ be the π-system on the interval $[0,1]$ given by*

$$\mathscr{P} = \{[0,x] : x \in [0,1]\},$$

and let and let $\mu_0 : \mathscr{P} \to [0,1]$ be given by $\mu_0([0,x]) = x$. Then there exists a probability measure $\mu : \mathscr{B} \to [0,1]$ (where $\mathscr{B} = \sigma(\mathscr{C})$ is the Borel sigma-algebra on $[0,1]$) that extends μ_0.

Note that indeed there always exists such a μ; it is called the Lebesgue measure. To prove this we naturally extend μ_0 to the algebra generated by $\mathscr{P}$, and then show that this extension is countably additive.

3.2 Basic properties of probability measures

Theorem 3.5. *Let $(\Omega, \mathscr{F}, \mu)$ be a probability space.*

1. *If $(F_1, F_2, \ldots)$ be a sequence of sets in $\mathscr{F}$ such that $F_n \subseteq F_{n+1}$ then*

$$\mu(\cup_n F_n) = \lim_n \mu(F_n).$$

2. *If $(F_1, F_2, \ldots)$ be a sequence of sets in $\mathscr{F}$ such that $F_n \supseteq F_{n+1}$ then*

$$\mu(\cap_n F_n) = \lim_n \mu(F_n).$$

Proof. 1. Let $G_1 = F_1$, and for $n > 1$ let $G_n = F_n \setminus F_{n-1}$. Then $\cup_{k=1}^{n} G_n = F_n$, $\cup_n F_n = \cup_n G_n$, and additionally the G_n's are disjoint. Hence

$$\mu(\cup_n F_n) = \mu(\cup_n G_n) = \sum_n \mu(G_n) = \lim_n \sum_{k=1}^{n} \mu(G_n) = \lim_n \mu\left(\cup_{k=1}^{n} G_n\right) = \lim_n \mu(F_n).$$

2. Left as an exercise.

□

Corollary 3.6. *Let $(\Omega, \mathscr{F}, \mu)$ be a probability space, and let $(F_1, F_2, \ldots)$ be a sequence of sets in $\mathscr{F}$.*

1. *If $\mu(F_n) = 0$ for all n then*

$$\mu(\cup_n F_n) = 0.$$

2. *If $\mu(F_n) = 1$ for all n then*

$$\mu(\cap_n F_n) = 1.$$

3.3 Cumulative distribution functions

Let $(\mathbb{R}, \mathscr{B}, \mu)$ be a probability space. The *cumulative distribution function* (c.d.f.) associated with μ is

$$F(x) = \mu\big((-\infty, x]\big).$$

Claim 3.7. *The following holds for any cumulative distribution function F:*

1. *F is monotone non-decreasing.*
2. $\sup_x F(x) = 1$.
3. $\inf_x F(x) = 0$.
4. *F is right-continuous.*

Proof. 1. For any $x > y$ by the additivity of μ we have that

$$F(y) = \mu\big((-\infty, y]\big) = \mu\big((-\infty, x] \cup (x, y]\big) = \mu\big((-\infty, x]\big) + \mu\big((x, y)\big) = F(x) + \mu((x, y]) \geq F(x).$$

2. Let $E_n = (-\infty, n)$. Then $E_n \subset E_{n+1}$ and $\cup_n E_n = \mathbb{R}$. Hence by Theorem 3.5

$$\lim_n F(n) = \lim_n \mu(E_n) = \mu(\mathbb{R}) = 1.$$

Since F is monotone non-decreasing it follows that $\sup_x F(x) = 1$.

3. The proof of this is similar to the proof that $\sup_x F(x) = 1$.

4. Fix some $x \in \mathbb{R}$. Let $E_n = (-\infty, x + \varepsilon_n)$, for any decreasing sequence ε_n of positive numbers that converges to zero. Then $E_{n+1} \subset E_n$ and $\cap_n E_n = (-\infty, x]$, and so by Theorem 3.5

$$\lim_n F(x + \varepsilon_n) = \lim_n \mu(E_n) = \lim_n \mu\big((-\infty, x]\big) = F(x).$$

□

Claim 3.8. *A probability measure μ on $(\mathbb{R}, \mathscr{B})$ is uniquely determined by its cumulative distribution function.*

Proof. Let $\mathscr{P}$ be the π-system from Example 2.1, and note that $\mathscr{B} = \sigma(\mathscr{P})$. Let $\mu_0 : \mathscr{P} \to [0,1]$ be given by

$$\mu_0\big((-\infty, x]\big) = F(x),$$

so that μ_0 is the restriction of μ to $\mathscr{P}$. Since μ is (trivially) countably additive, by the Hahn-Kolmogorov Theorem that there is at most one probability measure that extends μ_0 to $\mathscr{B}$. Thus μ is the unique measure with c.d.f. F. □

One can in fact show that for any F that satisfies the properties of Claim 3.7 is the cumulative distribution function of some μ. The proof uses the Hahn-Kolmogorov Theorem.

4 Events and random variables

4.1 Events

Given a measurable space $(\Omega, \mathscr{F})$, an *event* A is an element of $\mathscr{F}$. We sometimes call events *measurable sets*.

4.2 Sub-sigma-algebras

A *sub-sigma-algebra* of $\mathscr{F}$ is a subset of $\mathscr{F}$ that is also a sigma-algebra.

Given another measurable space $(\Theta, \mathscr{G})$, a function $f : \Omega \to \Theta$ is *measurable* if for all $A \in \mathscr{G}$ it holds that $f^{-1}(A) \in \mathscr{F}$.

Exercise 4.1. *Prove that f is measurable iff the collection*

$$\sigma(f) = \{f^{-1}(A) : A \in \mathscr{G}\} = f^{-1}(\mathscr{G}). \tag{4.1}$$

is a sub-sigma-algebra of $\mathscr{F}$.

Hence (assuming f is onto, otherwise restrict to its image), $f^{-1} : \mathscr{G} \to \sigma(f)$ is an isomorphism of sigma-algebras.

Fix a measurable space $(\Omega, \mathscr{F})$, and let f be a measurable function to some other measurable space. Given a sub-sigma-algebra $\mathscr{G} \subseteq \mathscr{F}$, we say that f is $\mathscr{G}$-measurable if $\sigma(f)$ is a sub-sigma-algebra of $\mathscr{G}$.

We say that a sigma-algebra $\mathscr{F}$ is *separable* if it generated by a countable subset. That is, if there exists some countable $\mathscr{C} \subset \mathscr{F}$ such that $\mathscr{F} = \sigma(\mathscr{C})$.

We say that $\mathscr{F}$ *separates points* if for all $\omega_1 \neq \omega_2$ there exists some $A \in \mathscr{F}$ such that $\omega_1 \in A$ and $\omega_2 \notin A$.

Theorem 4.2. *Let $(\Omega, \mathscr{F})$, $(\Theta_1, \mathscr{G}_1)$ and $(\Theta_2, \mathscr{G}_2)$ be measurable spaces with sigma-algebras that separate points. Let $f : \Omega \to \Theta_1$ and $g : \Omega \to \Theta_2$ be measurable functions. Then g is $\sigma(f)$-measurable iff there exists a measurable $h : \Theta_1 \to \Theta_2$ such that $g = h \circ f$.*

Exercise 4.3. *Prove for the case that $g = h \circ f$.*

Measurable functions to $(\mathbb{R}, \mathscr{B})$ will be of particular interest.

Claim 4.4. *Let $(\Omega, \mathscr{F})$ be a measurable space, and let $f : \Omega \to \mathbb{R}$. Then*

1. *If $\mathscr{C} \subset \mathscr{B}$ satisfies $\sigma(\mathscr{C}) = \mathscr{B}$, and if $f^{-1}(A) \in \mathscr{F}$ for all $A \in \mathscr{C}$ then f is measurable.*
2. *For each $x \in \mathbb{R}$ let $A_x \subset \Omega$ be given by $A_x = \{\omega : f(\omega) \leq x\}$. If each A_x is in $\mathscr{F}$ then f is measurable.*
3. *If Ω is a topological space with Borel sigma-algebra $\mathscr{F}$, and if f is continuous, then it is measurable.*

4. *If g is a measurable function from $(\mathbb{R},\mathscr{B})$ to itself and f is measurable then $g\circ f$ is measurable.*

Claim 4.5. *Let $(\Omega,\mathscr{F})$ be a measurable space, and let $\{f_n\}$ be a sequence of measurable functions to $(\mathbb{R},\mathscr{B})$ with $0\le f_n\le 1$ for all n. Then the following are measurable:*

1. $\inf_n f_n$.
2. $\liminf_n f_n$.
3. *The set $\{\omega : \lim_n f_n(\omega)$ exists $\}$.*

Claim 4.6. *The measurable functions $(\Omega,\mathscr{F})\to(\mathbb{R},\mathscr{B})$ are a vector space over the reals:*

1. *If f is measurable then λf is measurable, for all $\lambda\in\mathbb{R}$.*
2. *If f_1 and f_2 are measurable, then f_1+f_2 is measurable.*

4.3 Random variables

Given a probability space $(\Omega,\mathscr{F},\mu)$ and a measurable space $(\Theta,\mathscr{G})$, we say that two measurable functions $f,g\colon\Omega\to\Theta$ are equivalent if $\mu(\{\omega : f(\omega)=g(\omega)\})=1$. A *random variable* is an equivalence class of measurable functions. We will often consider the case that $(\Theta,\mathscr{G})=(\mathbb{R},\mathscr{B})$, in which case we will call X a *real random variable*. In fact, we will do this so often that we will often refer to real random variables as just "random variables".

A few notes:

1. Note we will often just think of random variables as measurable functions. We will say, for example, that a real random variable is non-negative, by which we will mean that there is a non-negative function in the equivalence class. We will also define random variables by just describing one element of the equivalence class.
2. It is easy to verify that sums, products, limits etc. of random variables are well defined, in the sense that (for example) the equivalence class of $f+g$ is equal to the equivalence class of $f'+g'$ whenever f and f' are equivalent and g and g' are equivalent.
3. We will want to verify that expressions of the form
$$\mu\big(\{X\in A\}\big)=\mu\big(\{w : X(\omega)\in A\}\big)$$
for a random variable X and measurable A are well defined, in the sense that they are independent of the choice of representative: $X(\omega)$ can be taken to mean $f(\omega)$, where f is any member of the equivalence class X.

Example 4.7. *Let $\Omega=\{0,1\}^{\mathbb{N}}$, and let $\mathbb{P}$ be the Bernoulli measure defined in Example 3.2. Define the random variable $X\colon\Omega\to\mathbb{R}$ by*
$$X(\omega)=\max\{n\in\mathbb{N} : \omega_k=0 \text{ for all } k\le n\}.$$
Note that X is not well defined at a single point in Ω, the all zeros sequence. We accordingly extend $\mathbb{R}$ to include ∞ (and $-\infty$) and assign $X(\omega)=\infty$ in this case.

Given a random variable $X : \Omega \to \Theta$, we define the *pushforward* measure $\nu = X_* \mu$ on $(\Theta, \mathscr{G})$ by

$$\nu(A) = \mu\left(X^{-1}(A)\right).$$

The measure ν is also called the *law* of X.

Exercise 4.8. *Calculate the cumulative distribution function of the random variable defined in Example 4.7.*

5 Independence and the Borel-Cantelli Lemmas

5.1 Independence

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $(\mathscr{F}_1, \mathscr{F}_2, \ldots)$ be sub-sigma-algebras. We say that these sigma-algebras are *independent* if for any $(A_1, A_2, \ldots)$ with $A_n \in \mathscr{F}_n$ and any finite sequence n_k it holds that

$$\mathbb{P}\left[\cap_k A_{n_k}\right] = \prod_k \mathbb{P}\left[A_{n_k}\right]. \tag{5.1}$$

We say that the random variables $(X_1, X_2, \ldots)$ are independent if $(\sigma(X_1), \sigma(X_2), \ldots)$ are independent.

We say that the events $(A_1, A_2, \ldots)$ are independent if their indicators functions $(\mathbb{1}_{\{A_1\}}, \mathbb{1}_{\{A_2\}}, \ldots)$ are independent. Note that $\sigma(\mathbb{1}_{\{A\}}) = \{\emptyset, A, A^c, \Omega\}$.

Claim 5.1. *Let the events $(A_1, A_2, \ldots)$ be independent. Then*

$$\mathbb{P}[\cap_n A_n] = \prod_n \mathbb{P}[A_n].$$

Proof. By independence we have that for any $m \in \mathbb{N}$

$$\mathbb{P}\left[\cap_{n=1}^m A_n\right] = \prod_{n=1}^m \mathbb{P}[A_n].$$

Denote $B_m = \cap_{n=1}^m A_n$. Then B_n is a decreasing sequence with $\cap_m B_m = \cap_n A_n$, and so by Theorem 3.5 we have that

$$\mathbb{P}[\cap_n A_n] = \mathbb{P}[\cap_m B_m] = \lim_m \mathbb{P}[B_m] = \lim_m \prod_{n=1}^m \mathbb{P}[A_n] = \prod_n \mathbb{P}[A_n].$$

□

It turns out that to prove independence it suffices to show (5.1) for generating π-systems.

5.2 A question about a sequence of random variables

Theorem 5.2. *Let $(X_1, X_2, \ldots)$ be a sequence of independent real random variables, each with the distribution $\mathbb{P}[X_n > x] = \mathrm{e}^{-x}$ when $x > 0$ and $\mathbb{P}[X_n > x] = 1$ when $x \le 0$. Let*

$$L = \limsup_n \frac{X_n}{\log n}.$$

Then $\mathbb{P}[L = 1] = 1$.

To prove this Theorem we will need the *Borel-Cantelli Lemmas*.

5.3 The Borel-Cantelli Lemmas

Lemma 5.3 (Borel-Cantelli Lemmas)**.** *Let* $(\Omega, \mathscr{F}, \mathbb{P})$ *be a probability space, and let* $(A_1, A_2, \ldots)$ *be a sequence of events.*

1. *If* $\sum_n \mathbb{P}[A_n] < \infty$ *then*

$$\mathbb{P}[\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n] = 0.$$

2. *If* $\sum_n \mathbb{P}[A_n] = \infty$ and $(A_1, A_2, \ldots)$ *are independent then*

$$\mathbb{P}[\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n] = 1.$$

To see why independence is needed for the second part, consider the case that all the events A_n are equal to some event A with $0 < \mathbb{P}[A] < 1$.

Proof of Lemma 5.3. 1. Note that

$$\{\omega : \omega \in A_n \text{ for infinitely many } n\} = \cap_n \cup_{m \geq n} A_m.$$

Let $B_n = \cup_{m \geq n} A_m$, so that we want to show that $\mathbb{P}[\cap_n B_n] = 0$. Note that B_n is a decreasing sequence (i.e., $B_{n+1} \subseteq B_n$) and therefore by Theorem 3.5 we have that

$$\mathbb{P}[\cap_n B_n] = \lim_n \mathbb{P}[B_n].$$

Since $B_n = \cup_{m \geq n} A_m$, we have that $\mathbb{P}[B_n] \leq \sum_{m \geq n} \mathbb{P}[A_m]$. But the latter converges to 0, and so we are done.

2. Note that

$$\begin{aligned} \{\omega : \omega \in A_n \text{ for infinitely many } n\}^c &= \{\omega : \omega \in A_n \text{ for finitely many } n\} \\ &= \{\omega : \omega \in A_n^c \text{ for all } n \text{ large enough}\} \\ &= \cup_n \cap_{m \geq n} A_m^c. \end{aligned}$$

We would hence like to show that $\mathbb{P}\left[\cup_n \cap_{m \geq n} A_m^c\right] = 0$.

Let $C_n = \cap_{m \geq n} A_m^c$. Then by independence and Claim 5.1 we have that

$$\mathbb{P}[C_n] = \mathbb{P}\left[\cap_{m \geq n} A_m^c\right] = \prod_{m \geq n} (1 - \mathbb{P}[A_m]).$$

Since $1 - x \leq e^{-x}$ this implies that

$$\mathbb{P}[C_n] \leq \exp\left(-\sum_{m \geq n} \mathbb{P}[A_m]\right) = 0.$$

Finally, by Corollary 3.6, $\mathbb{P}[\cap_n C_n] = 0$.

□

Proof of Theorem 5.2. Let A_n be the event that $X_n \geq \alpha \log n$. Then

$$\mathbb{P}[A_n] = n^{-\alpha},$$

and the events $(A_1, A_2, \ldots)$ are independent (exercise!). Also, note that

$$\sum_n \mathbb{P}[A_n] \begin{cases} = \infty & \text{if } \alpha \leq 1, \\ < \infty & \text{if } \alpha > 1. \end{cases}$$

Thus, from the Borel-Cantelli Lemmas it follows that

$$\mathbb{P}[X_n \geq \alpha \log n \text{ for infinitely many } n] = \begin{cases} 1 & \text{if } \alpha \leq 1, \\ 0 & \text{if } \alpha > 1. \end{cases}$$

Now, note that the event $\{L \geq \alpha\}$ is identical to the event

$$\cap_{m>0} \{X_n \geq (\alpha - 1/m) \log n \text{ for infinitely many } n\},$$

and so $\mathbb{P}[L \geq 1] = 1$, by Corollary 3.6. It also follows that $\mathbb{P}[L \geq 1 + 1/n] = 0$ for any $n > 0$, and so we have that $\mathbb{P}[L > 1] = 0$, again by Corollary 3.6. Hence $\mathbb{P}[L \leq 1] = 1$, and so $\mathbb{P}[L = 1] = 1$. □

6 The tail sigma-algebra

6.1 Motivating example

Consider a sequence of independent real random variables $(X_1, X_2, \ldots)$ such that there exists some $M \geq 0$ such that $\mathbb{P}[|X_n| \leq M] = 1$ for all n. That is, the sequence is uniformly bounded.

Define the random variables

$$Y_n = \frac{1}{n}\sum_{k=1}^{n} X_n \quad \text{and} \quad L = \limsup_n Y_n.$$

Claim 6.1. $\mathbb{P}[|L| \leq M] = 1$.

Proof. Clearly $\mathbb{P}[|Y_n| \leq M] = 1$. Hence $\mathbb{P}[|Y_n| \leq M \text{ for all } n] = 1$, and thus $\mathbb{P}[|L| \leq M] = 1$. □

Define the event $A = \{\lim_n Y_n \text{ exists }\}$.

Theorem 6.2. *There exists some $c \in [-M, M]$ such that $\mathbb{P}[L = c] = 1$, and $\mathbb{P}[A] \in \{0, 1\}$.*

An interesting observation is that L is independent of X_1. To see this, define

$$L' = \limsup_n \frac{1}{n}\sum_{k=2}^{n} X_n,$$

which is clearly independent of X_1. But

$$L = \limsup_n \frac{1}{n}\sum_{k=1}^{n} X_n = \limsup_n \frac{X_1}{n} + \frac{1}{n}\sum_{k=2}^{n} X_n = L'.$$

In fact, by the same argument, L is independent of $(X_1, X_2, \ldots, X_n)$ for *any* n. This makes L a *tail random variable*, as we now explain.

6.2 The tail sigma-algebra

For each $n \in \mathbb{N}$ define the sigma-algebra $\mathcal{T}_n$ by

$$\sigma(X_n, X_{n+1}, \ldots),$$

which is the smallest sigma-algebra that contains $(\sigma(X_n), \sigma(X_{n+1}), \ldots)$. Define the tail sigma-algebra by

$$\mathcal{T} = \cap_n \mathcal{T}_n.$$

A random variable is a *tail random variable* if it is $\mathcal{T}$-measurable.

Claim 6.3. *L is a tail random variable.*

Proof. Using a construction similar to the L' construction above, it is easy to see that for every n there exists a function f_n such that $L = f_n(X_n, X_{n+1}, \ldots)$. It follows that L is $\mathcal{T}_n$-measurable. Thus for every $A \in \sigma(A)$ it holds that $L^{-1}(A) \in \mathcal{T}_n$, for every n. Thus $A \in \cap_n \mathcal{T}_n = \mathcal{T}$. □

Let $(Z_1, Z_2, \ldots)$ be i.i.d random variables, each distributed uniformly over the set of symbols $S = \{a, b, c\}$. Let S^* be the set of finite strings over S, and define the random variable W_n taking values in S^* as follows:

- $W_1 = Z_1$.
- If W_n is empty, or if the last symbol in W_n is different than Z_{n+1}, then W_{n+1} is the concatenation $W_n Z_{n+1}$.
- If the last symbol in W_n is Z_{n+1} then W_{n+1} is equal to W_n, with this last symbol removed.

We will prove later in the course that with probability one it holds that $\lim_n |W_n| = \infty$, and hence we can define the random variable T to be the eventual first symbol in all W_n high enough. It is immediate that T is measurable in the tail sigma-algebra of the sequence $(W_1, W_2, \ldots)$.

It is also easy to see that $\mathbb{P}[T = a] = 1/3$, since by the symmetry of the definitions, $\mathbb{P}[T = a] = \mathbb{P}[T = b] = \mathbb{P}[T = c]$, and these must sum to one. By the same argument, the probability that W_n starts with some string w for all n high enough is $3^{-1} \cdot 2^{-|w|}$.

6.3 The zero-one law

Theorem 6.4 (Kolmogorov's Zero-One Law)**.** *Let $\mathcal{T}$ be the tail sigma-algebra of a sequence of independent random variables. Then $\mathbb{P}[A] \in \{0, 1\}$ for any $A \in \mathcal{T}$.*

Before proving this theorem we will prove a lemma.

Lemma 6.5. *Let the event A be independent of itself. Then $\mathbb{P}[A] \in \{0, 1\}$.*

Proof. $\mathbb{P}[A] = \mathbb{P}[A \cap A] = \mathbb{P}[A] \cdot \mathbb{P}[A]$. □

Proof of Theorem 6.4. Let $\mathcal{G}_n = \sigma(X_1, \ldots, X_{n-1})$, $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \ldots)$ and $\mathcal{T} = \cap_n \mathcal{T}_n$. We first claim that $\mathcal{G}_n$ and $\mathcal{T}_n$ are independent. To see this, define $\mathcal{T}_n^m = \sigma(X_n, \ldots, X_{n+m})$, and note that $\mathcal{T}_n^m$ and $\mathcal{G}_n$ are independent, and so $\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$ for any $A \in \mathcal{G}_n$ and any $B \in \mathcal{T}_n^m$. Now $\mathcal{C}_n = \cup_m \mathcal{T}_n^m$ is not a sigma-algebra, but it is a π-system. Since $\mathbb{P}[A \cap B] = \mathbb{P}[A] \cdot \mathbb{P}[B]$ for any $A \in \mathcal{G}_n$ and any $B \in \mathcal{C}_n$, it follows that $\mathcal{G}_n$ and $\sigma(\mathcal{C}_n) = \mathcal{T}_n$ are independent.

Since $\mathcal{T} \subset \mathcal{T}_n$ then $\mathcal{G}_n$ and $\mathcal{T}$ are independent. Hence $\mathcal{T}$ is independent of

$$\sigma(\cup_n \mathcal{G}_n) = \sigma(\cup_n \sigma(X_n)) = \sigma(X_1, X_2, \ldots).$$

Since $\mathcal{T} \subset \sigma(X_1, X_2, \ldots)$ it follows that $\mathcal{T}$ is independent of $\mathcal{T}$, and so $\mathbb{P}[A] \in \{0, 1\}$ for any $A \in \mathcal{T}$. □

Proof of Theorem 6.2. Since A is a tail random variable then $\mathbb{P}[A] \in \{0,1\}$.

For any $q \in \mathbb{Q}$ define the tail event $A_q = \{L \geq q\}$. By Kolmogorov's zero-one, law, the probability of each of these is either 0 or 1, and so there is some

$$c = \sup\{q : \mathbb{P}\left[A_q\right] = 1\} = \inf\{q : \mathbb{P}\left[A_q\right] = 0\}.$$

Since $\mathbb{Q}$ is countable, $\mathbb{P}[L \geq c] = \mathbb{P}[L \leq c] = 1$, and so $\mathbb{P}[L = c] = 1$. Finally, $c \in [-M, M]$, since $\mathbb{P}[L \in [-M, M]] = 1$. □

7 Expectations

7.1 Expectations in finite probability spaces

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a *finite* probability space, and let $f : \Omega \to \mathbb{R}$ be any function. The expectation of f is given by

$$\mathbb{E}[f] = \sum_{\omega \in \Omega} f(\omega)\mathbb{P}[\omega].$$

Another way of writing this is the following:

$$\mathbb{E}[f] = \sum_{x \in \mathrm{Im} f} x\mathbb{P}\left[f^{-1}(x)\right].$$

Note that this formulation does not reference points in Ω. Relatedly, it has the advantage that it naturally extends to any probability space, given that f has a finite (or countable) image. This is the basic idea that is behind our general definition of expectation.

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We say that a measurable non-negative $\tilde{f}$ is *simple* if it has a finite image,[1] and define its expectation $\mathbb{E}\left[\tilde{f}\right]$ by

$$\mathbb{E}\left[\tilde{f}\right] = \sum_{x \in \mathrm{Im} \tilde{f}} x\mathbb{P}\left[\tilde{f}^{-1}(x)\right].$$

7.2 Expectations of non-negative random variables

Given a (non-simple) non-negative real function f, we define its expectation by

$$\mathbb{E}[f] = \sup\left\{\mathbb{E}\left[\tilde{f}\right] : \tilde{f} \text{ is simple and } \tilde{f} \leq f\right\}.$$

Note that this supremum may be infinite.

It is straightforward to verify that for any non-negative functions f, g such that $\mathbb{P}[f = g] = 1$ it holds that $\mathbb{E}[f] = \mathbb{E}[g]$. We can therefore define the expectation of a *random variable* X as the expectation of any f in the equivalence class. We will henceforth consider expectations of random variables.

It is likewise straightforward to verify that for any two non-negative random variables X, Y:

- *Linearity of expectation:* For any $\lambda > 0$ it holds that

$$\mathbb{E}[X + \lambda Y] = \mathbb{E}[X] + \lambda\mathbb{E}[Y].$$

- If $X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

[1] Note that this definition is slightly different than the standard one.

7.3 Markov's inequality

Theorem 7.1 (Markov's Inequality)**.** *If X is a non-negative random variable with $\mathbb{E}[X] < \infty$ then for every $\lambda > 0$*

$$\mathbb{P}[X \geq \lambda] \leq \frac{\mathbb{E}[X]}{\lambda}.$$

Proof. Let $A = \{X \geq \lambda\}$, and let Y be given by

$$Y(\omega) = \lambda \cdot \mathbb{1}_{\{A\}}(\omega) = \begin{cases} \lambda & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Then $Y \leq X$, and so $\mathbb{E}[Y] \leq \mathbb{E}[X]$. Since $\mathbb{E}[Y] = \lambda \cdot \mathbb{P}[A]$, we get that

$$\lambda \cdot \mathbb{P}[X \geq \lambda] \leq \mathbb{E}[X],$$

and the claim follows by dividing both sides by λ. □

7.4 Pointwise convergence and convergence of expectations

Consider the non-negative random variables $(X_1, X_2, \ldots)$ defined on the interval $(0,1]$ (equipped with the Borel sigma-algebra and Lebesgue measure) which are given by

$$X_n(x) = \begin{cases} n & \text{if } x \leq 1/n, \\ 0 & \text{otherwise.} \end{cases}$$

Then

1. $\mathbb{E}[X_n] = 1$.
2. For every $x \in (0,1]$ it holds that $\lim_n X_n(x) = X(x)$, where X is the constant function $X(x) = 0$.
3. $\lim_n \mathbb{E}[X_n] \neq \mathbb{E}[X]$.

Hence it is not necessarily true that if $X_n \to X$ pointwise then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.

Theorem 7.2 (Monotone Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(X_1, X_2, \ldots)$ be a sequence of non-negative random variables such that $X_n(\omega)$ is increasing for every $\omega \in \Omega$. Let $X(\omega) = \lim_n X_n(\omega) \in [0, \infty]$. Then*

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X] \in [0, \infty].$$

Theorem 7.3 (Dominated Convergence Theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(X_1, X_2, \ldots)$ be a sequence of non-negative random variables. Let X, Y be a non-negative random variables with $\mathbb{E}[Y] < \infty$, and such that $\lim_n X_n(\omega) = X(\omega)$ for every $\omega \in \Omega$, and $X_n(\omega) \leq Y(\omega)$ for every $\omega \in \Omega$ and $n \in \mathbb{N}$. Then*

$$\lim_n \mathbb{E}[X_n] = \mathbb{E}[X].$$

7.5 $\mathscr{L}^p$

Given a random variable X, we define the random variables X^+ and X^- by

$$X^+(\omega) = \max\{X(\omega), 0\} \quad \text{and} \quad X^-(\omega) = \max\{-X(\omega), 0\},$$

so that X^+ and X^- are both non-negative, and $X = X^+ - X^-$. If $\mathbb{E}\left[X^+\right]$ and $\mathbb{E}[X^-]$ are both finite, we define

$$\mathbb{E}[X] = \mathbb{E}\left[X^+\right] - \mathbb{E}[X^-],$$

and say that $X \in \mathscr{L}^1(\Omega, \mathscr{F}, \mathbb{P})$, or just $X \in \mathscr{L}^1$. Note that $X \in \mathscr{L}^1$ iff $\mathbb{E}[|X|] < \infty$ iff $|X| \in \mathscr{L}^1$. For $p \geq 1$ we say that $X \in \mathscr{L}^p$ if $|X|^p \in \mathscr{L}^1$.

Exercise 7.4. *Show that $\mathscr{L}^p$ is a vector space.*

$X \mapsto \mathbb{E}[|X|^p]^{1/p}$ defines a norm on $\mathscr{L}^p$.

Theorem 7.5. *If $r > p \geq 1$ and $X \in \mathscr{L}^r$ then $X \in \mathscr{L}^p$ and moreover*

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq \mathbb{E}\left[|X|^r\right]^{1/r}.$$

In fact, if we equip $\mathscr{L}^p$ with this norm, then it is a *Banach space*; that is, it is complete with respect to the metric induced by this norm.

Theorem 7.6. *Let $(X_1, X_2, \ldots)$ be a sequence of random variables in $\mathscr{L}^p$ such that*

$$\lim_r \sup_{m,n \geq r} \{\mathbb{E}\left[|X_n - X_m|^p\right]\} = 0.$$

Then there exists an $X \in \mathscr{L}^p$ such that

$$\lim_n \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

7.6 $\mathscr{L}^2$

A particularly interesting case is $p = 2$. In this case we can define an *inner product* $(X, Y) := \mathbb{E}[X \cdot Y]$, which makes $\mathscr{L}^2$ a Hilbert space, with completeness given by Theorem 7.6.

Theorem 7.7. *Let $X, Y \in \mathscr{L}^2$. Then $X \cdot Y \in \mathscr{L}^1$.*

Proof. Note first that $|X|, |Y| \in \mathscr{L}^2$. Since $\mathscr{L}^2$ is a vector space then $\mathbb{E}\left[(|X| + |Y|)^2\right] < \infty$, and so

$$\mathbb{E}\left[X^2 + 2|X| \cdot |Y| + Y^2\right] < \infty.$$

By the linearity of expectation

$$\mathbb{E}\left[(|X| + |Y|)^2\right] = \mathbb{E}\left[X^2\right] + 2 \cdot \mathbb{E}[|X| \cdot |Y|] + \mathbb{E}\left[Y^2\right],$$

and so we have that $\mathbb{E}[|X| \cdot |Y|] < \infty$. Now,

$$\mathbb{E}[|X \cdot Y|] = \mathbb{E}[|X| \cdot |Y|],$$

and so $|X \cdot Y| \in \mathscr{L}^1$. Hence $X \cdot Y \in \mathscr{L}^1$. □

It follows from Theorems 7.6 and 7.7 that $\mathscr{L}^2$ is a real Hilbert space, when equipped with the inner product $(X,Y) := \mathbb{E}[X \cdot Y]$. We can therefore immediately conclude that for any $X, Y \in \mathscr{L}^2$

1. $\mathbb{E}[X \cdot Y]^2 \leq \mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]$, with equality iff for some $\lambda \in \mathbb{R}$ it a.s. holds that $X = \lambda \cdot Y$. This is the Cauchy-Schwartz inequality.

2. $\mathbb{E}\left[(X+Y)^2\right] = \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right]$ iff $\mathbb{E}[X \cdot Y] = 0$.

Given $X \in \mathscr{L}^2$, we define the random variable $\tilde{X} := X - \mathbb{E}[X]$, and denote $\mathrm{Var}(X) = \mathbb{E}\left[\tilde{X} \cdot \tilde{X}\right]$ and $\mathrm{Cov}(X,Y) = \mathbb{E}\left[\tilde{X} \cdot \tilde{Y}\right]$. We say that X and Y are uncorrelated if $\mathrm{Cov}(X,Y) = 0$. Using these definitions the facts above become

1. $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X) \cdot \mathrm{Var}(Y)$, with equality iff for some $\lambda \in \mathbb{R}$ it a.s. holds that $X = \lambda \cdot Y$.

2. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ iff X and Y are uncorrelated.

8 A strong law of large numbers and the Chernoff bound

8.1 Expectation of a product of independent random variables

Theorem 8.1. *Let $X, Y \in \mathscr{L}^1$ be independent. Then $X \cdot Y \in \mathscr{L}^1$ and*

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

To prove this, we first note that it holds for indicator functions by the definition of independence, then show that it holds for simple functions, and apply the monotone convergence theorem to show that it holds in general.

8.2 Jensen's inequality

Theorem 8.2 (Jensen's Inequality)**.** *Let X be a real random variable with $\mathbb{E}[X] = x_0 \in \mathbb{R}$. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then $\mathbb{E}\big[\varphi(X)\big] \geq \varphi(\mathbb{E}[X])$.*

Proof. Since φ is convex we can find $a, b \in \mathbb{R}$ such that $\varphi(x) \geq ax + b$ for all $x \in \mathbb{R}$, and $\varphi(x_0) = ax_0 + b$. Hence

$$\mathbb{E}\big[\varphi(X)\big] \geq \mathbb{E}[aX + b] = a\mathbb{E}[x] + b = ax_0 + b = \varphi(x_0) = \varphi(\mathbb{E}[X]).$$

□

8.3 SLLN in $\mathscr{L}^4$

Theorem 8.3. *Let $(X_1, X_2, \ldots)$ be a sequence of independent random variables uniformly bounded in $\mathscr{L}^4$ (so that $\mathbb{E}\big[X_n^4\big] < K$ for all n and some $K > 0$), and with $\mathbb{E}[X_n] = 0$. Let*

$$Y_n = \frac{1}{n} \sum_{k \leq n} X_n.$$

Then $\lim_n Y_n = 0$ a.s.

Proof. By independence

$$\mathbb{E}\big[X_k \cdot X_\ell^3\big] = \mathbb{E}\big[X_k \cdot X_\ell^2 \cdot X_m\big] = 0,$$

and so, by linearity we have that

$$\mathbb{E}\big[Y_n^4\big] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^4\right] = \frac{1}{n^4}\sum_{k \leq n} \mathbb{E}\big[X_k^4\big] + \frac{6}{n^4}\sum_{k < \ell \leq n} \mathbb{E}\big[X_k^2 \cdot X_\ell^2\big].$$

Again applying independence we get

$$\mathbb{E}\big[Y_n^4\big] = \frac{1}{n^4}\sum_{k \leq n} \mathbb{E}\big[X_k^4\big] + \frac{6}{n^4}\sum_{k < \ell \leq n} \mathbb{E}\big[X_k^2\big] \cdot \mathbb{E}\big[X_\ell^2\big].$$

By Jensen's inequality $\mathbb{E}\left[X_k^2\right]^2 < K$, and so

$$\mathbb{E}\left[Y_n^4\right] \leq \frac{K}{n^3} + \frac{6K}{n^2} \leq \frac{7K}{n^2}.$$

It follows from Markov's inequality that for any $\varepsilon > 0$

$$\mathbb{P}\left[Y_n^4 \geq \varepsilon^4\right] \leq \frac{7K}{\varepsilon^4 n^2},$$

and so, by Borel-Cantelli, $\limsup_n |Y_n| \leq \varepsilon$ for any $\varepsilon > 0$. Intersecting these probability one events for $\varepsilon = 1/2, 1/3, 1/4, \ldots$ yields that $\limsup_n |Y_n| = 0$ and thus $\lim_n Y_n = 0$. □

8.4 The Chernoff bound

With a little additional effort we can prove that if $\mathbb{E}[X_n] = \mu$ then $\lim_n Y_n = \mu$. A natural question is: what is the probability that Y_n is significantly far from μ, for finite n? For example, for $\eta > \mu$, what is the probability that $Y_n \geq \eta$?

Theorem 8.4 (Chernoff Bound)**.** *Let $(X_1, X_2, \ldots)$ be a sequence of i.i.d. random variables in $\mathscr{L}^\infty$, and with $\mathbb{E}[X_n] = \mu$. Then for every $\eta > \mu$ there is an $r > 0$ such that*

$$\mathbb{P}\left[Y_n \geq \eta\right] \leq \mathrm{e}^{-r \cdot n}.$$

Proof. Denote $p_n = \mathbb{P}\left[Y_n \geq \eta\right]$; we want to show that $p_n \leq \mathrm{e}^{-r \cdot n}$.

Note that the event $\{Y_n \geq \eta\}$ is identical to the event $\{\mathrm{e}^{t \cdot n \cdot Y_n} \geq \mathrm{e}^{t \cdot n \cdot \eta}\}$, for any $t > 0$. Since $\mathrm{e}^{t \cdot n \cdot Y_n}$ is a positive random variable, by the Markov inequality we have that

$$p_n = \mathbb{P}\left[\mathrm{e}^{t \cdot n \cdot Y_n} \geq \mathrm{e}^{t \cdot n \cdot \eta}\right] \leq \frac{\mathbb{E}\left[\mathrm{e}^{t \cdot n \cdot Y_n}\right]}{\mathrm{e}^{t \cdot n \cdot \eta}}.$$

Now,

$$\mathbb{E}\left[\mathrm{e}^{t \cdot n \cdot Y_n}\right] = \mathbb{E}\left[\prod_{k \leq n} \mathrm{e}^{t \cdot X_k}\right] = \prod_{k \leq n} \mathbb{E}\left[\mathrm{e}^{t \cdot X_k}\right],$$

where the penultimate equality uses independence. Let X be a random variable with the same distribution as each X_k. Then we have shown that

$$\mathbb{E}\left[\mathrm{e}^{t \cdot n \cdot Y_n}\right] = \mathbb{E}\left[\mathrm{e}^{t \cdot X}\right]^n.$$

We now define the *moment generating function of X* by $M_X(t) := \mathbb{E}\left[\mathrm{e}^{tX}\right]$. The name comes from the fact that

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}\left[X^n\right]. \tag{8.1}$$

Note that this means that $M'_X(0) = \mathbb{E}[X]$. Using M_X we can write

$$p_n \leq \exp\left(-(t \cdot \eta - \log M_x(t)) \cdot n\right)$$

If we define the *cumulant generating function of* X by $K_X(t) := \log M_X(t)$, then

$$p_n \leq \exp\left(-(t \cdot \eta - K_X(t)) \cdot n\right).$$

Since $K'_X(0) = M'_X(0)/M_X(0) = \mathbb{E}[X]$, and since K_X is smooth (as it turns out), it follows that for $t > 0$ small enough,

$$t \cdot \eta - K_X(t) = t \cdot \eta - t \cdot \mu - O(t^2) > 0.$$

Hence, if we define

$$r = \sup_t \{t \cdot \eta - K_X(t)\}$$

we get that $r > 0$ and

$$p_n \leq \mathrm{e}^{-r \cdot n}.$$

□

It turns out that the Chernoff bound is asymptotically tight. We show this in §21.

9 The weak law of large numbers

9.1 $\mathscr{L}^2$

Theorem 9.1. *Let $(X_1, X_2, \ldots)$ be a sequence of independent real random variables in $\mathscr{L}^2$, let $\mathbb{E}[X_n] = \mu$, $\mathrm{Var}(X_n) \leq \sigma^2$, and let $Y_n = \frac{1}{n}\sum_{k\leq n} X_n$. Then for every $\varepsilon > 0$ and $n \in \mathbb{N}$*

$$\mathbb{P}\left[|Y_n - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon},$$

and in particular

$$\lim_n \mathbb{P}\left[|Y_n - \mu| \geq \varepsilon\right] = 0.$$

In this case we say that Y_n *converges in probability* to μ. More generally, we say that a sequence of real random variables Y_n converges in probability to a real random variable Y if

$$\lim_n \mathbb{P}[|Y_n - Y| \geq \varepsilon] = 0.$$

Exercise 9.2. *Does convergence in probability imply pointwise convergence? Does pointwise convergence imply convergence in probability?*

To prove this Theorem we will need *Chebyshev's inequality*, which is just Markov's inequality in disguise.

Lemma 9.3 (Chebyshev's Inequality)**.** *For every $X \in \mathscr{L}^2$ and for every $\lambda > 0$ it holds that*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \lambda] \leq \frac{\mathrm{Var}(X)}{\lambda^2}.$$

Proof of Theorem 9.1. Note that $\mathbb{E}[Y_n] = \mu$, and that, by independence,

$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{k\leq n} X_k\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{k\leq n} X_k\right) = \frac{1}{n^2}\sum_{k\leq n}\mathrm{Var}(X_k) \leq \frac{\sigma^2}{n}.$$

Hence Chebyshev's inequality yields that for every $\lambda > 0$ we have that

$$\mathbb{P}\left[|Y_n - \mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon}$$

□

9.2 $\mathscr{L}^1$

We can relax the assumption $X \in \mathscr{L}^2$ to $X \in \mathscr{L}^1$ and still prove the weak law of large numbers. In fact, even the strong law holds in this setting (for i.i.d. random variables), but we will leave the proof of that for after we prove the Ergodic Theorem.

Theorem 9.4. *Let $(X_1, X_2, \ldots)$ be a sequence of i.i.d. real random variables in $\mathscr{L}^1$, let $\mathbb{E}[X_n] = \mu$, and let $Y_n = \frac{1}{n}\sum_{k\le n} X_n$. Then for every $\varepsilon > 0$*

$$\lim_n \mathbb{P}\left[|Y_n - \mu| \ge \varepsilon\right] = 0.$$

Proof. We assume $\mu = 0$; the reduction is straightforward.

Let $X = X_1$. For $N \in \mathbb{N}$, and a r.v. X denote

$$X^{\le N} = X \cdot \mathbb{1}_{\{|X| \le N\}} \quad \text{and} \quad X^{>N} = X \cdot \mathbb{1}_{\{|X| > N\}},$$

so that $X = X^{\le N} + X^{>N}$. By the Dominated Convergence Theorem

$$\mathbb{E}\left[|X^{>N}|\right] \to 0 \quad \text{and} \quad \mathbb{E}\left[X^{\le N}\right] \to \mathbb{E}[X] = 0, \tag{9.1}$$

since both are dominated by $|X|$.

Fix $\varepsilon, \delta > 0$. To prove the claim (under our assumption that $\mu = 0$) we show that $\mathbb{P}[|Y_n| \ge \varepsilon] < \delta$ for all n large enough. For any $N \in \mathbb{N}$ we can write Y_n as

$$Y_n = \frac{1}{n}\sum_{k\le n} X_k^{\le N} + X_k^{>N} = Y_n^{\le} + Y_n^{>},$$

where

$$Y_n^{\le} := \frac{1}{n}\sum_{k\le n} X_k^{\le N} \quad \text{and} \quad Y_n^{>} = \frac{1}{n}\sum_{k\le n} X_k^{>N}.$$

Note that $Y_n^{\le}$ is not the same as $Y_n^{\le N}$; we will not need the latter. Likewise, $Y_n^{>}$ is not the same as $Y_n^{>N}$.

Choose N large enough so that $\mathbb{E}\left[|X^{>N}|\right] < \varepsilon \cdot \delta/4$; this is possible by (9.1). Now,

$$\mathbb{E}\left[|Y_n^{>}|\right] = \mathbb{E}\left[\frac{1}{n}\left|\sum_{k\le n} X_k^{>N}\right|\right] \le \mathbb{E}\left[\frac{1}{n}\sum_{k\le n}\left|X_k^{>N}\right|\right] = \mathbb{E}\left[\left|X^{>N}\right|\right] < \varepsilon\cdot\delta/4.$$

Therefore, by Markov's inequality, we have that

$$\mathbb{P}\left[|Y_n^{>}| \ge \varepsilon/2\right] < \delta/2.$$

Since $X_k^{\le N}$ is bounded it is in $\mathscr{L}^2$. Therefore, by independence,

$$\mathrm{Var}\left(Y_n^{\le}\right) = \frac{\mathrm{Var}\left(X_k^{\le N}\right)}{n} \le \frac{N^2}{n}.$$

By linearity of expectations $\mathbb{E}\left[Y_n^{\le}\right] = \mathbb{E}\left[X_n^{\le N}\right]$, and thus tends to zero, by (9.1). It thus from Chebyshev's inequality that for n large enough $\mathbb{P}\left[|Y_n^{\le}| \ge \varepsilon/2\right] < \delta/2$. Since $\mathbb{P}[|Y_n| \ge \varepsilon] \le \mathbb{P}\left[|Y_n^{\le}| \ge \varepsilon/2 \text{ and } |Y_n^{>}| \ge \varepsilon/2\right]$, the claim follows by the union bound. □

10 Conditional expectations

10.1 Why things are not as simple as they seem

Consider a point chosen uniformly from the surface of the (idealized, spherical) earth, so that the probability of falling on a set is proportional to its area.

Say we condition on the point falling on the equator. What is the conditional distribution? It obviously has to be uniform: by symmetry, there cannot be a reason that it is more likely to be in one time zone than another.

Say now that we condition on the point falling on a particular meridian m. By the same reasoning, the conditional distribution is uniform, and so, for example, the probability that we are within 2 meters of the north pole is the same as the probability that we are within 1 meter from the equator. Integrating over m we get that regardless of the meridian, the probability of being 2 meters from the north pole is the same as the probability of being 1 meter from the equator. But the area within 2 meters of the north pole is about $4\pi m^2$, whereas the area within 1 meter of the equator is about $80000m^2$.

10.2 Conditional expectations in finite spaces

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with $|\Omega| < \infty$, $\mathscr{F} = 2^\Omega$, and $\mathbb{P}[\omega] > 0$ for all $\omega \in \Omega$. Let $\Omega = \{1, \ldots, n\}^2$, let Z be the random variable given by $Z(\omega_1, \omega_2) = \omega_1$, and let $\mathscr{G} = \sigma(Z)$ be the sigma-algebra generated by the sets $A_k = \{k\} \times \{0, \ldots, n\}$. Let X be a real random variable. Then the usual definition is the $\mathbb{E}[X|Z]$ is the *random variable* $\Omega \to \mathbb{R}$ given by

$$\mathbb{E}[X|Z](\omega) = \frac{\sum_{\omega' \in Z^{-1}(\omega)} X(\omega')\mathbb{P}\left[\omega'\right]}{\sum_{\omega' \in Z^{-1}(\omega)} \mathbb{P}\left[\omega'\right]}.$$

This notation can be confusing - $\mathbb{E}[X|Z]$ is a random variable and not a number! But given $A \in \mathscr{F}$ *with* $\mathbb{P}[A] > 0$, we denote by $\mathbb{E}[X|A]$ the *number*

$$\mathbb{E}[X|A] = \frac{1}{\mathbb{P}[A]} \mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right].$$

As $\mathscr{G} = \sigma(Y)$, it will often be less confusing to write instead $\mathbb{E}[X|\mathscr{G}]$, which denotes the same random variable.

Exercise 10.1. *Let* $Y = \mathbb{E}[X|\mathscr{G}]$.

1. $Y = \operatorname{argmin}_{W \in \mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P})} \mathbb{E}\left[(X - W)^2\right]$.
2. Y *is* $\mathscr{G}$*-measurable.*
3. *If* $A \in \mathscr{G}$ *with* $\mathbb{P}[A] > 0$ *then* $\mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[Y \cdot \mathbb{1}_{\{A\}}\right]$.

10.3 Conditional expectations in $\mathscr{L}^2$

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Given a sub-sigma-algebra $\mathscr{G} \subseteq \mathscr{F}$, we know by Theorem 7.6 that the subspace $\mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P}) \subseteq \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ is closed. We can therefore define the projection operator

$$P_{\mathscr{G}} : \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P}) \to \mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P})$$

by

$$P_{\mathscr{G}}(X) = \operatorname*{argmin}_{Y \in \mathscr{L}^2(\Omega, \mathscr{G}, \mathbb{P})} \mathbb{E}\left[(X-Y)^2\right].$$

Some immediate observations:

1. $P_{\mathscr{G}}(X)$ is $\mathscr{G}$-measurable.
2. If $Y \in \mathscr{L}^2(\Omega, \mathscr{F}, \mathbb{P})$ then $\mathbb{E}[(X - P_{\mathscr{G}}(X)) \cdot Y] = 0$, or $\mathbb{E}[X \cdot Y] = \mathbb{E}[P_{\mathscr{G}}(X) \cdot Y]$. Thus given $A \in \mathscr{G}$ with $\mathbb{P}[A] > 0$ we have that $\mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[P_{\mathscr{G}}(X) \cdot \mathbb{1}_{\{A\}}\right]$.

10.4 Conditional expectations in $\mathscr{L}^1$

Theorem 10.2. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space with a r.v. $X \in \mathscr{L}^1$ and a sub-sigma-algebra $\mathscr{G} \subseteq \mathscr{F}$. Then there exists a unique random variable Y with the following properties:*

1. *$Y \in \mathscr{L}^1(\Omega, \mathscr{G}, \mathbb{P})$.*
2. *For every $A \in \mathscr{G}$ it holds that $\mathbb{E}\left[Y \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right]$.*

We denote $\mathbb{E}[X|\mathscr{G}] := Y$. For $A \in \mathscr{F}$ with $\mathbb{P}[A] > 0$ we denote $\mathbb{E}[X|A] = \mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right]/\mathbb{P}[A]$.

Proof. We first prove uniqueness. Let Y and Z both satisfy the two conditions in the theorem, and assume by contradiction that $\mathbb{P}[Y > Z] > 0$. Then there is some $\varepsilon > 0$ such that $\mathbb{P}[Y - \varepsilon > Z] > 0$. Let $A = \{Y - \varepsilon > Z\}$, and note that $A \in \mathscr{G}$. Then

$$\begin{aligned}
\mathbb{E}\left[Y \cdot \mathbb{1}_{\{A\}}\right] &= \mathbb{E}\left[(Y-\varepsilon) \cdot \mathbb{1}_{\{A\}}\right] + \varepsilon \mathbb{P}[A] \\
&> \mathbb{E}\left[Z \cdot \mathbb{1}_{\{A\}}\right] + \varepsilon \cdot \mathbb{P}[A] \\
&> \mathbb{E}\left[Z \cdot \mathbb{1}_{\{A\}}\right].
\end{aligned}$$

But since $A \in \mathscr{G}$ we have that both $\mathbb{P}\left[Y \cdot \mathbb{1}_{\{A\}}\right]$ and $\mathbb{P}\left[Z \cdot \mathbb{1}_{\{A\}}\right]$ are equal to $\mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right]$ - contradiction.

We prove the reminder under the assumption that $X \geq 0$; the reduction is straightforward. Let $X_n = X \cdot \mathbb{1}_{\{X \leq n\}}$. Then X_n is bounded, and in particular is in $\mathscr{L}^2$. Let $Y_n = P_{\mathscr{G}}(X_n)$. We claim that Y is non-negative. To see this, assume by contradiction that $\mathbb{P}[Y_n < -\varepsilon] > 0$ for some $\varepsilon > 0$, and let $A = \{Y_n < -\varepsilon\}$. Then $\mathbb{E}\left[Y_n \cdot \mathbb{1}_{\{A\}}\right] < -\varepsilon \cdot \mathbb{P}[A] < 0$, but $\mathbb{E}\left[Y_n \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right] \geq 0$.

Now, Y_n is a monotone increasing sequence. To see this, note that X_n is monotone increasing, and that $P_{\mathcal{G}}$ is a linear operator, and so $Y_{n+1} - Y_n = P_{\mathcal{G}}(X_{n+1} - X_n)$ is non-negative, by the same proof as above.

Since Y_n is monotone increasing then so is $Y_n \cdot \mathbb{1}_{\{A\}}$, for any $A \in \mathcal{G}$. Therefore, if we define $Y = \lim_n Y_n$, then $\mathbb{E}\left[Y_n \cdot \mathbb{1}_{\{A\}}\right] \to \mathbb{E}\left[Y \cdot \mathbb{1}_{\{A\}}\right]$. But $\mathbb{E}\left[Y_n \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}[X_n \cdot A]$, and, since $X_n \cdot \mathbb{1}_{\{A\}}$ is also monotone increasing with $X \cdot \mathbb{1}_{\{A\}} = \lim_n X_n \cdot \mathbb{1}_{\{A\}}$, we have that

$$\mathbb{E}\left[Y \cdot \mathbb{1}_{\{A\}}\right] = \lim_n \mathbb{E}[Y_n \cdot A] = \lim_n \mathbb{E}\left[X_n \cdot \mathbb{1}_{\{A\}}\right] = \mathbb{E}\left[X \cdot \mathbb{1}_{\{A\}}\right].$$

Finally, each Y_n is $\mathcal{G}$-measurable by construction, and therefore so is Y. □

10.5 Some properties of conditional expectation

Exercise 10.3. *1. If X is $\mathcal{G}$-measurable (i.e., $\sigma(X) \subseteq \mathcal{G}$) then $\mathbb{E}[X|\mathcal{G}] = X$.*

2. The Law of Total Expectation. *If $\mathcal{G}_2 \subseteq \mathcal{G}_1$ then $\mathbb{E}[\mathbb{E}[X|\mathcal{G}_1]|\mathcal{G}_2] = \mathbb{E}[X|\mathcal{G}_2]$. In particular $\mathbb{E}[\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[X]$.*

3. If $Z \in \mathcal{L}^\infty(\Omega, \mathcal{G}, \mathbb{P})$ then

$$\mathbb{E}[Z \cdot X|\mathcal{G}] = Z \cdot \mathbb{E}[X|\mathcal{G}].$$

11 The Galton-Watson process

11.1 Definition

Consider an asexual organism (in the original work these were Victorian men) whose number of offspring X_0 is chosen at random from some distribution on $\mathbb{N}_0 = \{0,1,2,\ldots\}$. Each of its descendants $i \in \{1,\ldots,X_0\}$ has X_i offspring, with the random variables $(X_0, X_1, X_2, \ldots)$ distributed independently and identically. An interesting question is: what is the probability that the organism's progeny will live forever, and what is the probability that there will be a last one to its name?[2]

Formally, consider generations $\{1,2,\ldots\}$, and to each generation n associate an infinite sequence of random variables $(X_{n,1}, X_{n,2}, \ldots)$, with all the random variables $(X_{n,i})$ independent and identically distributed on $\mathbb{N}_0$. We will, to simplify some expressions, define $X = X_{1,1}$. We assume that

$$0 < \mathbb{E}[X] < \infty,$$

and denote $\mu = \mathbb{E}[X]$. We also assume that $\mathbb{P}[X = 0] > 0$.

To each generation n we define the number of organisms Z_n, which is also a random variable. It is defined recursively by $Z_1 = 1$ and $Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i}$. Clearly $Z_n = 0$ implies $Z_{n+1} = 0$.

11.2 The probability of extinction

We are interested in the event that $Z_n = 0$ for some n, or that, equivalently, $Z_n = 0$ for all n large enough. This is again equivalent to the event $\sum_n Z_n < \infty$, since each Z_n is an integer. We denote this event by E (for extinction), and denote $E_n = \{Z_n = 0\}$.

Note that the sequence E_n is increasing and $E = \cup_n E_n$. Therefore, by Theorem 3.5,

$$\mathbb{P}[E] = \lim_n \mathbb{P}[Z_n = 0].$$

We first calculate the expectation of Z_{n+1}. Since Z_n is independent of $(X_{n,1}, X_{n,2}, \ldots)$, it holds that

$$\begin{aligned}
\mathbb{E}[Z_{n+1}] &= \mathbb{E}\left[\sum_{i=1}^{Z_n} X_{n,i}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{Z_n} X_{n,i} \middle| Z_n\right]\right] \\
&= \mathbb{E}[Z_n \cdot \mathbb{E}[X|Z_n]] \\
&= \mathbb{E}[Z_n] \cdot \mathbb{E}[X],
\end{aligned}$$

[2] Neither of the names Galton and Watson have died out (as of 2018), although Galton is rather rare: https://forebears.io/surnames/galton.

and so

$$\mathbb{E}[Z_{n+1}] = \mu^n.$$

Claim 11.1. *If $\mu < 1$ then $\mathbb{P}[E] = 1$.*

Proof 1. By Markov's inequality, $\mathbb{P}[Z_n \geq 1] \leq \mu^n$. Thus by the Borel-Cantelli Lemma w.p. 1 there will be some n with $Z_n < 1$, and thus $Z_n = 0$. □

Proof 2. Note that $\mathbb{E}\left[\sum_n Z_n\right] = \sum_n \mathbb{E}[Z_n] < \infty$, and so $\mathbb{P}\left[\sum_n Z_n = \infty\right] = 0$. □

It is also true that $\mathbb{P}[E] = 1$ when $\mu = 1$. Note that in this case

$$\mathbb{E}[Z_{n+1}|Z_1, Z_2, \ldots, Z_n] = \mathbb{E}[Z_{n+1}|Z_n] = Z_n \cdot \mathbb{E}[X] = Z_n.$$

The first equality makes Z_n a *Markov chain*. The second makes it a *Martingale*; we will discuss both concepts formally. By the Martingale Convergence Theorem we have that Z_n converges almost surely to some r.v. Z_∞. But clearly Z_n cannot converge to anything but 0, and so $\mathbb{P}[E] = 1$.

Note that the event E is equal to the union of the event that $X_{1,1} = 0$ with the event that $X_{1,1} > 0$ but each of the sub-tree of the Z_2 offspring goes extinct. Since the process on each subtree is identical, and since the probability that all of such k offspring trees goes extinct is $\mathbb{P}[E]^k$, we have that $\mathbb{P}[E]$ must satisfy

$$\mathbb{P}[E] = \sum_{k \in \mathbb{N}_0} \mathbb{P}[X = k]\mathbb{P}[E]^k. \tag{11.1}$$

11.3 The probability generating function

We accordingly define $f : [0,1] \to [0,1]$, the *probability generating function* of X, by

$$f(t) = \sum_{k \in \mathbb{N}_0} \mathbb{P}[X = k] \cdot t^k = \mathbb{E}\left[t^X\right],$$

where we take $0^0 = 1$. Then (11.1) is equivalent to observing that $\mathbb{P}[E]$ is a fixed point of f. Note that 1 is always a fixed point, but in general there might be more.

Some observations:

1. $f(0) = \mathbb{P}[X = 0]$ and $f(1) = 1$.
2. $f'(t) = \sum_{k \in \mathbb{N}} \mathbb{P}[X = k] \cdot k \cdot t^{k-1} = \mathbb{E}\left[X \cdot t^{X-1}\right]$. Hence $f'(1) = \mathbb{E}[X] = \mu$. Note also that $f'(t) \geq 0$.
3. Likewise, the k^{th} derivative of f is non-negative. Thus f is convex.

Let $f_n(t) = \mathbb{E}\left[t^{Z_n}\right]$ be the generating function of Z_n. Then

$$\begin{aligned} f_{n+1}(t) &= \mathbb{E}\left[t^{Z_{n+1}}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[t^{Z_{n+1}}\middle| Z_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[t^{\sum_{k=1}^{Z_n} X_{n,k}}\middle| Z_n\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[t^X\right]^{Z_n}\right] \\ &= \mathbb{E}\left[f(t)^{Z_n}\right] \\ &= f_n(f(t)), \end{aligned}$$

where we again used the fact that Z_n is independent of $(X_{n,1}, X_{n,2}, \ldots)$. Since $f_1(t) = t$, f_{n+1} is the n-fold composition of f with itself:

$$f_{n+1} = f \circ f \circ \cdots \circ f.$$

Now $\mathbb{P}[Z_n = 0] = f_n(0)$. Since f is analytic,

$$\mathbb{P}[E] = \lim_n \mathbb{P}[E_n] = \lim_n f_n(0)$$

will be the fixed point of f that one converges to by applying f repeatedly to 0. Furthermore, $f(0) = \mathbb{P}[X = 0] > 0$, $f(1) = 1$, and f is increasing and convex. Thus f will have a unique fixed point. Finally, since $f'(1) = \mu$, this fixed point will be 1 iff $\mu \leq 1$.

12 Markov chains

12.1 Definition

Let the state space S be a countable or finite set. A sequence of S-valued random variables $(X_0, X_1, X_2, \ldots)$ is said to be a *Markov chain* if for all $x \in S$ and $n > 0$

$$\mathbb{P}[X_n = x | X_0, X_1, \ldots, X_{n-1}] = \mathbb{P}[X_n = x | X_{n-1}].$$

A Markov chain is said to be *time homogeneous* if $\mathbb{P}[X_n = x | X_{n-1}]$ does not depend on n. In this case it will be useful to study the associated stochastic S-indexed matrix $P(x,y) = \mathbb{P}[X_{n+1} = y | X_n = x]$. It is easy to see that $\mathbb{P}[X_{n+m} = y | X_n = x] = P^m(x,y)$, where P^m denotes the usual matrix exponentiation. We call P the transition matrix of the Markov chain.

In the context of a transition matrix P, we will denote by $\mathbb{P}^x$ the measure of the Markov chain for which $\mathbb{P}[X_0 = x] = 1$.

The next claim is needed to formally apply the Markov property.

Claim 12.1. *Let $(X_0, X_1, \ldots)$ be a time homogeneous Markov chain. Fix some measurable $f : S^{\mathbb{N}} \to \mathbb{R}$ and denote*

$$Y_n = f(X_n, X_{n+1}, \ldots).$$

Then for any $n, m \in \mathbb{N}$ and $x \in S$ such that $\mathbb{P}[X_n = x] > 0$ and $\mathbb{P}[X_m = x] > 0$ it holds that

$$\mathbb{E}[Y_{n+1} | X_n = x] = \mathbb{E}[Y_{m+1} | X_m = x].$$

Example: let $S = \mathbb{Z}$, let $X_0 = 0$, and let $P(x,y) = \frac{1}{2}\mathbb{1}_{\{|x-y|=1\}}$. This is called the *simple random walk* on $\mathbb{Z}$. More generally (in some direction), one can consider a graph $G = (S, E)$ with finite positive out-degrees $d(x) = |E \cap \{x\} \times S|$ and let

$$P(x,y) = \frac{\mathbb{1}_{\{(x,y) \in E\}}}{d(x)}.$$

The *lazy random walk* on $\mathbb{Z}$ has transition probabilities $P(x,y) = \frac{1}{3}\mathbb{1}_{\{|x-y| \le 1\}}$.

12.2 Irreducibility and aperiodicity

We say that a (time homogeneous) Markov chain is *irreducible* if for all $x, y \in S$ there exists some m so that $P^m(x,y) > 0$. We say that an irreducible chain is *aperiodic* if for some (equivalently, every) $x \in S$ it holds that $P^m(x,x) > 0$ for all m large enough.

Exercise 12.2. *Show that if an irreducible chain is not aperiodic then for every $x \in S$ there is a $k \in \mathbb{N}$ so that $P^m(x,x) = 0$ for all m not divisible by k.*

Exercise 12.3. 1. *Show that the simple random walk on $\mathbb{Z}$ is irreducible but not aperiodic.*

2. *Show that the lazy random walk on $\mathbb{Z}$ is irreducible and aperiodic.*

3. *Show that the simple random walk on a directed graph is irreducible iff the graph is strongly connected.*

4. *Show that the simple random walk on a connected, undirected graph is aperiodic iff the graph is not bipartite.*

12.3 Recurrence

We define the hitting time to $x \in S$ by

$$T_x = \min\{n > 0 : X_n = x\}.$$

This is a random variable taking values in $\mathbb{N} \cup \{\infty\}$. An irreducible Markov chain is said to be *recurrent* if $\mathbb{P}[T_x < \infty] = 1$ whenever $\mathbb{P}[X_0 = x] > 0$. A non-recurrent random walks is called *transient*.

Theorem 12.4. *Fix an irreducible Markov chain with $\mathbb{P}[X_0 = x] > 0$ for all $x \in S$. Then the following are equivalent.*

1. *The Markov chain is recurrent.*

2. *For some (all) $x \in X$ it holds that*

$$\mathbb{P}[X_n = x \text{ i.o.}] = 1.$$

3. *For some (all) $x \in X$ it holds that $\sum_m P^m(x,x) = \infty$.*

Proof. Choose any $x \in S$. Since $\mathbb{P}[T_x < \infty] = 1$, and since $\mathbb{P}[X_0 = y] > 0$ for any $y \in S$, we have that $\mathbb{P}[T_x < \infty | X_0 = y] = 1$, or that

$$\mathbb{P}[X_n = x \text{ for some } n > 0 | X_0 = y] = 1.$$

By irreducibility we have that $\mathbb{P}[X_m = y] > 0$ for any m, and so by the Markov property it follows that

$$\mathbb{P}[X_n = x \text{ for some } n > m | X_m = y] = 1.$$

Summing over y yields that

$$\mathbb{P}[X_n = x \text{ for some } n > m] = 1,$$

and so

$$\mathbb{P}[X_n = x \text{ i.o.}] = 1.$$

We have thus shown that (1) implies (2).

Note that $P^m(x,x) = \mathbb{P}[X_m = x | X_0 = x]$. Now, (2) implies that

$$\mathbb{P}[X_n = x \text{ i.o.} | X_0 = x] = 1$$

and so, by Borel-Cantelli, (2) implies (3).

Finally, to show that (3) implies (1), assume that the Markov chain is transient. Then $\mathbb{P}[T_x < \infty] < 1$, and so $\mathbb{P}[T_x < \infty | X_0 = x] < 1$. Denote the latter by p. Hence, by the Markov property,

$$p = \mathbb{P}[X_n = x \text{ for some } n > m | X_m = x].$$

Therefore, conditioned on $X_0 = x$, the probability that x is visited k more times is $p^k(1-p)$. In particular the expected number of visits is finite, and since this expectation is equal to $\sum_m P^m(x,x)$, the proof is complete. □

Exercise 12.5. *Prove that every irreducible Markov chain over a finite state space is recurrent.*

Exercise 12.6. *Let P be the transition matrix of a Markov chain over S, and for $\varepsilon > 0$ let $P_\varepsilon = (1-\varepsilon)P + \varepsilon I$, where I is the identity matrix. Thus P_ε is the ε-lazified version of P. Consider two Markov chains over S: both with $X_0 = x$, and one with transition matrix P and the other with transition matrix P_ε. Prove that either both are recurrent or both are transitive.*

12.4 The simple random walk on $\mathbb{Z}^d$

Corollary 12.7. *The simple random walk on $\mathbb{Z}$ is recurrent.*

Proof. Note that $\mathbb{P}[X_{2n+1} = 0] = 0$ and that

$$\mathbb{P}[X_{2n} = 0] = 2^{-2n}\binom{2n}{n}.$$

By Stirling

$$\binom{2n}{n} \geq \frac{2^{2n-1}}{\sqrt{n}},$$

and so

$$\mathbb{P}[X_{2n} = 0] \geq \frac{1}{2\sqrt{n}}.$$

Hence

$$\sum_m P^m(0,0) \geq \frac{1}{2\sqrt{m}} = \infty,$$

and the claim follows by Theorem 12.4. □

Consider now a random walk with a drift on $\mathbb{Z}$. For example, let $P(x,y) = p$ if $y = x+1$ and $P(x,y) = 1-p$ if $y = x-1$. In this case, assuming $X_0 = 0$, $X_n = \sum_{k \le n} Y_n$ where the Y_n are i.i.d. r.v. with $\mathbb{P}[Y_n = 1] = p$ and $\mathbb{P}[Y_n = -1] = 1-p$. It follows from the strong law of large numbers that a.s. $\lim_n X_n/n = 2p - 1 > 0$, and so in particular $\lim_n X_n = \infty$, and the random walk is transient. The same argument holds whenever the transition probabilities correspond to an $\mathscr{L}^1$ random variable with non-zero expectation, by the same argument (although we have yet to prove an $\mathscr{L}^1$ SLLN).

Exercise 12.8. *Prove that the simple random walk on $\mathbb{Z}^2$ (given by $P(x,y) = \frac{1}{4}\mathbb{1}_{\{|x-y|=1\}}$) is recurrent, but that the simple random walk on $\mathbb{Z}^d$ (given by $P(x,y) = \frac{1}{d}\mathbb{1}_{\{|x-y|=1\}}$) is transient for all $d \ge 3$.*

13 Martingales

13.1 Definition

A *filtration* $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ is a sequence of increasing sigma-algebras $\mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \cdots$. A natural (and in some sense only) example is the case that $\mathscr{F}_n = \sigma(Y_1, \ldots, Y_n)$ for some sequence of random variables $(Y_1, Y_2, \ldots)$.

A process $(X_1, X_2, \ldots)$ is said to be *adapted* to Φ if each X_n is $\mathscr{F}_n$-measurable. A sequence of real random variables $(X_1, X_2, \ldots)$ that is adapted to Φ and is in $\mathscr{L}^1$ is called a martingale with respect to Φ if for all $n \geq 1$

$$\mathbb{E}[X_{n+1}|\mathscr{F}_n] = X_n.$$

It is called a supermartingale if

$$\mathbb{E}[X_{n+1}|\mathscr{F}_n] \leq X_n.$$

Note that if $(X_1, X_2, \ldots)$ is a martingale then $\mathbb{E}[X_n] = \mathbb{E}[X_1]$ and by subtracting the constant $\mathbb{E}[X_1]$ from all X_n's we get that $(X_0, X_1, \ldots)$ is a martingale with $X_0 = 0$. A similar statement holds for supermartingales.

As a first example, let W_n be i.i.d. r.v. with $\mathbb{P}[W_n = +1] = \mathbb{P}[W_n = -1] = 1/2$, let $X_n = \sum_{k \leq n} W_n$, and let $\mathscr{F}_n = \sigma(X_1, \ldots, X_n)$. Then X_n is the amount of money made in n fair bets (or the locations of a simple random walk on $\mathbb{Z}$) and is a martingale with respect to $(\mathscr{F}_1, \mathscr{F}_2, \ldots)$. If we set $\mathbb{P}[W_n = +1] = 1/2 - \varepsilon$ and $\mathbb{P}[W_n = -1] = 1/2 + \varepsilon$ for some $\varepsilon > 0$ then X_n is a supermartingale.

13.2 Examples

As a second example we introduce *Pólya's urn*. Consider an urn in which there are initially a single black ball and a single white ball. In each time period we reach in, pull out a ball, and then put back two balls of the same color. Formally, let $(Y_1, Y_2, \ldots)$ be i.i.d. random variables distributed uniformly over $[0, 1]$, and let the number of black balls at time n be B_n, given by $B_1 = 1$ and

$$B_{n+1} = B_n + \mathbb{1}_{\{Y_n < B_n/(n+1)\}}.$$

Denote by $R_n = B_n/(n+1)$ the fraction of black balls. Then

$$\mathbb{E}[R_{n+1}|B_1, \ldots, B_n] = \mathbb{E}[R_{n+1}|B_n],$$

since the process $(B_1, B_2, \ldots)$ is a Markov chain. Furthermore

$$\begin{aligned}
\mathbb{E}[R_{n+1}|B_n] &= \frac{1}{n+2}\mathbb{E}[B_{n+1}|B_n] \\
&= \frac{1}{n+2}\left(B_n + \frac{B_n}{n+1}\right) \\
&= \frac{B_n}{n+1} \\
&= R_n,
\end{aligned}$$

and so R_n is a martingale with respect to $\mathscr{F}_n = \sigma(B_1, \ldots, B_n)$.

As a third example let $X \in \mathscr{L}^2$ (e.g., X is a standard Gaussian), let $Y_1, Y_2, \ldots$ be i.i.d. in $\mathscr{L}^2$, and let $Z_n = X + Y_n$. This can be thought of as a model of independent measurements (Z_n) with noise (Y_n) of a physical quantity (X). Under this interpretation,

$$\tilde{X}_n = \mathbb{E}[X | Y_1, \ldots, Y_n]$$

is a natural estimator of X. By the law of total expectations

$$\mathbb{E}\left[\tilde{X}_{n+1} \middle| Y_1, \ldots, Y_n\right] = \tilde{X}_n,$$

and thus $\tilde{X}_n$ is also a martingale.

13.3 Martingale convergence in $\mathscr{L}^2$

Theorem 13.1 (Martingale Convergence in $\mathscr{L}^2$)**.** *Let $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ be a filtration, and let $(X_1, X_2, \ldots)$ be a martingale w.r.t. Φ. Furthermore, assume that there exists a K such that $\mathbb{E}\left[X_n^2\right] < K$ for all n. Then there exists a random variable $X \in \mathscr{L}^2$ such that $\mathbb{E}\left[(X - X_n)^2\right] \to 0$.*

Proof. Set $X_0 = 0$, and for $n \geq 1$ let $Y_n = X_n - X_{n-1}$. Since $X_{n-1} = \mathbb{E}[X_n | \mathscr{F}_{n-1}]$, we have that Y_n is orthogonal to any $\mathscr{F}_{n-1}$-measurable r.v., and in particular is orthogonal to Y_m for any $m < n$. Now,

$$\sum_{k \leq n} Y_n = X_n$$

and so by the orthogonality of the Y_n's it follows that

$$\sum_{k \leq n} \mathbb{E}\left[Y_k^2\right] = \mathbb{E}\left[X_n^2\right] < K.$$

Thus

$$\sum_{k} \mathbb{E}\left[Y_n^2\right] < K,$$

and we have that X_n is a Cauchy sequence in $\mathscr{L}^2$. Therefore, since $\mathscr{L}^2$ is complete (Theorem 7.6) there exists some $X \in \mathscr{L}^2$ such that $\mathbb{E}\left[(X - X_n)^2\right] \to 0$. □

The next theorem shows that, in fact, convergence is pointwise, and an $\mathscr{L}^1$ assumption suffices.

13.4 The Martingale Convergence Theorem

Theorem 13.2 (Martingale Pointwise Convergence)**.** *Let $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ be a filtration, and let $(X_1, X_2, \ldots)$ be a supermartingale w.r.t. Φ. Furthermore, assume that there exists a K such that $\mathbb{E}[|X_n|] < K$ for all n. Then there exists a random variable $X \in \mathscr{L}^1$ such that almost surely* $\lim_n X_n = X$.

Before proving this theorem we will need the following lemmas.

Lemma 13.3. *Let $(X_0, X_1, X_2, \ldots)$ be a supermartingale w.r.t. $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ with $X_0 = 0$, let B_n be $\{0,1\}$-values random variables adapted to Φ, and let $Y_n = \sum_{k \le n} B_{k-1} \cdot (X_k - X_{k-1})$. Then Y_n is a supermartingale and $\mathbb{E}[Y_n] \le 0$.*

The idea behind this lemma is the following: imagine that you are gambling at a casino with non-positive expected wins from every gamble. Say that you have some system for deciding when to gamble and when to stay out (i.e., the B_n's). Then you do not expect to win more than you would have if you stayed in the game every time.

Proof.

$$\begin{aligned}
\mathbb{E}[Y_{n+1} | \mathscr{F}_n] &= \mathbb{E}\left[\sum_{k \le n+1} B_{k-1} \cdot (X_k - X_{k-1}) \middle| \mathscr{F}_n \right] \\
&= \mathbb{E}[Y_n + B_n \cdot (X_{n+1} - X_n) | \mathscr{F}_n] \\
&= Y_n + B_n \mathbb{E}[X_{n+1} - X_n | \mathscr{F}_n] \\
&= Y_n + B_n (\mathbb{E}[X_{n+1} | \mathscr{F}_n] - X_n) \\
&\le Y_n.
\end{aligned}$$

Thus Y_n is a supermartingale, and by induction $\mathbb{E}[Y_n] \le 0$. □

Lemma 13.4. *Let $(X_0, X_1, X_2, \ldots)$ be a supermartingale w.r.t. $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ with $X_0 = 0$. Fix some $a < b$, and let B_n be defined as follows: $B_0 = 0$, and B_{n+1} is the indicator of the union of the events*

1. *$B_n = 1$ and $X_n \le b$.*
2. *$B_n = 0$ and $X_n < a$.*

Let $U_n^{a,b}$ be the number of $k \le n$ such that $B_k = 0$ and $B_{k-1} = 1$. Then

$$\mathbb{E}\left[U_n^{a,b}\right] \le \frac{\mathbb{E}[(X_n - a)^-]}{b - a}.$$

Proof. By picture, it is clear that for

$$Y_n = \sum_{k \le n} B_{n-1} \cdot (X_n - X_{n-1})$$

it holds that

$$Y_n \ge (b - a) U_n^{a,b} - (X_n - a)^-.$$

By Lemma 13.3 we have that $\mathbb{E}[Y_n] \le 0$, and so the claim follows by taking expectations. □

Proof of Theorem 13.2. For a given $a < b$, let $U_\infty^{a,b} = \lim_n U_n^{a,b}$. The limit exists since this is a monotone increasing sequence, and it also follows that

$$\mathbb{E}\left[U_\infty^{a,b}\right] = \lim_n \mathbb{E}\left[U_n^{a,b}\right] \le \lim_n \frac{\mathbb{E}[(X_n - a)^-]}{b-a} \le \frac{|a| + K}{b-a} < \infty.$$

Thus $\mathbb{P}\left[U_\infty^{a,b} < \infty\right] = 1$, and it follows that with probability zero it occurs that $\limsup_n X_n \ge b$ and $\liminf_n X_n \le a$. Applying this to a countable dense set of pairs (a,b) we get that with probability zero $\limsup_n X_n > \liminf_n X_n$, and so $\limsup_n X_n = \liminf_n X_n$ almost surely. □

Exercise 13.5. *Let R_n be the fraction of black balls in Pólya's urn. Show that $\lim_n R_n$ is distributed uniformly on $(0,1)$. Hint: calculate the distribution of R_n.*

14 Stopping times

14.1 Definition

Let $\Phi = (\mathscr{F}_1, \mathscr{F}_2, \ldots)$ be a filtration, and let $(X_0, X_1, X_2, \ldots)$ be a supermartingale w.r.t. Φ. Denote $\mathscr{F}_\infty = \sigma(\cup_n \mathscr{F}_n)$.

A random variable T taking values in $\mathbb{N} \cup \{\infty\}$ is called a *stopping time* if for all $n \leq \infty$ it holds that the event $\{T \leq n\}$ is $\mathscr{F}_n$-measurable.

Example: $(X_1, X_2, \ldots)$ is a Markov chain over the state space S, and T_x is the hitting time to $x \in S$ given by

$$T_x = \min\{n > 0 : X_n = x\}.$$

Example: $(X_1, X_2, \ldots)$ is the simple random walk on $\mathbb{Z}$, and T is the first $n \geq 3$ such that $X_n < X_{n-1} < X_{n-2}$.

Given a stopping time T, we define the *stopped process*

$$(X_1^T, X_2^T, \ldots) = (X_1, X_2, \ldots, X_{T-1}, X_T, X_T, \ldots).$$

That is, $X_n^T = X_n$ if $n \leq T$, and $X_n^T = X_T$ if $n \geq T$. Equivalently, $X_n^T = X_{\min\{T,n\}}$. Intuitively, the stopped process corresponds to the process of a gambler's bank account, when the gambler decides stopping at time T.

Theorem 14.1. *If $(X_0, X_1, X_2, \ldots)$ is a (super)martingale (with $X_0 = 0$) then $(X_0^T, X_1^T, X_2^T, \ldots)$ is a (super)martingale.*

Proof. We prove for the case of supermartingales; the proof for martingales is identical.

Let $B_n = \mathbb{1}_{\{T \geq n\}}$ and $Y_n = \sum_{k \leq n} B_{k-1} \cdot (X_k - X_{k-1})$. Then by Lemma 13.3 we have that Y_n is a supermartingale. But $Y_n = X_n^T$. □

So the gambler's bank account is still a martingale, no matter what the stopping time is, and in particular $\mathbb{E}\left[X_n^T\right] \leq 0$ (with equality for martingales). However, consider a simple random walk on $\mathbb{Z}$, with stopping time T_1. That is, the gambler stops once she has earned a dollar. Then clearly $\mathbb{E}\left[X_{T_1}\right] = 1$. The following theorem gives conditions for when $\mathbb{E}[X_T] = 0$.

14.2 Optional Stopping Time Theorem

Theorem 14.2 (Doob's Optional Stopping Time Theorem)**.** *Let $(X_0, X_1, \ldots)$ be a supermartingale, and let T be a stopping time. Assume that $\mathbb{P}[T = \infty] = 0$, and that one of following holds:*

1. *$\exists N$ s.t. $\mathbb{P}[T \leq N] = 1$.*
2. *$\exists K$ s.t. $\mathbb{P}[|X_n| \leq K \text{ for all } n] = 1$.*
3. *$\mathbb{E}[T] < \infty$ and $\exists K$ s.t. $\mathbb{P}[|X_{n+1} - X_n| \leq K \text{ for all } n] = 1$.*
4. *X_n is non-negative.*

Then $\mathbb{E}[X_T] \leq \mathbb{E}[X_0]$*, with equality if* $(X_0, X_1, \ldots)$ *is a martingale.*

To prove this theorem we will need the following important lemma.

Lemma 14.3 (Fatou's Lemma)**.** *Let* $(Z_1, Z_2, \ldots)$ *be a sequence of non-negative real random variables. Then*

$$\mathbb{E}\left[\liminf_n Z_n\right] \leq \liminf_n \mathbb{E}[Z_n].$$

Recall from the Galton-Watson example that indeed this may be a strict inequality.

Exercise 14.4. *Prove Fatou's Lemma. Hint: use the Monotone Convergence Theorem.*

Proof. We prove that $\mathbb{E}[X_T] \leq 0$; the equality in case of the martingales follows easily.

Note that $\mathbb{E}\left[X_n^T\right] \leq 0$, by Theorem 14.1. Also $\lim_n X_n^T = X_T$, since $\mathbb{P}[T < \infty] = 1$ under all conditions.

1. $X_T = X_N^T$.
2. By the Bounded Convergence Theorem $\mathbb{E}\left[X^T\right] = \lim_n \mathbb{E}\left[X_n^T\right] \leq 0$.
3. $$\left|X_n^T\right| = \left|\sum_{k=1}^{\min\{T,n\}} X_k - X_{k-1}\right| \leq K \cdot T.$$

   Hence by the Dominated Convergence Theorem $\mathbb{E}[X_T] = \mathbb{E}\left[X_n^T\right]$.
4. By Fatou's Lemma,

$$\mathbb{E}[X_T] \leq \liminf_n \mathbb{E}\left[X_n^T\right] \leq 0.$$

□

Corollary 14.5. *Let* T_1 *be the hitting time to* 1 *of the simple random walk on* $\mathbb{Z}$*. Then* $\mathbb{E}[T_1] = \infty$*.*

15 Harmonic and superharmonic functions

15.1 Definition

Let $(X_0, X_1, \ldots)$ be a Markov chain over the state space S with transition matrix P. We say that a function $f : S \to \mathbb{R}$ is *P-harmonic* if $Pf = f$. Here $Pf : S \to \mathbb{R}$ is

$$[Pf](x) = \sum_{y \in S} P(x,y) f(y).$$

We say that f is *P-superharmonic* if $[Pf](x) \leq f(x)$ for all $x \in S$.

15.2 Harmonic functions and martingales

Claim 15.1. *Assume that P is irreducible, so that for all x there exists an n such that $\mathbb{P}[X_n = x] > 0$. Let $Z_n = f(X_n)$. Then Z_n is a (super)martingale iff f is (super)harmonic.*

Proof. We prove for the (super) case:

$$\mathbb{E}[f(X_{n+1}) | X_0, \ldots, X_n] = \mathbb{E}[f(X_{n+1}) | X_n] = \sum_{y \in S} P(X_n, y) f(y) \leq f(X_n)$$

iff f is superharmonic. □

15.3 Superharmonic functions and recurrence

Theorem 15.2. *Let P be irreducible. Then the following are equivalent.*

1. *Every Markov chain with transition matrix P is recurrent.*
2. *Some Markov chain with transition matrix P is recurrent.*
3. *Every non-negative P-superharmonic function is constant.*

Proof. The equivalence of (1) and (2) follows easily from Theorem 12.4.

To see that (1) implies (3), let T_y be the hitting time to y, and note that $\mathbb{P}_x\left[T_y < \infty\right] = 1$, by recurrence. Let f be a non-negative superharmonic function, and let $Z_n = f(X_n)$. Then we can apply the Optional Stopping Time Theorem to $Z_n^{T_y}$ to get that

$$\mathbb{E}_x\left[Z_{T_y}\right] \leq \mathbb{E}_x[Z_0].$$

The l.h.s. is equal to $f(y)$ and the r.h.s. is equal to $f(x)$, and so f is constant.

Assume (3), and note that

$$\begin{aligned}\mathbb{P}_x\left[T_y<\infty\right] &= \mathbb{P}_x\left[X_1=y, T_y<\infty\right]+\mathbb{P}_x\left[X_1\neq y, T_y<\infty\right]\\ &= \mathbb{P}_x\left[X_1=y\right]+\sum_{z\neq y}\mathbb{P}_x\left[X_1=z, T_y<\infty\right]\\ &= \mathbb{P}_x\left[X_1=y\right]+\sum_{z\neq y}\mathbb{P}_x\left[T_y<\infty\middle|X_1=z\right]\cdot\mathbb{P}_x\left[X_1=z\right]\\ &= P(x,y)+\sum_{z\neq y}P(x,z)\cdot\mathbb{P}_z\left[T_y<\infty\right]\\ &\geq \sum_z P(x,z)\cdot\mathbb{P}_z\left[T_y<\infty\right].\end{aligned}$$

Hence $f(x)=\mathbb{P}_x\left[T_y<\infty\right]$ is superharmonic, and thus constant by assumption. Say $p=\mathbb{P}_x\left[T_y<\infty\right]$. By irreducibility $p>0$. Hence, by the Markov property, for every N the expected number of visits at times $n>N$ is at least p, and so the expected number of visits is infinite. Thus the random walk is recurrent. □

15.4 Bounded harmonic functions

The following claim is a direct consequence of Claim 15.1 and the Martingale Convergence Theorem.

Claim 15.3. *Let $f: S\to\mathbb{R}$ be bounded and superharmonic. Then $Z_n=f(X_n)$ is a bounded supermartingale and therefore converges almost surely to $Z:=\lim_n Z_n$.*

Recall that $\mathcal{T}_n=\sigma(X_n,X_{n+1},\ldots)$ and that

$$\mathcal{T}=\cap_n\mathcal{T}_n$$

is the tail sigma-algebra. We think of our probability space as being $(\Omega,\mathcal{F},\mathbb{P})$ with $\Omega=S^{\mathbb{N}}$ and $\mathcal{F}$ the Borel sigma-algebra of the product of the discrete topologies. Then $A\in\mathcal{T}_n$ iff A is of the form $S^n\times B$ for some measurable $B\in\mathcal{F}$. Equivalently, $A\in\mathcal{T}_n$ iff for every $(x_0,x_1,\ldots)\in A$, and $(y_0,\ldots,y_{n-1})\in S^n$ it holds that

$$(y_0,\ldots,y_{n-1},x_n,x_{n+1},\ldots)\in A.$$

15.5 The shift-invariant sigma-algebra

Another important sigma-algebra is the *shift-invariant sigma-algebra* $\mathcal{I}$. To define it, let $\varphi: S^{\mathbb{N}}\to S^{\mathbb{N}}$ be the shift map given by $\varphi(x_0,x_1,x_2,\ldots)=(x_1,x_2,\ldots)$. Then $\mathcal{I}$ is the φ-invariant subsets of $S^{\mathbb{N}}$. That is,

$$\mathcal{I}=\{A\subset S^{\mathbb{N}}:\varphi^{-1}(A)=A\}.$$

Equivalently, $A\in\mathcal{I}$ iff for every $(x_0,x_1,\ldots)\in A$, and $y\in S$ it holds that

$$(y,x_1,x_2,\ldots)\in A$$

and

$$(x_2, x_3, \ldots) \in A.$$

Exercise 15.4. *Show that $\mathscr{I} \subset \mathscr{T}$, but that the two are not equal.*

Exercise 15.5. *Find an irreducible Markov chain on the state space $\mathbb{N}$ that has a random variable that is $\mathscr{T}$-measurable but not $\mathscr{I}$-measurable.*

Claim 15.6. *Let $Z = \lim_n f(X_n)$ for some bounded harmonic f. Then Z is $\mathscr{I}$-measurable.*

The proof is a (perhaps tedious) application of the definition.

Since Z is bounded we have that $Z \in \mathscr{L}^\infty(\mathscr{I})$.

Theorem 15.7. *For every $Z \in \mathscr{L}^\infty(\mathscr{I})$ there is a bounded harmonic function f such that $Z = \lim_n f(X_n)$.*

Proof. For $x \in S$ choose any n such that $\mathbb{P}[X_n = x] > 0$, and let $f(x) = \mathbb{E}[Z|X_n = x]$. This is well defined (i.e., independent of the choice of n) because Z is $\mathscr{I}$-measurable. It is straightforward to check that f is bounded.

If $\mathbb{P}[X_n = x] > 0$ then $\mathbb{P}[X_{n+1} = y] > 0$ for all y such that $P(x,y) > 0$, and so

$$\begin{aligned}
[Pf](x) &= \sum_y P(x,y) f(y) \\
&= \sum_y \mathbb{P}[X_{n+1} = y | X_n = x] \cdot \mathbb{E}[Z | X_{n+1} = y] \\
&= \sum_y \mathbb{P}[X_{n+1} = y | X_n = x] \cdot \mathbb{E}[Z | X_{n+1} = y, X_n = x] \\
&= \mathbb{E}[Z | X_n = x] \\
&= f(x),
\end{aligned}$$

where the equality before last follows from the Markov property. Thus f is harmonic.

To see that $Z = \lim_n f(X_n)$, note that, by the martingale convergence theorem,

$$Z = \lim_n \mathbb{E}[Z | X_1, \ldots, X_n].$$

By the Markov property

$$\mathbb{E}[Z | X_1, \ldots, X_n] = \mathbb{E}[Z | X_n] = f(X_n),$$

and so $Z = \lim_n f(X_n)$. □

To summarize, denote by $h^\infty(S,P) \subset \ell^\infty(S)$ the bounded P-harmonic functions. If $f \in h^\infty(S,P)$, then $Z = \lim_n f(X_n)$ is in $\mathscr{L}^\infty(\mathscr{I})$. Conversely, if $Z \in \mathscr{L}^\infty(\mathscr{I})$ then $f = \mathbb{E}[Z|X_n = x]$ is in $h^\infty(S,P)$.

It turns out that the map $\Phi\colon \mathscr{L}^\infty(\mathscr{T}) \to h^\infty(S,P)$ given by $\Phi\colon Z \mapsto f$ is a linear isometry.

16 The Choquet-Deny Theorem

16.1 The asymptotic direction of a random walk

As motivation, consider the simple random walk $(X_1, X_2, \ldots)$ on $\mathbb{Z}^3$. Let $P_n = X_n/|X_n|$ be the projection of X_n to the unit sphere (and assume $P_n = 0$ whenever $X_n = 0$). Since this random walk is transient, it is easy to deduce that $\lim_n |X_n| = \infty$. It follows that $\lim_n |P_{n+1} - P_n| = 0$; that is, the projection moves more and more slowly. A natural question is: does P_n converge?

Theorem 16.1 (Choquet-Deny Theorem)**.** *Let $(Y_1, Y_2, \ldots)$ be i.i.d. random variables taking values in some countable abelian group G. Let $X_n = \sum_{k \le n} Y_n$. Then $(X_1, X_2, \ldots)$ is a time homogeneous Markov chain over the state space G. If $(X_1, X_2, \ldots)$ is also irreducible then every $W \in \mathscr{L}^\infty(\mathscr{I})$ is constant.*

For the proof of this theorem we will need an important classical result from convex analysis.

16.2 The Krein-Milman Theorem

Theorem 16.2 (Krein-Milman Theorem)**.** *Let X be a Hausdorff locally convex topological space. A point $x \in C$ is* extreme *if whenever x is equal to the non-trivial convex combination $\alpha y + (1-\alpha)z$ then $y = z$.*

Let C be compact convex subset of X. Then every $x \in C$ can be written as the limit of convex combinations of extreme points in C.

Proof of Theorem 16.1. Denote by P the transition matrix of $(X_1, X_2, \ldots)$, and let $\mu(g) = \mathbb{P}[Y_n = g]$. Then $P(g,k) = \mu(k-g)$. Thus, if f is P-harmonic then

$$f(g) = \sum_{k \in G} f(k)P(g,k) = \sum_{k \in G} f(k)\mu(k-g) = \sum_{k \in G} f(g+k)\mu(k).$$

Let $H = h^{[0,1]}(G,P)$ be the set of all P-harmonic functions with range in $[0,1]$. We note that harmonicity is invariant to multiplication by a constant and addition, and so if we show that every $f \in h^{[0,1]}(G,P)$ is constant then we have shown that every $f \in H$ is constant. It then follows that every $W \in \mathscr{L}^\infty(\mathscr{I})$ is constant, by the fact that Φ is an isometry.

We state three properties of H that are easy to verify.

1. H is invariant to the G action: for any $f \in H$ and $g \in G$, the function $f^g : G \to \mathbb{R}$ given by $[f^g](k) = f(k-g)$ is also in H.

2. H is compact in the topology of pointwise convergence.

3. H is convex.

As a convex compact space, H is the closed convex hull of its extreme points; this is the Krein-Milman Theorem. Thus H has extreme points. Let $f \in H$ be an extreme point. Then, since f is harmonic,

$$f(g) = \sum_{k \in G} f(g+k)\mu(k) = \sum_{k \in G} f^{-k}(g)\mu(k).$$

By the first property of H each f^{-k} is also in H, and thus we have written f as a convex combination of functions in H. But f is extreme, and so $f = f^{-k}$ for all k in the support of μ. But since the Markov chain is irreducible, the support of μ generates G. Hence f is invariant to the G-action, and therefore constant.

□

An immediate corollary of the Choquet-Deny Theorem is that every event in $\mathscr{I}$ has probability either 0 or 1. As an application, consider the question on the simple random walk on $\mathbb{Z}^3$. We would like to show that P_n does not converge pointwise. Note that the event that P_n converges is a shift-invariant event, and therefore has measure in $\{0,1\}$. Assume by contradiction that it has measure 1, and let $P = \lim_n P_n$. For each Borel subset B of the sphere, the event that $P \in B$ is shift-invariant, and therefore has measure in $\{0,1\}$. It follows that there is some p such that $P = p$ almost surely. But by the symmetry of the problem the probability that $P = p$ is the same as he probability that $P = -p$, which is impossible.

Exercise 16.3. *Derive Kolmogorov's zero-one law from Theorem 16.1.*

17 Basics of information theory

17.1 Shannon entropy

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Let X be a simple random variable taking values in some measurable space $(\Theta, \mathscr{G})$. We define the *Shannon entropy* of X by

$$H(X) = -\sum_{\theta \in \Theta} \mathbb{P}[X = \theta] \log \mathbb{P}[X = \theta],$$

where we use the convention $0 \log 0 = 0$. Since X is simple, there is some finite subset of Θ, which we will denote by $\operatorname{supp} X$, for which $\mathbb{P}[X = \theta] > 0$. Furthermore, $\sum_{\theta \ in \Theta} \mathbb{P}[X = \theta] = 1$.

Denote by $\mathbb{P}[X]$ the random variable given by $\mathbb{P}[X](\omega) = \mathbb{P}[X = X(\omega)]$. Then we can write the entropy as

$$H(X) = \mathbb{E}[-\log \mathbb{P}[X]].$$

Exercise 17.1. *Show that if $|\operatorname{supp} X| = n$ then $H(X) \le \log n$, with equality iff X is distributed uniformly on its support. Hint: use Jensen's inequality, and the ℓ^1-ℓ^2 inequality, which states that for every $z \in \mathbb{R}^n$ it holds that $\|x\|_1 \le \sqrt{n}\|x\|_2$.*

The first important property of Shannon entropy is the following form of monotonicity:

Claim 17.2. *Let X, Y be simple random variables. Suppose Y is $\sigma(X)$-measurable (i.e., $Y = f(X)$ for some function f). Then $H(Y) \le H(X)$.*

Proof. Note that $\mathbb{P}[Y] \le \mathbb{P}[X]$ almost surely. Hence

$$H(Y) = \mathbb{E}[-\log \mathbb{P}[Y]] \le \mathbb{E}[-\log \mathbb{P}[X]] = H(X).$$

□

Given two random variables X and X' taking values in Θ, Θ', we can consider the pair (X, X') as a single random variable taking values in $\Theta \times \Theta'$. We denote the entropy of this random variable as $H(X, X')$. The second important property of Shannon entropy is additivity with respect to independent random variables.

Claim 17.3. *Let X, Y be independent simple random variables. Then $H(X, Y) = H(X) + H(Y)$.*

Proof. By independence, $\mathbb{P}[X, Y] = \mathbb{P}[X] \cdot \mathbb{P}[Y]$. Hence

$$H(X, Y) = \mathbb{E}[-\log \mathbb{P}[X, Y]] = \mathbb{E}[-\log \mathbb{P}[X] - \log \mathbb{P}[Y]] = H(X) + H(Y).$$

□

17.2 Conditional Shannon entropy

Let $\mathcal{G}$ be a sub-sigma-algebra of $\mathcal{F}$. For a simple random variable X, define the random variable $\mathbb{P}[X|\mathcal{G}](\omega) = \mathbb{P}[X = X(\omega)|\mathcal{G}](\omega)$, and denote the conditional Shannon entropy by

$$H(X|\mathcal{G}) = \mathbb{E}[-\log \mathbb{P}[X|\mathcal{G}]].$$

For a simple random variable X and any random variable Y, we denote $H(X|Y) = H(X|\sigma(Y))$.

Claim 17.4. *$H(X|\mathcal{G}) \le H(X)$, with equality if and only if X is independent of $\mathcal{G}$.*

Proof. By the law of total expectation, $\mathbb{P}[X|\mathcal{G}] = \mathbb{E}[\mathbb{P}[X]|\mathcal{G}]$. Since $x \mapsto -\log(x)$ is a convex function, it follows from Jensen's inequality that

$$\begin{aligned} H(X|\mathcal{G}) &= \mathbb{E}[-\log \mathbb{P}[X|\mathcal{G}]] \\ &= \mathbb{E}[-\log \mathbb{E}[\mathbb{P}[X]|\mathcal{G}]] \\ &\le \mathbb{E}[\mathbb{E}[-\log \mathbb{P}[X]|\mathcal{G}]] \\ &= \mathbb{E}[-\log \mathbb{P}[X]] \\ &= H(X). \end{aligned}$$

When X is independent of $\mathcal{G}$, $\mathbb{P}[X] = \mathbb{P}[X|\mathcal{G}]$, and we therefore have equality. It thus remains to be shown if X is not independent of $\mathcal{G}$ then the inequality is strict. Indeed, in that case $\mathbb{P}[X] \neq \mathbb{P}[X|\mathcal{G}]$ with positive probability, and thus Jensen's inequality is strict with positive probability, from which it follows that our inequality is also strict. □

The same proof shows more generally that if $\mathcal{G}_1 \subseteq \mathcal{G}_2$ then $H(X|\mathcal{G}_1) \ge H(X|\mathcal{G}_2)$.

Exercise 17.5. *Suppose $\mathcal{G} = \cap_{i=n}^{\infty} \mathcal{G}_n$, and $\mathcal{G}_{n+1} \subseteq \mathcal{G}_n$. Then*

$$H(X|\mathcal{G}) = \lim_n H(X|\mathcal{G}_n) = \sup_n H(X|\mathcal{G}_n).$$

17.3 Mutual information

We denote the *mutual information* of X and $\mathcal{G}$ by $I(X;\mathcal{G}) = H(X) - H(X|\mathcal{G})$. By the above, I is non-negative, and is equal to 0 if and only if X is independent of $\mathcal{G}$. For two random variables X, Y, we denote $I(X;Y) = I(X;\sigma(Y))$.

Claim 17.6. *Let X, Y be simple random variables. Then*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = I(Y;X).$$

Proof. By definition,

$$I(X;Y) = \mathbb{E}[-\log \mathbb{P}[X]] - \mathbb{E}[-\log \mathbb{P}[X|Y]]$$

By Bayes' Law, $\mathbb{P}[X|Y]\mathbb{P}[Y] = \mathbb{P}[X,Y]$. Hence $\log\mathbb{P}[X|Y] = \log\mathbb{P}[X,Y] - \log\mathbb{P}[Y]$, and

$$\begin{aligned} I(X;Y) &= \mathbb{E}[-\log\mathbb{P}[X]] - \mathbb{E}[-\log\mathbb{P}[X,Y] + \log\mathbb{P}[Y]] \\ &= \mathbb{E}[-\log\mathbb{P}[X]] - \mathbb{E}[-\log\mathbb{P}[X,Y]] + \mathbb{E}[-\log\mathbb{P}[Y]] \\ &= H(X) - H(X,Y) + H(Y). \end{aligned}$$

□

It follows that

$$H(X|Y) = H(X) - I(X;Y) = H(X) - I(Y;X) = H(X) + H(Y|X) - H(Y),$$

and so

$$H(X|Y) = H(Y|X) - H(Y) + H(X). \tag{17.1}$$

17.4 The information processing inequality

Let $X_1, X_2, X_3, \ldots$ be a Markov chain, with each X_n simple.

Claim 17.7. *$I(X_3;X_1,X_2) = I(X_3;X_2)$. Likewise, for $m > n$, $I(X_n;\sigma(X_m,X_{m+1},\ldots)) = I(X_n;X_m)$.*

The claim is a consequence of the fact that by the Markov property, $\mathbb{P}[X_3|X_1,X_2] = \mathbb{P}[X_3|X_2]$.

18 Random walks on groups

18.1 Finitely generated groups

Let G be a group. We will denote the group operation by $g \cdot h$, or just gh. A subset S of G is said to *generate* G if each $g \in G$ is equal to a product of elements in S. We say that G is *finitely generated* if it admits a finite generating set S. Note that finitely generated groups are countable.

Example 18.1.

- $G = \mathbb{Z}$*, where the operation is addition, and* $S = \{-1, 1\}$.
- $\mathrm{SL}(2, \mathbb{Z})$*: the integer matrices with determinant* 1*, with the operation of matrix multiplication. It is not immediate that the following is a generating set:*

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

- *The free group generated by* $\{a, b, a^{-1}, b^{-1}\}$.

18.2 Random walks on finitely generated groups

Let G be finitely generated by S. Let $X_1, X_2, \ldots$ be i.i.d. random variables taking value in G, and such that $\mathbb{P}[X_n = g] > 0$ iff $g \in S$. Denote by $\mu(g) = \mathbb{P}[X_n = g]$ the distribution of X_n. Let $Z_n = X_1 \cdot X_2 \cdots X_n$. Note that Z is a Markov chain, with transition matrix $P(g, h) = \mu(g^{-1}h)$.

18.3 Random walk entropy

Let $h_n = \frac{1}{n} H(X_1, \ldots, X_n)$.

Claim 18.2. $H(Z_{n+m}) \leq H(Z_n) + H(Z_m)$.

Proof.

$$Z_{n+m} = (X_1 \cdots X_n) \cdot (X_{n+1} \cdots X_{n+m}),$$

and so

$$H(Z_{n+m}) \leq H(X_1 \cdots X_n, X_{n+1} \cdots X_{n+m}).$$

These two random variables are independent, and so

$$H(Z_{n+m}) \leq H(X_1 \cdots X_n) + H(X_{n+1} \cdots X_{n+m}).$$

The distribution of $Z_m = X_1 \cdots X_m$ is identical to that of $X_{n+1} \cdots X_{n+m}$, and so

$$H(Z_{n+m}) \leq H(Z_n) + H(Z_m).$$

□

This claim shows that the sequence $H(Z_n)$ is *subadditive*. Fekete's Subadditive Lemma states that if $(a_n)_n$ is a subadditive sequence then $\frac{a_n}{n}$ converges, and that furthermore

$$\lim_n \frac{a_n}{n} = \lim_n a_{n+1} - a_n.$$

. We accordingly define the *random walk entropy* h_μ by

$$h(\mu) = \lim_{n\to\infty} \frac{1}{n} H(Z_n).$$

Note that $\frac{1}{n}H(Z_n) \le \frac{1}{n}H(X_1,\dots,X_n) = H(X_1)$, and thus $h(\mu)$ is finite.

18.4 The Kaimanovich-Vershik Theorem

Theorem 18.3. *The Markov chain $Z_1, Z_2, \dots$ has a trivial tail sigma-algebra if and only if $h(\mu) = 0$.*

Proof. We calculate the mutual information $I(Z_1; \mathcal{T})$, where $\mathcal{T}$ is the tail sigma-algebra. Recall that $\mathcal{T} = \cap_n \mathcal{T}_n$, where $\mathcal{T}_n = \sigma(Z_n, Z_{n+1}, \dots)$. Hence, by Exercise 17.5,

$$H(Z_1|\mathcal{T}) = \lim_n H(Z_1|Z_n, Z_{n+1}, \dots).$$

By the Markov property it follows that

$$H(Z_1|\mathcal{T}) = \lim_n H(Z_1|Z_n).$$

By (17.1)

$$H(Z_1|\mathcal{T}) = \lim_n H(Z_n|Z_1) - H(Z_n) + H(Z_1).$$

Now, $Z_1 = X_1$, and $Z_n = X_1 \cdots X_n$, and so

$$H(Z_1|\mathcal{T}) = \lim_n H(X_1 \cdots X_n|X_1) - H(Z_n) + H(Z_1).$$

Note that conditioned on $X_1 = g$, the distribution of $X_1 \cdots X_n$ is identical to the distribution of $gX_1 \cdots X_{n-1}$, which has the same entropy as $X_1, \dots, X_{n-1} = Z_{n-1}$. Hence $H(X_1 \cdots X_n|X_1) = H(Z_{n-1})$, and we get that

$$H(Z_1|\mathcal{T}) = \lim_n H(Z_{n-1}) - H(Z_n) + H(Z_1).$$

Thus

$$I(Z_1; \mathcal{T}) = \lim_n H(Z_n) - H(Z_{n-1}) = h(\mu).$$

It follows that if $h(\mu) > 0$ then $\mathcal{T}$ is not independent of Z_1, and in particular $\mathcal{T}$ is non-trivial.

For the other direction, a calculation similar to the one above shows that $I(Z_1, \dots, Z_n; \mathcal{T}) = nh(\mu)$. Thus, if $h(\mu) = 0$, then $\mathcal{T}$ is independent of $(Z_1, \dots, Z_n)$ for all n, and, as in the proof of Kolmogorov's zero-one law, is trivial. □

For a finitely generated group G with generating set S, we denote by $|g|$ the minimum number of elements of S whose product is equal to g. It is easy to see that $|gh| \le |g| + |h|$. We denote $B_r = \{g \in G : |g| \le r\}$. We say that G has *subexponential growth* if $|B_r|$ is smaller than any exponent. That is, if $\lim_r \frac{1}{r} \log |B_r| = 0$.

Corollary 18.4. *If G has subexponential growth then $\mathcal{T}$ is trivial.*

Proof. Since Z_n is supported on B_r, $H(Z_n) \le \log |B_n|$. Hence

$$h(\mu) = \lim_n \frac{1}{n} H(Z_n) \le \lim_n \frac{1}{n} \log |B_n|.$$

Hence if G is subexponential then $h(\mu) = 0$ and $\mathcal{T}$ is trivial. □

19 Characteristic functions and the Central Limit Theorem

19.1 Convergence in distribution

Let $(X, X_1, X_2, \ldots)$ be real random variables. Denote their c.d.f.s by $F, F_1, F_2, \ldots$. We say that X_n *converges in distribution* to X if $F_n(x)$ converges to $F(x)$ for every x that is a continuity point of F.

Claim 19.1. *The following are equivalent.*

1. X_n *converges in distribution to* X.
2. $\lim_n \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$ *for every bounded continuous* $h : \mathbb{R} \to \mathbb{R}$.

19.2 The characteristic function

Let X be a real random variables. The *characteristic function* $\varphi_X : \mathbb{R} \to \mathbb{C}$ of X is given by

$$\varphi_X(t) = \mathbb{E}\left[e^{itX}\right] = \mathbb{E}[\cos(tX)] + i \cdot \mathbb{E}[\sin(tX)].$$

This expectation exists for any real random variable X and any real t, since the sine and cosine functions are bounded.

Note that

$$\begin{aligned}\varphi_{aX+b}(t) &= \mathbb{E}\left[e^{it(aX+b)}\right] \\ &= \mathbb{E}\left[e^{itaX} \cdot e^{itb}\right] \\ &= \varphi_{aX} \cdot e^{itb}.\end{aligned}$$

Exercise 19.2. *φ_X is continuous, and is differentiable n times if $X \in \mathscr{L}^n$. In this case $\varphi_X^{(n)}(0) = i^n \cdot \mathbb{E}[X^n]$.*

If X and Y are independent, then

$$\begin{aligned}\varphi_{X+Y}(t) &= \mathbb{E}\left[e^{it(X+Y)}\right] \\ &= \mathbb{E}\left[e^{itX}\right] \cdot \mathbb{E}\left[e^{itY}\right] \\ &= \varphi_X(t) \cdot \varphi_Y(t).\end{aligned}$$

A real random variable X is said to have a *probability distribution function* (or p.d.f.) $f_X : \mathbb{R} \to \mathbb{R}$ if for any measurable $h : \mathbb{R} \to \mathbb{R}$ it holds that

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\,\mathrm{d}x,$$

whenever the l.h.s. exists. In this case the cumulative distribution function

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x,$$

so that f is the derivative of F. Also,

$$\varphi_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{itx} f_X(x)\,\mathrm{d}x,$$

So that φ_X is the Fourier transform of f_X.

We saw in Claim 3.8 that F uniquely determines the distribution of X.

Theorem 19.3 (Lévy's Inversion Formula)**.** *Let X be a real random variable. For every $b > a$ such that $\mathbb{P}[X = a] = \mathbb{P}[X = b] = 0$ it holds that*

$$F(b) - F(a) = \lim_{T\to\infty} \int_{-T}^{T} \frac{\mathrm{e}^{-ita} - \mathrm{e}^{-itb}}{it} \varphi_X(t)\,\mathrm{d}t.$$

Since there are at most countably many $c \in \mathbb{R}$ such that $\mathbb{P}[X = c] > 0$, F is determined by φ_X.

Theorem 19.4 (Lévy's Continuity Theorem)**.** *Let $(X, X_1, X_2, \ldots)$ be real random variables. Then the following are equivalent:*

1. *X_n converges in distribution to X.*
2. *$\varphi_{X_n}(t)$ converges to $\varphi_X(t)$ for every $t \in \mathbb{R}$.*

19.3 The characteristic function of normalized i.i.d. sums

Let X be a standard Gaussian (or normal) random variable. This is a real random variable with p.d.f. $f_X(x) = \frac{1}{2\pi}\mathrm{e}^{-x^2/2}$. It is easy to calculate that

$$\varphi_X(t) = \mathrm{e}^{-\frac{1}{2}t^2}.$$

Thus if X_1 and X_2 are independent standard Gaussian then

$$\varphi_{(X_1+X_2)/\sqrt{2}}(t) = \mathrm{e}^{-\frac{1}{2}t^2},$$

and more generally the same holds for $(X_1 + \cdots + X_n)/\sqrt{n}$.

If $(X_1, X_2, \ldots)$ are (not necessarily Gaussian) i.i.d. and $Y_n = \sum_{k\leq n} X_k$ then

$$\varphi_{Y_n}(t) = \varphi_{X_n}(t)^n.$$

If we define

$$Z_n = \frac{1}{\sqrt{n}} Y_n = \frac{1}{\sqrt{n}} \sum_{k\leq n} X_n$$

then

$$\varphi_{Z_n}(t) = \varphi_X(t/\sqrt{n})^n.$$

Now, let $\mathbb{E}[X] = 0$ and $\mathbb{E}\left[X^2\right] = 1$. Since $X \in \mathscr{L}^2$ then φ_X is twice differentiable and

$$\varphi_X(0) = 1 \qquad \varphi'(0) = 0 \qquad \varphi''(0) = 1.$$

It is an exercise to show that it follows that

$$\varphi_X(t) = 1 - \frac{1}{2}t^2 + o(t^2),$$

where here we mean by $o(t^2)$ that as $t \to 0$ it holds that

$$|\varphi_X(t) - 1 - \frac{1}{2}t^2| \cdot t^2 \to 0.$$

Thus we have that

$$\varphi_{Z_n}(t) = \varphi_X(t/\sqrt{n})^n = (1 - \tfrac{1}{2}t^2/n + o(t^2/n^2))^n,$$

and thus

$$\lim_n \varphi_{Z_n}(t) = \mathrm{e}^{-\frac{1}{2}t^2}.$$

As we know, $\mathrm{e}^{-\frac{1}{2}t^2}$ is the characteristic function of a standard Gaussian. Thus we have proved that if G is a standard Gaussian then for any $t \in \mathbb{R}$ it holds that

$$\mathbb{E}\left[\mathrm{e}^{itZ_n}\right] \to \mathbb{E}\left[\mathrm{e}^{itG}\right].$$

Hence φ_{Z_n} converges pointwise to φ_G. Using Lévy's Continuity Theorem, we have thus proved the Central Limit Theorem:

Theorem 19.5. *Let $(X_1, X_2, \ldots)$ be i.i.d. real random variables with $\mathbb{E}[X_n] = 0$ and $\mathbb{E}\left[X_n^2\right] = 1$. Then the sequence $Z_n = \frac{1}{\sqrt{n}} \sum_{k \le n} X_n$ converges in distribution to a standard Gaussian.*

20 The Radon-Nikodym derivative and absolute continuity

20.1 The Radon-Nikodym derivative

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Given a non-negative r.v. X with $\mathbb{E}[X] = 1$, we can define a the measure $\mathbb{Q} = X \cdot \mathbb{P}$ by

$$\mathbb{Q}[A] = \mathbb{E}\left[\mathbb{1}_{\{A\}} \cdot X\right].$$

It is easy to show that X is the unique r.v. such that $\mathbb{Q} = X \cdot \mathbb{P}$.

In this case we call X the *Radon-Nikodym derivative* of $\mathbb{Q}$ with respect to $\mathbb{P}$, and denote

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(\omega) = X(\omega).$$

20.2 Absolute continuity

Note that

$$\mathbb{P}[A] = 0 \quad \text{implies} \quad \mathbb{Q}[A] = 0, \tag{20.1}$$

so that not every measure $\mathbb{Q}$ can be written as $X \cdot \mathbb{P}$ for some X. When $\mathbb{Q}$ and $\mathbb{P}$ satisfy (20.1) then we say that $\mathbb{Q}$ is *absolutely continuous relative* to $\mathbb{P}$.

Example 20.1.

- *The uniform distribution on* $[0,1]$ *is absolutely continuous relative to the uniform distribution on* $[0,2]$.
- *If* $\mathbb{P}[A] > 0$ *then* $\mathbb{P}[\cdot|A]$ *is absolutely continuous relative to* $\mathbb{P}$.
- *The point mass* $\delta_{1/2}$ *is not absolutely continuous relative to the uniform distribution on* $[0,1]$.
- *The i.i.d. q measure on* $\{0,1\}^{\mathbb{N}}$ *is not absolutely continuous relative to the i.i.d. p measure on* $\{0,1\}^{\mathbb{N}}$*, unless* $p = q$.

Lemma 20.2. *If* $\mathbb{Q}$ *is absolutely continuous relative to* $\mathbb{P}$*, then for each* $\varepsilon > 0$ *there exists a* $\delta > 0$ *such that, for every measurable A,* $\mathbb{P}[A] < \delta$ *implies* $\mathbb{Q}[A] < \varepsilon$.

Proof. Assume the contrary, so that there is some ε and a sequence of events $(A_1, A_2, \ldots)$ with $\mathbb{P}[A_n] < 2^{-n}$ and $\mathbb{Q}[A_n] \geq \varepsilon$. Let $A = \cap_n \cup_{m>n} A_m$ be the event that infinitely many of these events occur. Then by Borel-Cantelli $\mathbb{P}[A] = 0$. On the other hand $\mathbb{Q}[A] \geq \varepsilon$, in contradiction to absolute continuity. □

20.3 The Radon-Nikodym Theorem

Recall that $\mathscr{F}$ is *separable* if it generated by a countable subset $\{F_1, F_2, \ldots\}$. We can assume w.l.o.g. that this subset is a π-system.

Theorem 20.3 (Radon-Nikodym Theorem)**.** *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space with $\mathscr{F}$ separable, and let $\mathbb{Q}$ be absolutely continuous relative to $\mathbb{P}$. Then there exists a r.v. X such that $\mathbb{Q} = X \cdot \mathbb{P}$.*

Proof. Let $\mathscr{F}_n = \sigma(F_1, \ldots, F_n)$. Then $\mathscr{F}_n$ is a finite sigma-algebra, and as such is the set of all possible unions of $\{B_1^n, \ldots, B_k^n\}$, a finite partition of Ω. Define the $\mathscr{F}_n$-measurable r.v. X_n as follows. For a given $\omega \in \Omega$ there is a unique $B^n \in \{B_1^n, \ldots, B_k^n\}$ such that $\omega \in B^n$. Set

$$X_n(\omega) = X_n(B^n) = \frac{\mathbb{Q}[B^n]}{\mathbb{P}[B^n]},$$

where we take $0/0 = 0$. It is easy to verify that $\mathbb{E}[X_n] = 1$, and that for every $B \in \mathscr{F}_n$ it holds that

$$\mathbb{Q}[B] = \mathbb{E}\left[\mathbb{1}_{\{B\}} \cdot X_n\right],$$

so that on $\mathscr{F}_n$ it holds that X_n is the Radon-Nikodym derivative $\mathrm{d}\mathbb{Q}/\mathrm{d}\mathbb{P}$.

Now, since $(\mathscr{F}_1, \mathscr{F}_2, \ldots)$ is a filtration, each element in B^n is the disjoint union of (at most) two sets B_i^{n+1} and B_j^{n+1}. Hence

$$\begin{aligned}
\mathbb{E}[X_{n+1}|\mathscr{F}_n](\omega) &= \frac{X_{n+1}(B_i^{n+1}) \cdot \mathbb{P}\left[B_i^{n+1}\right] + X_{n+1}(B_j^{n+1}) \cdot \mathbb{P}\left[B_j^{n+1}\right]}{\mathbb{P}\left[B_i^{n+1}\right] + \mathbb{P}\left[B_j^{n+1}\right]} \\
&= \frac{\frac{\mathbb{Q}[B_i^{n+1}]}{\mathbb{P}[B_i^{n+1}]} \cdot \mathbb{P}\left[B_i^{n+1}\right] + \frac{\mathbb{Q}[B_j^{n+1}]}{\mathbb{P}[B_j^{n+1}]} \cdot \mathbb{P}\left[B_j^{n+1}\right]}{\mathbb{P}[B^n]} \\
&= \frac{\mathbb{Q}[B_i^{n+1}] + \mathbb{Q}[B_j^{n+1}]}{\mathbb{P}[B^n]} \\
&= \frac{\mathbb{Q}[B^n]}{\mathbb{P}[B^n]} \\
&= X_n(\omega),
\end{aligned}$$

and thus $(X_1, X_2, \ldots)$ is a martingale w.r.t. the filtration $(\mathscr{F}_1, \mathscr{F}_2, \ldots)$. Since it is non-negative then it converges almost surely to some r.v. X.

We now claim that $(X_1, X_2, \ldots)$ are *uniformly integrable*, in the sense that for every ε there exists a K such that for all n it holds that

$$\mathbb{E}\left[X_n \cdot \mathbb{1}_{\{X_n > K\}}\right] < \varepsilon.$$

To see this, recall that $\mathbb{E}[X_n] = 1$, note that X_n is non-negative, and apply Markov's inequality to arrive at

$$\mathbb{P}[X_n > K] < \frac{1}{K}.$$

Now, by Lemma 20.2, if we choose K large enough then this implies that $\mathbb{Q}[X_n > K] < \varepsilon$. But the event $\{X_n > K\}$ is in $\mathscr{F}_n$, since X_n is $\mathscr{F}_n$-measurable. Hence

$$\mathbb{E}\left[X_n \cdot \mathbb{1}_{\{X_n > K\}}\right] = \mathbb{Q}[X_n > K] < \varepsilon.$$

This proves that $(X_1, X_2, \ldots)$ are uniformly integrable. An important result (which is not hard but which we will not prove) is that if $X_n \to X$ almost surely, then uniform integrability implies that this convergence is also in $\mathscr{L}^1$, in the sense that $\mathbb{E}[|X_n - X|] \to 0$. It follows that for any $F_i \in \{F_1, F_2, \ldots\}$

$$\lim_n \mathbb{E}\left[\mathbb{1}_{\{F_i\}} \cdot (X_n - X)\right] \leq \lim_n \mathbb{E}\left[\mathbb{1}_{\{F_i\}} \cdot |X_n - X|\right] = 0,$$

and thus

$$\mathbb{E}\left[\mathbb{1}_{\{F_i\}} \cdot X\right] = \lim_n \mathbb{E}\left[\mathbb{1}_{\{F_i\}} \cdot X_n\right] = \mathbb{Q}[F_i].$$

Thus the measure $X \cdot \mathbb{P}$ agrees with $\mathbb{Q}$ on the generating algebra $\{F_1, F_2, \ldots\}$, and thus $\mathbb{Q} = X \cdot \mathbb{P}$. □

21 Large deviations

Let $(X_1, X_2, \ldots)$ be i.i.d. real random variables. Denote $X = X_1$ and let $\mu = \mathbb{E}[X]$. Let

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

By the law of large numbers we expect that Z_n should be close to μ for large n. What is the probability that it is larger than some $\eta > \mu$? We already proved the Chernoff lower bound. We here prove an asymptotically matching upper bound.

21.1 The cumulant generating function

Recall that the *moment generating function* of X is

$$M(t) = \mathbb{E}\left[e^{tX}\right],$$

and that its *cumulant generating function* is

$$K(t) = \log M(t) = \log \mathbb{E}\left[e^{tX}\right].$$

Of course, these may be infinite for some t. Let I, the domain of both, be the set on which they are finite, and note that $0 \in I$.

Claim 21.1. *I is an interval, and K is convex on I.*

For the proof of this claim we will need Hölder's inequality. For $p \in [1, \infty]$ and a real r.v. X denote

$$|X|_p = \mathbb{E}\left[|X|^p\right]^{1/p}.$$

Lemma 21.2 (Hölder's inequality)**.** *For any $p, q \in [1, \infty]$ with $1/p + 1/q = 1$ and r.v.s X, Y it holds that*

$$|X \cdot Y|_1 \le |X|_p \cdot |Y|_q .$$

Exercise 21.3. *Prove Hölder's inequality. Hint: use Young's inequality, which states that for every real $x, y \ge 0$ and $p, q > 1$ with $1/p + 1/q = 1$ it holds that*

$$xy \le \frac{x^p}{p} + \frac{y^q}{q}.$$

Proof of Claim 21.1. Assume $a, b \in I$. Then for any $r \in (0,1)$

$$K(ra + (1-r)b) = \log \mathbb{E}\left[e^{(ra+(1-r)b)X}\right] = \log \mathbb{E}\left[\left(e^{aX}\right)^r \left(e^{bX}\right)^{1-r}\right].$$

By Hölder's inequality

$$\begin{aligned} K(ra+(1-r)b) &\le \log \mathbb{E}\left[\left(\mathrm{e}^{aX}\right)^r\right]^{1/r} + \log \mathbb{E}\left[\left(\mathrm{e}^{bX}\right)^{1-r}\right]^{1/(1-r)} \\ &= \log \mathbb{E}\left[\mathrm{e}^{aX}\right]^r + \log \mathbb{E}\left[\mathrm{e}^{bX}\right]^{1-r} \\ &= r \log \mathbb{E}\left[\mathrm{e}^{aX}\right] + (1-r)\log \mathbb{E}\left[\mathrm{e}^{bX}\right] \\ &= rK(a) + (1-r)K(b). \end{aligned}$$

Since K is non-negative it follows that it is finite on $ra+(1-r)b$, and thus I is an interval on which it is convex. □

Applying the Dominated Convergence Theorem inductively can be used to show that M and K are smooth (i.e., infinitely differentiable) on the interior of I.

21.2 The Legendre transform

Let the *Legendre transform* of K be given by

$$K^\star(\eta) = \sup_{t>0}(t\eta - K(t)).$$

It turns out that the fact that K is smooth and convex implies that $K^\star$ is also smooth and convex. Therefore, if the supremum in this definition is obtained at some t, then $K'(t) = \eta$. Conversely, if $K'(t) = \eta$ for some t, then this t is unique and $K^\star(\eta) = t\eta - K(t)$.

Theorem 21.4 (Chernoff bound)**.**

$$\mathbb{P}\left[Z_n \ge \eta\right] \le \mathrm{e}^{-K^\star(\eta)n}.$$

Proof. For any $t \ge 0$

$$\begin{aligned} \mathbb{P}\left[Z_n \ge \eta\right] &\le \mathbb{P}\left[tZ_n \ge t\eta\right] \\ &= \mathbb{P}\left[\mathrm{e}^{t\sum_{k\le n} X_k} \ge \mathrm{e}^{t\eta n}\right] \\ &\le \frac{\mathbb{E}\left[\mathrm{e}^{\sum_{k\le n} tX_k}\right]}{\mathrm{e}^{t\eta n}} \\ &= \mathrm{e}^{-(t\eta - K(t))n}. \end{aligned}$$

Optimizing over t yields the claim. □

21.3 Large deviations

Theorem 21.5. *If $\eta = K'(t)$ for some t in the interior of I then*

$$\mathbb{P}\left[Z_n \ge \eta\right] = \mathrm{e}^{-K^\star(\eta)n + o(n)}.$$

Proof. One side is given by the Chernoff bound. It thus remains to prove the upper bound. Let $Z_n = \sum_{k=1}^n X_k$. We want to prove that

$$\mathbb{P}\left[Z_n \geq \eta n\right] = \mathrm{e}^{-K^\star(\eta)n + o(n)}.$$

For a given n and $t \geq 0$, define the measure $\tilde{\mathbb{P}}$ by

$$\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}} = \mathrm{e}^{ntZ_n - nK(t)}.$$

Under this measure $(X_1, \ldots, X_n)$ are still i.i.d., and that their distribution does not depend on n:

$$\begin{aligned}
\tilde{\mathbb{P}}[X_1 \in A_1, \ldots, X_n \in A_n] &= \tilde{\mathbb{E}}\left[\prod_{i=1}^n \mathbb{1}_{\{X_i \in A_i\}}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{ntZ_n - nK(t)} \prod_{i=1}^n \mathbb{1}_{\{X_i \in A_i\}}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{t(X_1 + \cdots + X_n) - nK(t)} \prod_{i=1}^n \mathbb{1}_{\{X_i \in A_i\}}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n \mathbb{1}_{\{X_i \in A_i\}} \mathrm{e}^{tX_i - K(t)}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[\mathbb{1}_{\{X_i \in A_i\}} \mathrm{e}^{tX_i - K(t)}\right].
\end{aligned}$$

Using the fact that the expectation of a random variable is equal to the derivative at zero of its cumulant generating function, a simple calculation shows that

$$\tilde{\mathbb{E}}[X_1] = \tilde{\mathbb{E}}[Z_n] = K'(t).$$

Choose any $\eta < \overline{\eta}$ and t such that $K'(t) \in (\eta, \overline{\eta})$, so that $\tilde{\mathbb{E}}[Z_n] = K'(t) \in (\eta, \overline{\eta})$. Then

$$\begin{aligned}
\mathbb{P}\left[\eta \leq Z_n\right] &\geq \mathbb{P}\left[\eta \leq Z_n \leq \overline{\eta}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\{\eta \leq Z_n \leq \overline{\eta}\}}\right] \\
&= \tilde{\mathbb{E}}\left[\mathbb{1}_{\{\eta \leq Z_n \leq \overline{\eta}\}} \mathrm{e}^{-(ntZ_n - nK(t))}\right] \\
&\geq \tilde{\mathbb{E}}\left[\mathbb{1}_{\{\eta \leq Z_n \leq \overline{\eta}\}} \mathrm{e}^{-(nt\overline{\eta} - nK(t))}\right] \\
&= \mathrm{e}^{-(t\overline{\eta} - K(t))n} \tilde{\mathbb{E}}\left[\mathbb{1}_{\{\eta \leq Z_n \leq \overline{\eta}\}}\right] \\
&= \mathrm{e}^{-(t\overline{\eta} - K(t))n} \tilde{\mathbb{P}}\left[\eta \leq Z_n \leq \overline{\eta}\right].
\end{aligned}$$

Since $\tilde{\mathbb{E}}[Z_n] \in (\eta, \overline{\eta})$, and since $(X_1, \ldots, X_n)$ are i.i.d. under $\tilde{\mathbb{P}}$ with a distribution that does not depend on n, by the law of large numbers,

$$\tilde{\mathbb{P}}[\eta \leq Z_n \leq \overline{\eta}] \to_n 1,$$

and so

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left[\eta\le Z_n\right]\ge-(t\overline{\eta}-K(t)).$$

Since this holds for any $\overline{\eta}>\eta$ and $\overline{\eta}>K'(t)>\eta$, it also holds for $\overline{\eta}=\eta$ and t^* such that $K'(t^*)=\eta$. So

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbb{P}\left[\eta\le Z_n\right]\ge-(t^*\eta-K(^*t))$$

or

$$\mathbb{P}\left[\eta\le Z_n\right]\ge\mathrm{e}^{-(t^*\eta-K(t^*))n+o(n)}.$$

Finally, since K is convex and smooth, and since $K'(t^*)=\eta$, then t is the maximizer of $t\eta-K(t)$, and thus $t^*\eta-K(t^*)=K^\star(\eta)$. We have thus shown that

$$\mathbb{P}\left[\eta\le Z_n\right]\ge\mathrm{e}^{-K^\star(\eta)n+o(n)}.$$

□

22 Stationary distributions and processes

Given a transition matrix P on some state space S, and given a Markov chain $(X_1, X_2, \ldots)$ over this P, the law of X_2 is given by

$$\begin{aligned}\mathbb{P}[X_2 = t] &= \sum_s \mathbb{P}[X_1 = s, X_2 = t] \\ &= \sum_s \mathbb{P}[X_1 = s]\mathbb{P}[X_2 = t | X_1 = s] \\ &= \sum_s \mathbb{P}[X_1 = s]P(s,t).\end{aligned}$$

Thus, if we think of the distributions of X_1 and X_2 as vectors $v_1, v_2 \in \ell^1(S)$, then we have that $v_2 = v_1 P$.

A non-negative left eigenvector of P is called a *stationary distribution* of P. It corresponds to a distribution of X_1 that induces the same distribution on X_2. By the Perron-Frobenius Theorem, if S is finite then P has a stationary distribution. Furthermore, if P is also irreducible then this distribution is unique.

Exercise 22.1. *The uniform distribution on $\mathbb{Z}/n\mathbb{Z}$ is the unique stationary distribution of the μ random walk (recall that μ is generating).*

Let $(Y_1, Y_2, \ldots)$ be a general process. We say that this process is *stationary* (or *shift-invariant*) if its law is the same as the law of $(Y_2, Y_3, \ldots)$. Equivalently, for every n, the law of $(Y_{k+1}, \ldots, Y_{k+n})$ is independent of k.

Exercise 22.2. *Show that the two definitions are indeed equivalent.*

Claim 22.3. *If $(Y_1, Y_2, \ldots)$ is a Markov chain, and if the distribution of Y_1 is stationary, then $(Y_1, Y_2, \ldots)$ is a stationary process.*

Returning to our scenery reconstruction problem, we can use what we learned above to deduce that $(Z_1, Z_2, \ldots)$ is a stationary process. It easily follows that

$$(F_1, F_2, \ldots) = (f(Z_1), f(Z_2), \ldots)$$

is also a stationary process.

23 Stationary processes and measure preserving transformations

We say that a stationary process $(Y_1, Y_2, \ldots)$ is *ergodic* if its shift-invariant sigma-algebra is trivial. That is, if for every shift-invariant event A it holds that $\mathbb{P}[A] \in \{0, 1\}$.

Some examples:

- An i.i.d. process is obviously stationary. By Kolmogorov's zero-one law its tail sigma-algebra is trivial, and so its shift-invariant sigma-algebra is also trivial. Thus it is ergodic.

- Let $(Y_1, Y_2, \ldots)$ be binary random variables such that

  $$\mathbb{P}[(Y_1, Y_2, \ldots) = (1, 1, \ldots)] = 1/2$$

  and

  $$\mathbb{P}[(Y_1, Y_2, \ldots) = (0, 0, \ldots)] = 1/2.$$

  This process is stationary but not ergodic; the event $\lim_n Y_n = 1$ is shift-invariant and has probability 1/2.

- Let $(Y_1, Y_2, \ldots)$ be binary random variables such that

  $$\mathbb{P}[(Y_1, Y_2, \ldots) = (1, 0, 1, 0, \ldots)] = 1/2$$

  and

  $$\mathbb{P}[(Y_1, Y_2, \ldots) = (0, 1, 0, 1, \ldots)] = 1/2.$$

  This process is stationary and ergodic.

- Let P be chosen uniformly over $[0, 1]$, and let $(Y_1, Y_2, \ldots)$ be binary random variables, which conditioned on P are i.i.d. Bernoulli with parameter P. This process is stationary but not ergodic. For example, the event that

  $$\lim_n \frac{1}{n} \sum_{k \le n} Y_k \le 1/2$$

  is a shift-invariant event that has probability 1/2.

- Let $(Y_1, Y_2, \ldots)$ be a Markov chain, with the distribution of Y_1 equal to some stationary distribution. Then this process is stationary. It is ergodic iff the distribution of Y_1 is not a non-trivial convex combination of two different stationary distributions.

- Let Y_1 be distributed uniformly on $[0, 1)$. Fix some $0 < \alpha < 1$, and let $Y_{n+1} = Y_n + \alpha \mod 1$. This is a stationary process, and it is ergodic iff α is irrational. We will show this later.

A generalization of the last example is the following. Let $(\Omega, \mathscr{F}, \nu)$ be a probability space, and let $T\colon \Omega \to \Omega$ be a measurable transformation that *preserves* ν. That is, $\nu(A) = \nu(T^{-1}(A))$ for all $A \in \mathscr{F}$. We say that $A \in \mathscr{F}$ is T-invariant if $T^{-1}(A) = A$, and note that the collection of T-invariant sets is a sub-sigma-algebra. Let Y_1 have law ν, and let each $Y_{n+1} = T(Y_n)$. Then $(Y_1, Y_2, \ldots)$ is a stationary process.

Claim 23.1. *$(Y_1, Y_2, \ldots)$ is ergodic iff for every T-invariant $A \in \mathscr{F}$ it holds that $\nu(A) \in \{0, 1\}$.*

Proof. The map $\pi\colon \Omega \to \Omega^{\mathbb{N}}$ given by $\pi(\omega) = (\omega, T(\omega), T^2(\omega), \ldots)$ is a bijection that pushes the measure ν to the law $\mathbb{P}$ of $(Y_1, Y_2, \ldots)$, and thus these two probability spaces are isomorphic. Furthermore, if we denote the shift by $\sigma\colon \Omega^{\mathbb{N}} \to \Omega^N$, then π is *equivariant*, in the sense that $\pi \circ T = \sigma \circ \pi$. It follows that the T-invariant sigma-algebra is mapped to the shift-invariant sigma-algebra, and thus one is trivial iff the other is trivial. □

Of course, if we have a process $(Y_1, Y_2, \ldots)$ taking values in Ω^n, then stationarity is precisely invariance w.r.t. the shift transformation $T\colon \Omega \to \Omega$ given by $T(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots)$. Thus stationary processes and measure preserving transformations are two manifestations of the same object. We say that T is ergodic if the T-invariant sigma-algebra is trivial. That is, if for every measurable A such that $T^{-1}(A) = A$ it holds that A has measure in $\{0, 1\}$.

Claim 23.2. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, with $T\colon \Omega \to \Omega$ an ergodic measure preserving transformation.*

If $Z\colon \Omega \to \mathbb{R}$ is a T-invariant random variable (i.e., $Z(\omega) = Z(T(\omega))$ for almost every $\omega \in \Omega$) then there is some $z \in \mathbb{R}$ such that $\mathbb{P}[Z = z] = 1$.

Exercise 23.3. *Prove this claim. Hint: If Z is T-invariant then for any $a < b \in \mathbb{R}$, the event $Z \in [a, b]$ is T-invariant, and thus has measure either 0 or 1.*

Consider the map $R_\alpha\colon S^1 \to S^1$ given by $R_\alpha(\mathrm{e}^{2\pi i z}) = \mathrm{e}^{2\pi i(z+\alpha)}$. This is a measure preserving transformation of S^1, equipped with the uniform measure.

Proposition 23.4. *R_α is ergodic iff α is irrational.*

Proof. If $\alpha = k/m$ is rational, then the set $\{\mathrm{e}^{2\pi i z} : z \in \cup_{n=0}^{m-1}[n/m, n/m + 1/2m]\}$ is R_α-invariant and has measure 1/2. Hence R_α is not ergodic.

If α is irrational, let $f\colon S^1 \to \{0, 1\}$ be the indicator of an R_α-invariant set. We can use the Fourier transform to write f as

$$f(z) = \sum_{k \in \mathbb{Z}} c_n \mathrm{e}^{2\pi i k z}$$

for some coefficients (c_n). Then $[R_\alpha f](z)$ is

$$[R_\alpha f](z) = \sum_{k \in \mathbb{Z}} c_k \mathrm{e}^{2\pi i k(z-\alpha)} = \sum_{k \in \mathbb{Z}} d_k \mathrm{e}^{2\pi i k z},$$

where $d_k = c_k \mathrm{e}^{-2\pi i k \alpha}$. Since A is R_α-invariant then $R_\alpha f = f$, and so $c_k = d_k$. Since α is irrational, $\mathrm{e}^{-2\pi i k \alpha} \neq 1$ unless $k = 0$, and so we have that $c_k = 0$ unless $k = 0$. Thus f is constant, and so it must be the indicator of a set of measure either 0 or 1. □

24 The Ergodic Theorem

Theorem 24.1 (The Pointwise Ergodic Theorem)**.** *Let* $(\Omega, \mathscr{F}, \mathbb{P})$ *be a probability space, with* $T\colon \Omega \to \Omega$ *a measure preserving transformation. If* T *is ergodic then for every* $X \in \mathscr{L}^1(\Omega, \mathscr{F}, \mathbb{P})$ *it holds that for* ν*-almost every* $\omega \in \Omega$

$$\lim_n \frac{1}{n} \sum_{k=0}^{n-1} X(T^k(\omega)) = \mathbb{E}[X].$$

In the language of stationary processes, one can say that if $(Y_1, Y_2, \ldots)$ is a stationary process with trivial shift-invariant sigma-algebra, and if $f(Y_1, Y_2, \ldots) \in \mathscr{L}^1$, then almost surely

$$\lim_n \frac{1}{n} \sum_{k=1}^{n} f(Y_k, Y_{k+1}, \ldots) = \mathbb{E}[f(Y_1, Y_2, \ldots)].$$

This Theorem was originally proved by Birkhoff [1]. We give a proof due to Katznelson and Weiss [2].

Proof. We assume without loss of generality that X is non-negative; otherwise apply the proof separately to X^+ and X^-. Define $X^*\colon \Omega \to \Omega$ by

$$X^*(\omega) = \lim_n \frac{1}{n} \sum_{k=0}^{n-1} X(T^k(\omega))$$

whenever this limit exists. We want to show that it exists w.p. 1, and that $\mathbb{P}[X^* = \mathbb{E}[X]] = 1$.

Define $\overline{X}\colon \Omega \to \mathbb{R}$ by

$$\overline{X}(\omega) = \limsup_n \frac{1}{n} \sum_{k=0}^{n-1} X(T^k(\omega)),$$

and likewise

$$\underline{X}(\omega) = \liminf_n \frac{1}{n} \sum_{k=0}^{n-1} X(T^k(\omega)).$$

Note that both are T-invariant, and so there are some $\overline{x}$ and $\underline{x}$ such that

$$\mathbb{P}\left[\overline{X} = \overline{x}, \underline{X} = \underline{x}\right] = 1.$$

Proving that

$$\overline{x} \le \mathbb{E}[X] \le \underline{x} \tag{24.1}$$

will thus finish the proof.

Fix some $\varepsilon > 0$. Let $N(\omega)$ be the first positive integer such that

$$\frac{1}{N(\omega)} \sum_{k=0}^{N(\omega)-1} X(T^k(\omega)) + \varepsilon \ge \overline{x} \tag{24.2}$$

Since $N(\omega)$ is a.s. finite, there is some $K \in \mathbb{N}$ such that the set $A = \{\omega : N(\omega) > K\}$ has measure less than $\varepsilon/\overline{x}$. Define

$$\tilde{X}(\omega) = \begin{cases} X(\omega) & \omega \notin A \\ \max\{X(\omega), \overline{x}\} & \omega \in A, \end{cases}$$

and also

$$\tilde{N}(\omega) = \begin{cases} N(\omega) & \omega \notin A \\ 1 & \omega \in A. \end{cases}$$

Note that in analogy to (24.2) we have that

$$\frac{1}{\tilde{N}(\omega)} \sum_{k=0}^{\tilde{N}(\omega)-1} \tilde{X}(T^k(\omega)) + \varepsilon \geq \overline{x},$$

or, rearranging, that

$$\sum_{k=0}^{\tilde{N}(\omega)-1} \tilde{X}(T^k(\omega)) \geq \tilde{N}(\omega)(\overline{x} - \varepsilon). \tag{24.3}$$

Now X and $\tilde{X}$ only differ on A, and when they do differ then it is at most by $\overline{x}$, since X is non-negative. Hence

$$\begin{aligned} \mathbb{E}\left[\tilde{X}\right] &= \mathbb{E}\left[X + (\tilde{X} - X)\right] \\ &= \mathbb{E}[X] + \mathbb{E}\left[\tilde{X} - X\right] \\ &= \mathbb{E}[X] + \mathbb{E}\left[(\tilde{X} - X) \cdot \mathbb{1}_{\{A\}}\right] \\ &\leq \mathbb{E}[X] + \mathbb{E}\left[\overline{x} \cdot \mathbb{1}_{\{A\}}\right] \\ &\leq \mathbb{E}[X] + \overline{x} \cdot \varepsilon/\overline{x} \\ &= \mathbb{E}[X] + \varepsilon. \end{aligned} \tag{24.4}$$

Now, let $L = K\overline{x}/\varepsilon$. For each $\omega \in \Omega$, let $\omega_0 = \omega$ and let

$$\omega_{j+1} = T^{\tilde{N}(\omega_j)}(\omega_j).$$

It follows that

$$\omega_j = T^{\tilde{N}(\omega_0) + \tilde{N}(\omega_1) + \cdots + \tilde{N}(\omega_{j-1})}(\omega).$$

Let $J(\omega)$ be the maximal j such that

$$\tilde{N}(\omega_0) + \tilde{N}(\omega_1) + \cdots + \tilde{N}(\omega_j) < L,$$

and let

$$\tilde{N}_L(\omega) = \tilde{N}(\omega_0) + \tilde{N}(\omega_1) + \cdots + \tilde{N}(\omega_{J(\omega)}).$$

Note that $\tilde{N}_L(\omega) > L-K$. Then we can write

$$\sum_{k=0}^{L-1} \tilde{X}(T^k(\omega)) = \sum_{k=0}^{\tilde{N}(\omega_0)} \tilde{X}(T^k(\omega_0)) + \cdots + \sum_{k=0}^{\tilde{N}(\omega_{J(\omega)})} \tilde{X}(T^k(\omega_{J(\omega)})) + \sum_{k=\tilde{N}_L(\omega)}^{L-1} \tilde{X}(T^k(\omega))$$

Applying (24.3) to each term but the last yields

$$\sum_{k=0}^{L-1} \tilde{X}(T^k(\omega)) \geq \tilde{N}_L(\omega)(\overline{x}-\varepsilon) + \sum_{k=\tilde{N}_L(\omega)}^{L-1} \tilde{X}(T^k(\omega)),$$

and using the fact that X is non-negative means

$$\sum_{k=0}^{L-1} \tilde{X}(T^k(\omega)) \geq \tilde{N}_L(\omega)(\overline{x}-\varepsilon).$$

Since $\tilde{N}_L(\omega) > L-K$ we can apply this estimate too, and, rearranging, arrive at

$$\frac{1}{L}\sum_{k=0}^{L-1} \tilde{X}(T^k(\omega)) \geq \overline{x} - \frac{K}{L}\overline{x} - \varepsilon$$

which by the choice of L we can write as

$$\frac{1}{L}\sum_{k=0}^{L-1} \tilde{X}(T^k(\omega)) \geq \overline{x} - 2\varepsilon.$$

Now, by T-invariance the expectation of the l.h.s. is just equal to the expectation of $\tilde{X}$. Hence

$$\mathbb{E}\left[\tilde{X}\right] \geq \overline{x} - 2\varepsilon.$$

Putting this together with (24.4) yields

$$\overline{x} \leq \mathbb{E}\left[\tilde{X}\right] + 2\varepsilon \leq \mathbb{E}[X] + 3\varepsilon,$$

and taking ε to zero yields $\overline{x} \leq \mathbb{E}[X]$. This completes the first half of the proof of (24.1); the second follows by a similar argument. □

Exercise 24.2. *Use the Ergodic Theorem to prove the strong law of large numbers.*

25 The weak topology and the simplex of invariant measures

Let X be a compact metrizable topological space. By the Riesz Representation Theorem we can identify $\mathscr{P}(X)$, the set of probability measures on X, with the positive bounded linear functionals on $C(X)$ that assign 1 to the constant function 1. The space X^* of bounded linear functionals on $C(X)$ comes equipped with the compact, metrizable weak* topology, under which $\varphi_n \to \varphi$ if $\varphi_n(f) \to \varphi(f)$ for all $f \in C(X)$. The restriction of this topology to the (closed) set of probability measures yields what probabilists call the weak topology on the probability measures on X.

In the important case that $X = \{0,1\}^{\mathbb{N}}$ we have that $\nu_n \to \nu$ weakly if for every clopen A it holds that $\nu_n(A) \to \nu(A)$. In the case $X = \mathbb{R} \cup \{-\infty, \infty\}$ we have that $\nu_n \to \nu$ if $\limsup_n \nu_n(A) \leq \nu(A)$ for all closed A, or if $\liminf_n \nu_n(A) \geq \nu(A)$ for all open A.

Let $X = \{0,1\}^{\mathbb{Z}}$, and denote by $\mathscr{I}(X)$ the set of stationary (or shift-invariant) probability measures on X.

Claim 25.1. *$\mathscr{I}(X)$ is a closed subset of $\mathscr{P}(X)$.*

Proof. Denote the shift by $\sigma : X \to X$. Assume that ν_n is a sequence in $\mathscr{I}(X)$ that converges to some $\nu \in \mathscr{P}(X)$. We prove the claim by showing that ν is stationary.

Let A be a clopen subset of X. Then

$$\nu(A) = \lim_n \nu_n(A) = \lim_n \nu_n(\sigma(A)) = \nu(\sigma(A)),$$

where the last equality follows from the fact that A being clopen implies that $\sigma(A)$ is clopen. Thus ν is invariant on a generating sub-algebra of the sigma-algebra, and by a standard argument it is invariant. □

Clearly, $\mathscr{I}(X)$ is a convex set. The next proposition shows (a more general claim which implies) that its extreme points $\mathscr{I}_e(X)$ are the ergodic measures.

Proposition 25.2. *A T-invariant measure ν on $(\Omega, \mathscr{F})$ is ergodic iff it is extreme.*

Proof. Assume that ν is not ergodic. Then there is some T-invariant $A \in \mathscr{F}$ such that $p := \nu(A) \in (0,1)$. Let ν_1 be given by $\nu_1(B) = \nu(B \mid A) = \frac{1}{p}\nu(B \cap A)$, and let $\nu_2(B) = \nu(B \mid A^c)$. Then

$$\begin{aligned}
\nu_1(T^{-1}B) &= \frac{1}{p}\nu((T^{-1}(B)) \cap A)) \\
&= \frac{1}{p}\nu((T^{-1}(B)) \cap T^{-1}(A))) \\
&= \frac{1}{p}\nu((T^{-1}(B \cap A))) \\
&= \frac{1}{p}\nu(B \cap A) \\
&= \nu_1(B).
\end{aligned}$$

And thus ν_1 is T-invariant. The same argument applies to ν_2, since A^c is also T-invariant. Finally, $\nu = p\nu_1 + (1-p)\nu_2$.

For the other direction, assume $\nu = p\nu_1 + (1-p)\nu_2$ for some $p \in (0,1)$. Clearly, ν_1 is absolutely continuous relative to ν, and so we write $\nu_1 = X \cdot \nu$ for some $X \in \mathscr{L}^1(\nu)$.

We now claim that X is T-invariant; we prove this for the case that T is invertible (although it is true in general). In this case, for any $A \in \mathscr{F}$

$$\begin{aligned}\nu_1(A) = \nu_1(T(A)) &= \int_\Omega \mathbb{1}_{\{A\}}(T^{-1}(\omega)) \cdot X(\omega)\, \mathrm{d}\nu(\omega) \\ &= \int_\Omega \mathbb{1}_{\{A\}}(T^{-1}(\omega)) \cdot X(\omega)\, \mathrm{d}\nu(T^{-1}(\omega)) \\ &= \int_\Omega \mathbb{1}_{\{A\}}(\omega) \cdot X(T(\omega))\, \mathrm{d}\nu(\omega),\end{aligned}$$

and so $X \circ T$ is also a Radon-Nikodym derivative $\mathrm{d}\nu_1/\mathrm{d}\nu$. But by the uniqueness of this derivative X and $X \circ T$ agree almost everywhere. It is a now nice exercise to show that there exists some X' that is equal to X almost everywhere and is T-invariant. It then follows by Claim 23.2, and by the fact that $\mathbb{E}[X] = 1$, that $\mathbb{P}[X = 1] = 1$, and thus $\nu = \nu_1$. □

This Theorem has an interesting consequence.

Exercise 25.3. *Assume ν, μ are both T-invariant ergodic measures on $(\Omega, \mathscr{F})$. Show that there exist two disjoint set $A, B \in \mathscr{F}$ such that $\nu(A) = 0$ and $\mu(A) = 1$, while $\nu(B) = 1$ and $\mu(B) = 0$.*

Thus ν and μ "live in different places."

In fact, it is possible to show that there is a map $\beta : \mathscr{I}_e(X) \to \mathscr{F}$ with the properties that

1. $\mu(\beta_\mu) = 1$ for all $\mu \in \mathscr{I}_e(X)$.
2. For all $\mu \neq \nu \in \mathscr{I}_e(X)$ it holds that $\beta_\mu \cap \beta_\nu = \emptyset$.

Using this, it is possible to show that $\mathscr{I}(X)$ is in fact a *simplex*: a compact convex set in which there is a *unique* way to write each element as a convex *integral* of the extreme points.

Proposition 25.4. *The ergodic measures $\mathscr{I}_e(X)$ are dense in $\mathscr{I}(X)$.*

Thus the simplex $\mathscr{I}(X)$ has the interesting property that its extreme points are dense. It turns out that there is only one such simplex (up to affine homeomorphisms), which is called the *Poulsen simplex*.

Proof. It suffices to show that for $\nu, \mu \in \mathscr{I}_e(X)$ and $\theta = \frac{1}{2}\nu + \frac{1}{2}\mu$ it is possible to find $\theta_n \in \mathscr{I}_e(X)$ s.t. $\lim_n \theta_n = \theta$.

To this end, fix n and define θ_n as follows. Let the law of the r.v.s $(X_k)_{k\in\mathbb{Z}}$ be μ, and the law of $(Y_k)_{k\in\mathbb{Z}}$ be ν. For $m \in \mathbb{Z}$, let $(X_0^m, \ldots, X_{n-1}^m)$ be independent of all previously defined random variables, and with law equal to that of $(X_0, \ldots, X_{n-1})$. Define $(Y_0^m, \ldots, Y_{n-1}^m)$ analogously.

Define $(W_k)_{k\in\mathbb{Z}}$ by

$$W_k = \begin{cases} X^{\lfloor k/n \rfloor}_{k \bmod n} & \text{if } \lfloor k/n \rfloor \text{ is even} \\ Y^{\lfloor k/n \rfloor}_{k \bmod n} & \text{if } \lfloor k/n \rfloor \text{ is odd.} \end{cases}$$

Finally, choose N uniformly at random from $\{0, 1, \ldots, 2n-1\}$, and define $(Z_k)_{k\in\mathbb{Z}}$ by

$$Z_k = W_{k+N}.$$

Let θ_n be the law of (Z_k).

It is straightforward (if tedious) to verify that (Z_k) is stationary. We leave it as an exercise to show that it is ergodic. Thus to finish the proof we have to show that $\lim_n \theta_n = \theta$.

Fix $M \in \mathbb{N}$, and consider the event that $N \in \{1, \ldots, M\}$. As n tends to infinity, the probability of this event tends to zero. Thus, if we condition on N, with probability that tends to 1/2 we have that the law of $(Z_1, \ldots, Z_M)$ is equal to the law of $(X_1, \ldots, X_M)$, and likewise for $(Y_1, \ldots, Y_M)$. This completes the proof. □

26 Percolation

Let $\mathcal{V}$ be a countable set, and let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a locally finite, simple symmetric graph. That is, $\mathcal{E}$ is a symmetric relation on $\mathcal{V}$ with $\mathcal{E} \cap \{v\} \times \mathcal{V}$ finite for each $v \in \mathcal{V}$. We also assume that $\mathcal{G}$ is connected, so that the transitive closure of $\mathcal{E}$ is $\mathcal{V} \times \mathcal{V}$.

The *i.i.d. p percolation measure on* $\{0,1\}^{\mathcal{E}}$ is simply the product Bernoulli measure, in which we choose each edge independently with probability p. We will denote this measure by $\mathbb{P}_p[\cdot]$, and will denote by E the random edge set with this law. $G = (\mathcal{V}, E)$ will be the corresponding random graph.

Note that G will in general not be connected. For each $v \in \mathcal{V}$ we denote by $K(v)$ the (random) connected component that v belongs to in G. We denote by $\{v \leftrightarrow \infty\}$ the event that $K(v)$ is infinite. We denote by K_∞ the event that there is some v for which $K(v)$ is infinite.

Claim 26.1. *The probability of K_∞ is either 0 or 1. In the former case, for every $v \in \mathcal{V}$, $\mathbb{P}_p[v \leftrightarrow \infty] = 0$, while in the latter $\mathbb{P}_p[v \leftrightarrow \infty] > 0$.*

Proof. Enumerate $\mathcal{E} = (e_1, e_2, \ldots)$, and let $A_n = \{e_i \in E\}$. Then $(A_1, A_2, \ldots)$ is an i.i.d. sequence. Clearly K_∞ is $\sigma(A_1, A_2, \ldots)$-measurable, and also clearly it is a tail event. Hence the first part of the claim follows by Kolmogorov's 0-1 law.

Since the event K_∞ contains $\{v \leftrightarrow \infty\}$, it is immediate that $\mathbb{P}_p[K_\infty] = 0$ implies $\mathbb{P}_p[v \leftrightarrow \infty] = 0$. Assume now that $\mathbb{P}_p[K_\infty] = 1$. Then there is some $w \in \mathcal{V}$ such that, with positive probability, $\mathbb{P}_p[w \leftrightarrow \infty]$. Let $P = (e_1, e_2, \ldots, e_n)$ be a path between v and w.

Consider the random variable $\tilde{E}$ taking values in $\{0,1\}^{\mathcal{E}}$ defined as follows: for every edge $e \notin P$, we set $e \in \tilde{E}$ iff $e \in E$. And we set $e \in \tilde{E}$ for all $e \in P$. We (you) prove in the exercise below that the law of $\tilde{E}$ is absolutely continuous relative to the law of E.

Denote $\tilde{G} = (\mathcal{V}, \tilde{E})$, and denote by $\tilde{K}(v)$ the connected component of v in $\tilde{G}$. Now, $\tilde{K}(v) = \tilde{K}(w)$, since v and w are connected in $\tilde{G}$. Also, $\tilde{K}(w)$ contains $K(w)$, since $\tilde{E}$ contains E. Hence the event $\{|\tilde{K}(w)| = \infty\}$ occurs with positive probability, and so the same holds for $\tilde{K}(v) = \tilde{K}(w)$. Finally, by absolute continuity, the same holds for $K(v)$, and so $\mathbb{P}_p[v \leftrightarrow \infty] > 0$. □

Exercise 26.2. *Prove that the law of $\tilde{E}$ is absolutely continuous relative to the law of E.*

Claim 26.3. *If $q > p$ then $\mathbb{P}_q[K_\infty] \geq \mathbb{P}_p[K_\infty]$.*

To prove this claim we prove a stronger theorem, and in the process introduce the technique of *coupling*. Let $\Omega = \{0,1\}^{\mathbb{N}}$, with $\mathcal{F}$ the Borel sigma-algebra. We consider the natural partial order on Ω given by $\omega \geq \omega'$ if $\omega_n \geq \omega'_n$ for all $n \in \mathbb{N}$. We say that $A \in \mathcal{F}$ is *increasing* if for all $\omega \geq \omega'$ it holds that $\omega' \in A$ implies $\omega \in A$. Let $\mathbb{P}_p[\cdot]$ denote the i.i.d. p measure on Ω.

Theorem 26.4. *If A is increasing then $q > p$ implies $\mathbb{P}_q[A] \geq \mathbb{P}_p[A]$.*

Proof. Let $(X_1, X_2, \ldots)$ be i.i.d. random variables, each distributed uniformly on $[0,1]$. For each n let $Q_n = \mathbb{1}_{\{X_n \leq q\}}$ and $P_n = \mathbb{1}_{\{X_n \leq p\}}$. Note that $\mathbb{P}[Q_n = 1] = q$ and $\mathbb{P}[P_n = 1] = p$, and

that $(Q_1, Q_2, \ldots)$ is i.i.d., as is $(P_1, P_2, \ldots)$. Hence the law of $(Q_1, Q_2, \ldots)$ (resp., $(P_1, P_2, \ldots)$) is $\mathbb{P}_q[\cdot]$ (resp., $\mathbb{P}_p[\cdot]$). Note also that

$$(Q_1, Q_2, \ldots) \geq (P_1, P_2, \ldots),$$

since $q > p$. Hence for any increasing event $A \subset \{0,1\}^{\mathbb{N}}$ it holds that $(P_1, P_2, \ldots) \in A$ implies $(Q_1, Q_2, \ldots) \in A$, and thus $\mathbb{P}_q[A] \leq \mathbb{P}_p[A]$. □

The construction in this proof is an example of *coupling*. Formally, a coupling of two probability spaces $(\Omega, \mathscr{F}, \mathbb{P})$ and $(\Omega', \mathscr{F}', \mathbb{P}')$, is a probability space $(\Omega \times \Omega', \sigma(\mathscr{F} \times \mathscr{F}'), \mathbb{Q})$ such that the projections on the two coordinates pushes $\mathbb{Q}$ forward to $\mathbb{P}$ and $\mathbb{P}'$.

Since $\mathbb{P}_p[K_\infty] \in \{0,1\}$, since $\mathbb{P}_p[K_\infty]$ is weakly increasing in p, we are interested in the *critical percolation probability*

$$p_c = \sup\{p : \mathbb{P}_p[K_\infty] = 0\}.$$

An interesting (and often hard) question is whether $\mathbb{P}_{p_c}[K_\infty]$ is zero or one.

Let $\mathscr{G}$ be the infinite k-ary tree with root o. In this case we can calculate p_c, by noting that the event $\{o \leftrightarrow \infty\}$ can be thought of as the event that the Galton-Watson tree with children distribution $B(k,p)$ is infinite. We know that this happens with positive probability iff $p > 1/k$. Hence in this case $p_c = 1/k$, and $\mathbb{P}_{p_c}[K_\infty] = 0$.

27 The mass transport principle

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a locally finite, countable graph. A *graph automorphism* is a bijection $f : \mathcal{V} \to \mathcal{V}$ such that $(v,w) \in \mathcal{E}$ iff $(f(v), f(w)) \in \mathcal{E}$. The automorphisms of a graph form the group $\mathrm{Aut}(\mathcal{G})$ under composition. We say that $\mathcal{G}$ is *transitive* if its automorphism group acts on it transitively. That is, if for all $v, w \in \mathcal{V}$ there is a graph automorphism f s.t. $f(v) = w$. Intuitively, this means that the geometry of the graph "looks the same" from the point of view of every vertex.

An important example is when Γ is finitely generated by a symmetric finite subset S, and $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \Gamma$ is the corresponding Cayley graph. In this case it is easy to see that the Γ action on itself is an action by graph automorphisms, which is furthermore already transitive. We will restrict our discussion to this setting, even though it all extends to *unimodular transitive graphs*; these are graphs with a unimodular automorphism group.

A map $f : \Gamma \times \Gamma \to [0, \infty)$ is a mass-transport if it is invariant under the diagonal Γ-action:

$$f(h,k) = f(gh, gk)$$

for all $g, h, k \in \Gamma$. It is useful to think about f as indicating how much "mass" is passed from h to k, where the amount passed can depend on identities of h and k, but in a way that (in some sense) only depends on the geometry of the graph and not on their names.

Theorem 27.1 (Mass Transport Principle for Groups)**.** *For every mass transport $f : \Gamma \times \Gamma \to [0, \infty)$ and $g \in \Gamma$ it holds that*

$$\sum_{k \in G} f(g,k) = \sum_{k \in G} f(k,g).$$

That is, the total mass flowing out of g is equal to the total mass flowing in.

Proof. By invariance

$$\sum_{k \in G} f(g,k) = \sum_{k \in G} f(k^{-1}g, e).$$

Changing variables to $h = k^{-1}g$ yields

$$= \sum_{h \in G} f(h, e).$$

Applying invariance again yields

$$\sum_{h \in G} f(gh, g)$$

and again changing variables to $k = gh$ yields the desired result. □

As an application, consider the following random subgraph E of the standard Cayley graph of $\mathbb{Z}^2$. For each $z, w_1, w_2 \in \mathbb{Z}^2$ such that $w_1 = z + (0,1)$ and $w_2 = z + (1,0)$, we independently set $(z, w_1) \in E$, $(z, w_2) \notin E$ w.p. 1/2, and $(z, w_1) \notin E$, $(z, w_2) \in E$ w.p. 1/2.

For distinct $z, w \in \mathbb{Z}^2$, we say that w is a descendant of z (and z is an ancestor of w) in E if there is a path between w and z, and if $w \leq z$ in both coordinates.

Note that, by construction,

1. E has no cycles, and each node is adjacent to at least one edge, and so E is a spanning forest.

2. Each w has infinitely many ancestors.

3. If $w \leq z$ then the number of descendants of w is independent of the number of descendants of z.

Proposition 27.2. *The number of descendants of each $w \in \mathbb{Z}$ is almost surely finite, with infinite expectation.*

Proof. Let $f(w,z)$ equal the probability that z is an ancestor of w, and note that by the invariance of the definitions f is a mass transport.

The sum $\sum_z f(w,z)$ is the expected number of ancestors of w, which is infinite, since w a.s. has infinitely many ancestors. It follows by the mass transport principle that $\sum_z f(z,w)$, the expected number of descendants of w, is likewise infinite.

It is easy to see that the expected number of direct descendants of any w is 1. By the independence property mentioned above, the process $(N_1, N_2, \ldots)$ - where N_k is the number of descendants at distance k from w - is a non-negative martingale. It thus converges, and moreover must converge to an integer, and so converges to 0. □

Note that we can define the same process on $\mathbb{Z}^d$, where now each vertex has d potential ancestors, with the proof applying as is. We can further generalize to groups that have a set S such that $S \cup S^{-1}$ generates Γ and the graph induced by S has no cycles.

Exercise 27.3. *On $\mathbb{Z}^d$, prove that E is a spanning tree iff $d \leq 3$.*

28 Majority dynamics

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a locally finite, countable (or finite) graph, and denote by $N(v)$ the neighbors of $v \in \mathcal{V}$. We would like $|N(v)|$ to be odd, and so we either add or remove v to $N(v)$ to achieve this.

Let $(V_1, V_2, \ldots)$ be random subsets of $\mathcal{V}$, with the property that if $v, w \in V_n$ then $(v, w) \notin E$; that is, each V_n is an independent set.

Let $\Theta = \{-1, +1\}^{\mathcal{V}}$, and consider the following sequence of random variables $(X_1, X_2, \ldots)$, each taking values in Θ. First, let X_1 be chosen i.i.d. 1/2. Given X_n, we define X_{n+1} as follows. For $v \notin V_n$ let $X_{n+1}(v) = X_n(v)$. For $v \in V_n$ let

$$X_{n+1}(v) = \operatorname{sgn} \sum_{w \in N(v)} X_n(w).$$

Note that since $|N(v)|$ is odd then the sum is never 0, and there is no ambiguity with taking the signum. This process is called majority dynamics, or zero temperature Glauber dynamics.

Proposition 28.1. *If $\mathcal{G}$ is finite then X_n converges (hence stabilizes) almost surely.*

Proof. Let

$$H_n = \sum_{(v,w) \in E} X_n(v) \cdot X_n(w),$$

so that H_n is the number of edges in the graph along which there is agreement, minus the number of edges along which there is disagreement.

Note that in majority dynamics, whenever a node changes its label and non of its neighbors do, the total number of disagreements in the graph strictly decreases, and thus H_n decreases. Since V_n is an independent set, changes at a node are always done while keeping its neighbors constant, and thus H_n decreases by at least 2 with each change in X_n. Since H_n is bounded from below by $-|E|$, X_n must stabilize. □

In fact, this proof can be generalized to the case that $\mathcal{G}$ is infinite, but with bounded degrees and subexponential growth. This is no longer true on general graphs:

Exercise 28.2. *Prove that on the 3-regular tree X_n does not in general stabilize.*

It is also not true if the V_n are not independent sets.

Let $\mathcal{G}$ be a Cayley graph of a finitely generated group Γ, and choose $(V_1, V_2, \ldots)$ from a distribution that is invariant to the Γ-action. For example, to choose V_n one can choose an independent uniform number for each node, and include in V_n only those nodes whose numbers are higher than all of their neighbors'.

Theorem 28.3. *In this setting X_n converges pointwise almost surely.*

Proof. Let

$$h_n = \sum_{g \in N(e)} \mathbb{E}[X_n(e) \cdot X_n(g)]$$

Now,

$$\begin{aligned} h_{n+1} - h_n &= \sum_{g \in N(e)} \mathbb{E}[X_{n+1}(e) \cdot X_{n+1}(g)] - \sum_{g \in N(e)} \mathbb{E}[X_n(e) \cdot X_n(g)] \\ &= \sum_{g \in N(e)} \mathbb{E}[X_{n+1}(e) \cdot X_{n+1}(g) - X_n(e) \cdot X_n(g)] \end{aligned}$$

If $e \notin V_n$ and also $g \notin V_n$ then the term in the expectation is zero. Hence

$$\begin{aligned} = \sum_{g \in N(e)} & \mathbb{E}\left[X_{n+1}(e) \cdot X_{n+1}(g) - X_n(e) \cdot X_n(g) \cdot \mathbb{1}_{\{e \in V_n\}}\right] \\ & + \mathbb{E}\left[X_{n+1}(e) \cdot X_{n+1}(g) - X_n(e) \cdot X_n(g) \cdot \mathbb{1}_{\{g \in V_n\}}\right], \end{aligned}$$

since the two conditioned events $\{e \in V_n\}$ and $\{g \in V_n\}$ are mutually exclusive. Furthermore, this implies

$$\begin{aligned} = \sum_{g \in N(e)} & \mathbb{E}\left[X_{n+1}(e) \cdot X_n(g) - X_n(e) \cdot X_n(g) \cdot \mathbb{1}_{\{e \in V_n\}}\right] \\ & + \mathbb{E}\left[X_n(e) \cdot X_{n+1}(g) - X_n(e) \cdot X_n(g) \cdot \mathbb{1}_{\{g \in V_n\}}\right], \end{aligned}$$

which by the mass transport principle

$$= 2 \sum_{g \in N(e)} \mathbb{E}\left[X_{n+1}(e) \cdot X_n(g) - X_n(e) \cdot X_n(g) \cdot \mathbb{1}_{\{e \in V_n\}}\right].$$

Rearranging yields

$$= 2\mathbb{E}\left[(X_{n+1}(e) - X_n(e)) \cdot \sum_{g \in N(e)} X_n(g) \cdot \mathbb{1}_{\{e \in V_n\}}\right],$$

and since, conditioned on $e \in V_n$, $X_{n+1}(e) = \operatorname{sgn} \sum_{g \in N(e)} X_n(g)$, then

$$= 2\mathbb{E}\left[2\mathbb{1}_{\{X_{n+1}(e) \neq X_n(e)\}} \cdot \left|\sum_{g \in N(e)} X_n(g)\right| \cdot \mathbb{1}_{\{e \in V_n\}}\right],$$

Now, by definition $\mathbb{1}_{\{X_{n+1}(e) \neq X_n(e)\}} \cdot \mathbb{1}_{\{e \in V_n\}} = \mathbb{1}_{\{X_{n+1}(e) \neq X_n(e)\}}$. Also, since $|N(e)|$ is odd, $|\sum_{g \in N(e)} X_n(g)| \geq 1$, and so

$$\geq 4\mathbb{P}[X_{n+1}(e) \neq X_n(e)].$$

Hence h_n is non-decreasing. Since it is bounded by $|N(e)|$ it converges to some $h_\infty < \infty$. Furthermore

$$h_\infty - h_1 \geq 4 \sum_{n=1}^{\infty} \mathbb{P}[X_{n+1}(e) \neq X_n(e)],$$

and so the expected number of n such that $X_{n+1}(e) \neq X_n(e)$ is finite, and in particular $X_n(e)$ stabilizes w.p. 1. By invariance this holds for every $X_n(g)$, and we have proved our claim. □

29 Scenery Reconstruction: I

Fix n, and let $(X = X_1, X_2, \ldots)$ be i.i.d. random variables on the abelian group $\mathbb{Z}/n\mathbb{Z}$. Denote by $\mu(k) = \mathbb{P}[X = k]$ their law. Let X_0 be uniformly distributed on $\mathbb{Z}/n\mathbb{Z}$, and let $Z_n = \sum_{k=0}^{n} X_n$ be the corresponding random walk. We assume throughout that the support of μ generates $\mathbb{Z}/n\mathbb{Z}$.

Some important examples to keep in mind:

- $\mu(1) = 1$.
- $\mu(1) = 1 - \varepsilon$, $\mu(2) = \varepsilon$.

Fix some $f \in \{0,1\}^n$, and let $F_n = f(Z_n)$. The law of $(F_1, F_2, \ldots)$ depends on f; we think of these distributions as a family indexed by f. We denote by $\mathbb{P}_f[\cdot]$ the distribution when we fix a particular f. Note that $\mathbb{P}_f[\cdot]$ does not change if we shift f.

Exercise 29.1. *Prove this.*

Denote by $[f]$ the equivalence class of f under shifts. That is, $f' \in [f]$ if there is some $k \in \mathbb{Z}/n\mathbb{Z}$ such that for every $m \in \mathbb{Z}/n\mathbb{Z}$ it holds that $f'(k+m) = f(m)$.

The question of scenery reconstruction is the following: is it possible to determine $[f]$ given $(F_1, F_2, \ldots)$? In particular we say that we can reconstruct f if there is some measurable

$$\hat{f} : \{0,1\}^{\mathbb{N}} \to \{0,1\}^n$$

such that for every $f \in \{0,1\}^n$ it holds that

$$\mathbb{P}_f\left[\hat{f}(F_1, F_2, \ldots) \in [f]\right] = 1. \tag{29.1}$$

Equivalently, if

$$\mathbb{P}\left[\hat{f}(f(Z_1), f(Z_2), \ldots) \in [f]\right] = 1.$$

In statistics, $\hat{f}$ is called an *estimator* of f, and the existence of such an $\hat{f}$ is called *identifiability* (of f). This clearly depends on μ, and so we say that μ is reconstructive if this holds.

One can reformulate (29.1) in finitary terms. It is equivalent to the existence of a sequence $(\hat{f}_1, \hat{f}_2, \ldots)$ with $\hat{f}_k$ being $\sigma(F_1, \ldots, F_k)$-measurable and with

$$\lim_k \mathbb{P}_f\left[\hat{f}_k(F_1, \ldots, F_k) \in [f]\right] = 1$$

for all $f \in \{0,1\}^n$.

A very interesting question is how quickly does this converge to one (when it does), for μ chosen *uniformly over* n; for example for $\mu(1) = 0.99$, $\mu(2) = 0.01$.

Question 29.2. *Let $N(n,\varepsilon)$ be the smallest k such that there is an $\hat{f}_k : \{0,1\}^k \to \{0,1\}^n$ with*

$$\mathbb{P}_f\left[\hat{f}_k(F_1,\ldots,F_k) \in [f]\right] \geq 1-\varepsilon$$

for all f. For fixed ε (say 1/3), how does $N(n,\varepsilon)$ grow with n?

This is not known; it is not even known if $N(\cdot,\varepsilon)$ is exponential or polynomial. The question of whether a given μ is reconstructive is much better understood.

Theorem 29.3. *Let n be a prime > 5, and let $\mu \in \mathbb{Q}^n$. Then μ is reconstructive iff $\varphi_\mu(k) \neq \varphi_\mu(m)$ for all $k \neq m$. Here φ_μ is given by*

$$\varphi_\mu(k) = \varphi_X(k) = \mathbb{E}\left[e^{\frac{2\pi i}{n}\cdot k\cdot X}\right] = \sum_{\ell \in \mathbb{Z}/n\mathbb{Z}} e^{\frac{2\pi i}{n}\cdot k\cdot X}\mu(k).$$

where $k \cdot X$ is multiplication mod n.

The first direction (the case that $\varphi_\mu(k) \neq \varphi_\mu(m)$ for all $k \neq m$) does not require the extra assumptions on n and μ. This is due to Matzinger and Lember [3].

To prove this theorem we will need to study a few new concepts.

30 Scenery reconstruction: II

Fix $n \in \mathbb{N}$, $f \in \{0,1\}^n$ and μ a generating probability measure on $\mathbb{Z}/n\mathbb{Z}$, and recall our process in which X_0 is uniform on $\mathbb{Z}/n\mathbb{Z}$, $(X_1, X_2, \ldots)$ are i.i.d. with law μ, $Z_n = X_0 + X_1 + \cdots + X_n$ and $F_n = f(Z_n)$. Recall also that are we are interested in guessing (correctly, almost surely) what $[f]$ is (the equivalence class of functions that are shifts of f) from a single random instance of $(F_1, F_2, \ldots)$.

Define the $a : \mathbb{Z}/n\mathbb{Z} \to \mathbb{R}$, *autocorrelation of* f by

$$a(k) = \frac{1}{n} \sum_{m=0}^{n} f(m) \cdot f(m+k),$$

and note that a is the same for any $f' \in [f]$. Imagine that we are willing to settle on reconstructing a rather than $[f]$. We will show that if the values of the characteristic function φ_μ are unique then we can reconstruct the $a(k)$'s.

To this end, we define $A : \mathbb{N} \to \mathbb{R}$, the *autocorrelation of* F by

$$\alpha_k = \mathbb{E}[F_T \cdot F_{T+k}]$$

for some $T \in \mathbb{N}$; by stationarity, the choice of T is immaterial. We will show that if we know the α_k's the we can infer the a_k's. But this will not help us, unless there some measurable $\hat{\alpha}_k : \{0,1\}^{\mathbb{N}} \to \mathbb{R}$ such that

$$\mathbb{P}_f[\hat{\alpha}_k(F_1, F_2, \ldots) = \alpha_k] = 1.$$

A natural candidate for $\hat{\alpha}_k$ is the empirical average; we take lim sup rather than lim to make sure $\hat{\alpha}_k$ is well defined:

$$\hat{\alpha}_k = \limsup_m \frac{1}{m} \sum_{T=1}^{m} F_T \cdot F_{T+k}.$$

A statement such as "$\hat{\alpha}_k = \alpha_k$ almost surely" sounds a lot like the strong law of large numbers. We will show later that this is indeed true, and that it follows from the *Ergodic Theorem*, which is a generalization of the SLLN.

Let $\mu * \mu$ be the *convolution* of μ with itself, which is given by

$$[\mu * \mu](k) = \sum_m \mu(k-m) \cdot \mu(m).$$

This is a probability distribution which is exactly the law of $X_1 + X_2$. Define analogously the k-fold convolution $\mu^{(k)}$, which is the law of $X_1 + \cdots + X_k$.

Claim 30.1. *For every $k \in \mathbb{N}$ it holds that*

$$\alpha_k = \sum_m \mu^{(k)}(m) \cdot a_m.$$

Proof. We set $T = 0$, condition on X_0 and Z_k and thus

$$\begin{aligned}\alpha_k &= \mathbb{E}[f(Z_0)\cdot f(Z_k)] \\ &= \sum_{m,\ell} \mathbb{E}[f(X_0)\cdot f(Z_k)|X_0 = \ell, Z_k = \ell + m]\cdot \mathbb{P}[X_0 = \ell, Z_k = \ell + m] \\ &= \sum_{m,\ell} f(\ell)\cdot f(\ell+m)\cdot \frac{1}{n}\cdot \mu^{(k)}(m) \\ &= \sum_m a_m \cdot \mu^{(k)}(m).\end{aligned}$$

□

It follows that if we denote by α the column vector $(\alpha_0,\ldots,\alpha_{n-1})$, by a the column vector $(a_0,\ldots,a_{n-1})$, and by M the $n\times n$ matrix $M_{k,m} = \mu^{(k)}(m)$ then $\alpha = Ma$. Assuming (as we will show later) that we can determine α, it follows that we can determine a if M is invertible.

Claim 30.2. *M is invertible iff the values of the characteristic function φ_μ are unique.*

Proof. We apply the Fourier transform to each row of M. Since the Fourier transform is an orthogonal linear transformation, the resulting matrix $\hat{M}$ is invertible iff M is invertible.

Now, over $\mathbb{Z}/n\mathbb{Z}$ the Fourier transform is identical to the characteristic function. Since the k^{th} row of M is the law of $X_1+\cdots+X_k$, the k^{th} row of $\hat{M}$ is given by

$$\varphi_{X_1+\cdots+X_k}(m) = \mathbb{E}\left[e^{\frac{2\pi i}{n}\cdot m\cdot(X_1+\cdots+X_k)}\right] = \varphi_X(m)^k.$$

Thus $\hat{M}$ is a Vandermonde matrix, and is invertible iff φ_X has unique values. □

Recall that we are interested in reconstructing $[f]$ rather than a. To this end we need to define the two-fold autocorrelation

$$a_{k,\ell} = \frac{1}{n}\sum_{m=0}^{n} f(m)\cdot f(m+k)\cdot f(m+k+\ell),$$

and its analogue

$$\alpha_{k,\ell} = \mathbb{E}[F_T\cdot F_{T+k}\cdot F_{T+k+\ell}].$$

It is then easy to show that there is also a linear relation between these two objects, with the corresponding matrix being $M\otimes M$, the tensor product of M with itself. This is invertible iff M is invertible, and so we get the same result. However, this still does not suffice, and we need to add still more indices and calculate n-fold autocorrelations. The appropriate matrices are again invertible iff M is, and moreover $[f]$ is uniquely determined by the n-fold autocorrelation.

A Homework problems

1. Let I be a set, and let $\{\mathscr{F}_i\}_{i\in I}$ be a collection of sigma-algebras of subsets of Ω. Show that $\cap_{i\in I}\mathscr{F}_i$ is a sigma-algebra.

2. Let $\mathscr{C}$ be a collection of subsets of Ω. Show that there exists a unique minimal (under inclusion) sigma-algebra $\mathscr{F}\supseteq\mathscr{C}$.

3. Given $n,k\in\mathbb{N}=\{1,2,\ldots\}$, denote by $A_{n,k}=\{nz+k : z\in\mathbb{Z}\}$ the set of all integers that are equal to k mod n. Let $\mathscr{P}=\{A_{n,k} : n,k\in\mathbb{N}\}\cup\{\emptyset\}$. Let $\mathscr{A}$ be the collection of finite unions of elements of $\mathscr{P}$.

   (a) Show that $\mathscr{P}$ is a π-system and that $\mathscr{A}$ is an algebra.

   (b) Find a finitely additive probability measure on $\mathscr{A}$.

   (c) Is $\mathscr{A}$ a sigma-algebra?

   (d) **Bonus.** Is your measure sigma-additive?

4. Consider $\mathscr{A}_{\text{clopen}}$, as defined in Example 2.5 in the lecture notes. Prove that there exists an additive $\mu_0\colon\mathscr{A}_{\text{clopen}}\to[0,1]$ with

$$\mu_0(A_x)=2^{-|x|}.$$

   Bonus: Prove that there exists a countably additive such μ_0, so that whenever $(A_1,A_2,\ldots)$ are disjoint elements of $\mathscr{A}_{\text{clopen}}$ with $\cup_n A_n\in\mathscr{A}_{\text{clopen}}$ then $\mu_0(\cup_n A_n)=\sum_n\mu_0(A_n)$.

5. Let $\Omega=\mathbb{Z}$ and $\mathscr{F}=2^{\mathbb{Z}}$ be the power set of $\mathbb{Z}$. A finitely additive probability measure $\mu\colon\mathscr{F}\to[0,1]$ is *shift-invariant* if for all $A\in\mathscr{F}$ it holds that $\mu(A)=\mu(A+1)$, where

$$A+1=\{n+1 : n\in A\}.$$

   (a) Prove that if μ is shift-invariant then it is not sigma-additive.

   (b) *Bonus.* Prove that there exists such a shift-invariant μ.

6. Recall that by the *Prime Number Theorem*, the probability that a number chosen uniformly from $\{1,2,\ldots,n\}$ is prime is of order $1/\log n$. Formally, if Y_n is distributed uniformly on $\{1,\ldots,n\}$, and if $\mathscr{P}$ denotes the set of primes, then

$$\lim_n\frac{\mathbb{P}[Y_n\in\mathscr{P}]}{1/\log n}=1.$$

   Inspired by this result (due to Jacques Hadamard and Charles Jean de la Vallée Poussin), we will choose a *random* subset P of the natural numbers that will resemble the primes in the sense described above.

Let $(X_3, X_4, \ldots)$ be a sequence of independent random variables taking values in $\{0,1\}$, whose distributions are given by $\mathbb{P}[X_n = 1] = p_n$ and $\mathbb{P}[X_n = 0] = 1 - p_n$, with

$$p_n = \frac{n+1}{\log(n+1)} - \frac{n}{\log n}$$

Let P be a random variable taking values in the space of subsets of $\mathbb{N}$ (which can be identified with $\{0,1\}^{\mathbb{N}}$), and given by

$$P = \{n : X_n = 1\}.$$

Recall that the (unproven) Goldbach conjecture states that every even number greater than 2 is the sum of two primes. A weaker (and still unproven) conjecture is that there is some N so that every even $n \geq N$ is the sum of two primes. We will show that our random P will satisfy an analogue of this conjecture.

(a) Let $Y_1, Y_2, \ldots$ be a sequence of random variables, each independent of P, and such that $\mathbb{P}[Y_n = i] = 1/n$ if $1 \leq i \leq n$ and $\mathbb{P}[Y_n = i] = 0$ otherwise. Prove that

$$\lim_n \frac{\mathbb{P}[Y_n \in P]}{1/\log n} = 1.$$

(b) Let G_n be the event that n is not the sum of any two elements of P. Show that $\mathbb{P}[G_n] < n^{-2}$ for all n large enough.[3]

(c) Let G be the event that there exists some integer K such that every $n \geq K$ is the sum of two elements of P. Using the Borel-Cantelli Lemma, show that $\mathbb{P}[G] = 1$.

7. Consider the sequence of independent random variables $X_1, X_2, \ldots$, where the cumulative distribution function of X_n is given by

$$\mathbb{P}[X_n \leq x] = \begin{cases} 2^{-n} \exp(2^{-n} x) & \text{if } x < 0 \\ 2^{-n} & \text{if } 0 \leq x < 0.01 \\ 1 & \text{if } 0.01 \leq x. \end{cases}$$

8. Consider a magical casino in which there is an infinite sequence of slot machines. A gambler starts out with 0 dollars in her account, and proceeds to gamble on each machine in turn. When she gambles on the n^{th} machine her total yield is X_n. Hence her balance at time n is S_n, where $S_0 = 0$ and

$$S_{n+1} = S_n + X_{n+1} = X_1 + \cdots + X_{n+1}.$$

.

[3] In fact, $\mathbb{P}[G_n]$ decreases much more rapidly than that, but this bound will suffice for our needs.

(a) Prove that if a random variable Y has a cumulative distribution function F, and if $F(x_0) - \lim_{x \nearrow x_0} F(x) = p$, then $\mathbb{P}[Y = x_0] = p$. Conclude that $\mathbb{P}[X_n = 0.01] = 1 - 2^{-n}$.

(b) Show that $\mathbb{E}[S_n] < 0$ for all $n > 0$. Use the definition of expectation and its properties as given in the lecture notes (as opposed to theorems not taught in this course).

(c) Show that $\mathbb{P}[\lim S_n = \infty] = 1$.

9. Prove the Dominated Convergence Theorem using the Monotone Convergence Theorem.

10. Consider the Bernoulli measure on $\{0,1\}^{\mathbb{N}}$ which is the unique extension (as we have shown in class / homework) of $\mu_0 \colon \mathscr{A}_{\text{clopen}} \to [0,1]$ with

$$\mu_0(A_x) = 2^{-|x|}.$$

(a) For $n \in \mathbb{N}$ let the real random variable $X_n : \Omega \to \mathbb{R}$ be given by $X_n(\omega) = \omega_n$. Show that X_n is indeed a random variable, and prove that the random variables $(X_1, X_2, \ldots)$ are independent.

(b) Let the real random variable $Y : \Omega \to \mathbb{R}$ be given by

$$Y(\omega) = \sum_{n=1}^{\infty} 2^{-n} \omega_n.$$

Prove that Y is indeed a random variable, and that its law is the uniform (Lebesgue) measure on $[0,1)$. (Hint: you can use the fact that the algebra of diadic intervals $[m2^{-n}, (m+1)2^{-n})$, $m < 2^n$ generates the Borel sigma-algebra on $[0,1)$.)

(c) Construct independent random variables $(Y_1, Y_2, \ldots)$, each with the uniform distribution on $[0,1)$.

11. Prove that there exists a *simply normal number*: a real number $x \in [0,1]$ such that for any $d > 1$ and $a \in \{0, \ldots, d-1\}$ the digit a occurs in the base d representation of x with asymptotic frequency $1/d$:

$$\lim_n \frac{\text{number of times } a \text{ occurs in the first } n \text{ digits of } x}{n} = \frac{1}{d}.$$

(Hint: Use the previous problem to show that this holds with probability 1 for X if it is chosen uniformly on $[0,1]$.)

12. *The Law of Total Expectation.* Let $X \in \mathscr{L}^2$. Show that if $\mathscr{G}_2 \subseteq \mathscr{G}_1$ then $\mathbb{E}[\mathbb{E}[X|\mathscr{G}_1]|\mathscr{G}_2] = \mathbb{E}[X|\mathscr{G}_2]$.

13. Prove that if $X, Y \in \mathscr{L}^1$ are independent, then $\mathbb{E}[X|Y]$ is the constant random variable $\mathbb{E}[X]$.

14. Find a sequence of independent random variables $(X_1, X_2, \ldots)$ with $\mathbb{P}[X_n \in \{-n, n, 0\}] = 1$, $\mathbb{E}[X_n] = 0$, and such that the weak LLN holds but not the strong: for $Y_n = \frac{1}{n}\sum_{k \le n} X_k$ it holds that $\mathbb{P}[|Y_n| \ge \varepsilon] \to 0$ but $\mathbb{P}[\lim_n Y_n = 0] \ne 1$.

15. *Elchanan Mossel's die paradox.* Toss a six sided die until 6 comes up, and then stop. Conditioned on all tosses coming out even, what is the expected number of tosses?

16. The *independent random walk* on $\mathbb{Z}^d$ is given by

$$P(x,y) = \begin{cases} 2^{-d} & \text{if } |x_i - y_i| = 1 \text{ for all } i \in \{1, \ldots, d\} \\ 0 & \text{otherwise.} \end{cases}$$

    Equivalently, the independent random walk on $\mathbb{Z}^d$ is a Markov process taking values on $\mathbb{Z}^d$ where in each coordinate the process is independent of the other coordinates, and equal to a simple random walk. For which d is the independent random walk transient, and for which is it recurrent?

17. Let $\mathcal{T}_d$ be the set of nodes of the undirected d-regular tree. This is the unique undirected graph with no cycles and in which each node has degree d.

    The simple random walk on $\mathcal{T}_d$ is the Markov chain $(X_0, X_1, \ldots)$ that starts from some $X_0 = s_0 \in \mathcal{T}_d$ and, at each step, transitions to each of the d neighboring nodes with probability $1/d$.

    (a) Prove that for $d \ge 3$, the simple random walk on $\mathcal{T}_d$ is transient.

    (b) Prove that for $d \ge 3$, the simple random walk on $\mathcal{T}_d$ has a non-trivial tail σ-algebra. Recall that this means that there is some event A that is in $\sigma(X_n, X_{n+1}, X_{n+2}, \ldots)$ for every n, and such that $\mathbb{P}[A] \in (0,1)$.

18. Let P be a transition matrix over a countable state space S. A probability measure μ on S is said to be P-stationary if for all $x \in S$

$$\sum_{y \in S} P(y,x)\mu(y) = \mu(x).$$

    (a) Prove that if μ is P-stationary, X_0 has distribution μ, and $(X_0, X_1, \ldots)$ is a Markov chain with transition matrix P, then each X_n has distribution μ.

    (b) Prove that if P is irreducible and transient then it has no stationary probability measures.

    (c) Consider the transition matrix P on the state space $\{0, 1, 2, \ldots\}$ given by

$$P(i,j) = \begin{cases} 0 & \text{if } |i-j| > 1 \\ 2/3 & \text{if } j = i-1 \text{ and } i > 0 \\ 1/3 & \text{if } j = i+1 \text{ and } i > 0 \\ 1 & \text{if } j = 1 \text{ and } i = 0. \end{cases}$$

    Prove that there exists a P-stationary probability measure.

19. Let $G = (V, E)$ be an undirected, finite graph: V is finite, and $E \subset V \times V$ satisfies $(v, w) \in E$ iff $(w, v) \in E$. Denote by $d(v) = |\{w : (v, w) \in E\}|$ the degree of $v \in V$. Assume that G is connected.

    Let $B \subseteq V$ be some non-empty subset of the vertices, which we will refer to as the boundary of the graph. Consider the Markov chain $X_1, X_2, \ldots$ with transition matrix P on the state space V given by

$$P(v, w) = \begin{cases} \frac{1}{d(v)} & \text{if } v \notin B \text{ and } (v, w) \in E \\ 1 & \text{if } v \in B \text{ and } w = v \\ 0 & \text{otherwise.} \end{cases}$$

    That is, on $V \setminus B$ the Markov chain moves to adjacent vertices with equal probabilities, and it stops once it reaches B.

    (a) Let $T_B = \min\{n > 0 : X_n \in B\}$ be the hitting time to B. Prove that it is almost surely finite.

    (b) Prove that $f(v) = \mathbb{E}_v[T_B]$ is P-superharmonic.

    (c) Prove that every function $f_B : B \to \mathbb{R}$ has a unique extension to a P-harmonic $f : V \to \mathbb{R}$. Hint: use T_B.

    (d) Suppose B consistes of two vertices: $B = \{b_0, b_1\}$. Let $f_B : B \to \mathbb{R}$ be given by $f_B(b_0) = 0$ and $f_B(b_1) = 1$. Let $f : V \to \mathbb{R}$ be the unique extension of f_B to a P-harmonic function. Prove that $f(v) = \mathbb{P}_v[X_{T_B} = b_1]$. That is, that $f(v)$ is the probability that the Markov chain that starts at v hits the boundary at b_1, rather than at b_0.

    (e) Let K, L be two positive integers. A gambler arrives at a casino with K dollars in her pocket. She plays until she runs out of money, or until she has $K + L$ dollars in her pocket. At each game she either loses a dollar or gains a dollar, each with probability 1/2. What is the probability that she leaves the casino with $K + L$ dollars?

20. Let X be a random variable with $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2 < \infty$. Assume that x is non-atomic, i.e., $\mathbb{P}[X = x] = 0$ for all $x \in \mathbb{R}$. Equivalently, the cumulative distribution function $F_X(x) = \mathbb{P}[X \le x]$ continuous. The median of X is the unique $m \in \mathbb{R}$ such that $F_X(m) = 1/2$. Show that $|\mu - m| \le \sigma$: the median is at most one standard deviation away from the mean.

21. Let X be a bounded random variable with mean 0. Let $X, X_1, X_2, \ldots$ be an i.i.d. sequence, and let $Z_n = \frac{1}{n}\sum_{k=1}^{n} X_k$.

    (a) Show that conditioned on a large deviation, the deviation is as small as possible, in the sense that for every $\eta > 0$ with $\mathbb{P}\left[X \ge \eta\right] > 0$, and for every $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left[\eta \le Z_n \le \eta + \varepsilon \middle| \eta \le Z_n\right] = 1.$$

(b) Suppose X is symmetric: $\mathbb{P}[X \le a] = \mathbb{P}[-X \le a]$ for all $a \in \mathbb{R}$.
Let $\eta > 0 > \zeta$ with $\mathbb{P}\left[X \ge \eta\right] > 0$ and $\mathbb{P}[X \le \zeta] > 0$. Let A_n be the event $\{Z_n \ge \eta \text{ or } Z_n \le \zeta\}$. Calculate the rate $\lim_n -\frac{1}{n}\log\mathbb{P}[A_n]$ in terms of $K_X^\star$, η and ζ.

22. Let G be a group generated by the finite, symmetric set S of size d. The transition matrix P of the associated simple random walk is given by

$$P(g,h) = \begin{cases} \frac{1}{d} & \text{if } h = gs \text{ for some } s \in S \\ 0 & \text{otherwise.} \end{cases}$$

Note that $P(g,h) = P(h,g)$, since S is symmetric. For $f : G \to \mathbb{R}$ we denote by Pf the function on G given by $[Pf][g] = \sum_h P(g,h)f(h) = \sum_h P(h,g)f(h)$. Note that $f \mapsto Pf$ is a linear operator.

For $f : G \to \mathbb{R}$ We denote $\|f\|_p = \left(\sum_g |f(g)|^p\right)^{1/p}$ and $\ell^p(G) = \{f : G \to \mathbb{R} : \|f\|_p < \infty\}$.

Given a subset $F \subseteq G$, we denote

$$\partial F = \{g \in F : gs \notin F \text{ for some } s \in S\}.$$

We say that G is *amenable* if

$$\inf_{\text{finite} F \subset G} \frac{|\partial F|}{|F|} = 0$$

(If you are interested, you can verify that amenability does not depend on the choice of generating set.)

(a) Prove that if $f \ge 0$ then $\|Pf\|_1 = \|f\|_1$.

(b) Prove that if $f \ge 0$ then $\|Pf\|_2 \le \|f\|_2$.
Hint: Write $P = \frac{1}{d}(P_1 + \cdots + P_d)$, where for each P_i there is an $s_i \in S$ such that $P_i(g,h) = 1$ iff $h = gs_i$. Then use the fact (which follows from the triangle inequality) that $\|\sum_i \alpha_i f_i\|_2 \le \sum_i |\alpha_i| \cdot \|f_i\|_2$.

(c) Prove that if G is amenable then for each $\varepsilon > 0$ there is an $f \in \ell^2(G)$ such that $\|Pf\|_2 \ge (1-\varepsilon)\|f\|_2$.

(d) Kesten's Theorem states that if G is nonamenable then there exists an $r < 1$ such that $\|Pf\|_2 < r\|f\|_2$. Use this theorem to prove that a simple random walk on a nonamenable group has positive random walk entropy.

23. **For your own amusement (please do not submit).** The *lamplighter group* $LL(\mathbb{Z})$ is defined as follows. It is the set of pairs (A,z), where A is a finite subset of $\mathbb{Z}$, and z is an element of Z. The group operation is $(A,z)\cdot(B,w) = (A \triangle (B+z), z+w)$.

(a) Prove that $LL(\mathbb{Z})$ is indeed a group.

(b) Prove that $LL(\mathbb{Z})$ is generated by $S = \{s_1, s_2, s_3\}$ where $s_1 = (\{0\}, 0)$, $s_2 = (\emptyset, 1)$ and $s_3 = (\emptyset, -1)$.

(c) Prove that $LL(\mathbb{Z})$ is amenable.

(d) Find a random walk supported on $S = \{s_1, s_2, s_3\}$ (not necessarily with the uniform distribution) that has positive random walk entropy.

24. Prove that for $0 < p \leq 1$ the law of $\tilde{E}$ is absolutely continuous relative to the law of E. (See the lecture notes chapter on percolation for the definitions of these random variables.)

25. Let $\mathcal{G}$ be the 3-regular tree. This is the unique connected graph with no cycles in which each vertex is adjacent to three edges. Show that for any p s.t. $p_c < p < 1$ it holds under $\mathbb{P}_p$ that, w.p. 1, there are infinitely many infinite connected components in the p percolation on $\mathcal{G}$.

26. Use the Ergodic Theorem to prove the strong law of large numbers: let $(X_1, X_2, \ldots)$ be i.i.d. random variables in $\mathscr{L}^1$. Then

$$\mathbb{P}\left[\lim_n \frac{1}{n}\sum_{k=1}^{n} X_k = \mathbb{E}[X_1]\right] = 1.$$

27. Prove the following claim.

    Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, with $T\colon \Omega \to \Omega$ an ergodic measure preserving transformation. If $Z\colon \Omega \to \mathbb{R}$ is a T-invariant random variable (i.e., $Z(\omega) = Z(T(\omega))$ for all $\omega \in \Omega$) then there is some $z \in \mathbb{R}$ such that $\mathbb{P}[Z = z] = 1$.

28. Prove that irrational rotations are ergodic transformations of $[0,1)$, equipped with the Lebesgue measure. Hint: use the fact that a random variable X on $[0,1)$ is uniquely determined by the values of its characteristic function $\varphi_X(k) = \mathbb{E}\left[e^{2\pi i k X}\right]$ for integer k.

29. *Bonus.* Show that $(Z_k)_{k\in\mathbb{Z}}$ is stationary and ergodic (see the lecture on the simplex of invariant measures for the definition).

30. Let G be a group, let $(X_1, X_2, \ldots)$ be i.i.d. random variables taking values in G. Let $Z_n = X_1 \cdot X_2 \cdots X_n$ be a random walk on G.

    Show if G is finite, and if the support of X_1 generates G, then the uniform distribution on G is the unique stationary distribution of this Markov chain.

    Recall that a probability measure on the state space S of a Markov chain with transition matrix P is said to be stationary if for all $x \in S$

$$\mu(x) = \sum_{y\in S} \mu(y)P(y,x).$$

31. Let $G = \langle a, b \rangle$ be the free group with two generators. Let $(X_1, X_2, \ldots)$ be i.i.d. and distributed uniformly on $\{a, b, a^{-1}, b^{-1}\}$, so that $Z_n = X_1 \cdots X_n$ is the simple random walk on G.

    (a) Show that with probability 1, the length of the word Z_n tends to infinity. Here the length of $g \in G$ is the length of its shortest representation as a product of the symbols in $\{a, b, a^{-1}, b^{-1}\}$.

    (b) Show that the Choquet-Deny Theorem does not apply to this (non-abelian) group: there is a bounded harmonic function, or, equivalently, the shift-invariant sigma-algebra is non-trivial.

32. *Bonus.* Recall that a random variable is an equivalence class of measurable functions that are equal almost everywhere. A random variable is $\mathscr{G}$-measurable if in its equivalence class there is a $\mathscr{G}$-measurable function.

    (a) Let P be the transition matrix of an irreducible Markov chain on a state space S such that $P(x,x) > \varepsilon$ for some $\varepsilon > 0$ and all $x \in S$. Show that a random variable is $\mathscr{T}$-measurable iff it is $\mathscr{I}$-measurable.

    (b) Show that, for a random walk on a countable abelian group, a random variable is $\mathscr{I}$-measurable iff it is $\mathscr{T}$-measurable.

33. *Kolmogorov from Choquet-Deny.* Let $(X_1, X_2, \ldots)$ be i.i.d. finitely supported random variables. Use the Choquet-Deny Theorem, together with 32b above, to show that their tail sigma-algebra is trivial.

References

[1] G. D. Birkhoff. Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

[2] Y. Katznelson and B. Weiss. A simple proof of some ergodic theorems. *Israel Journal of Mathematics*, 42(4):291–296, 1982.

[3] H. Matzinger and J. Lember. Reconstruction of periodic sceneries seen along a random walk. *Stochastic processes and their applications*, 116(11):1584–1599, 2006.