

# LN on Advanced Probability Theory

Vladislav Kargin

Fall 2019

# Contents

<b>1</b>	<b>Foundations</b>	<b>4</b>
1.1	Probability spaces, measures and $\sigma$ -fields . . . . .	5
1.1.1	Intuition . . . . .	5
1.1.2	Probability Space . . . . .	5
1.1.3	The Caratéodory Theorem . . . . .	8
1.1.4	Product measures and Kolmogorov's extension theorem . . . . .	11
1.1.5	Absolute continuity and singularity of measures . . . . .	13
1.2	Random Variables . . . . .	15
1.2.1	Random variables as measurable functions . . . . .	15
1.2.2	Generation of $\sigma$ -algebras . . . . .	17
1.2.3	Distributions . . . . .	17
1.2.4	Expected value . . . . .	20
1.2.5	Independence . . . . .	22
1.2.6	Kolmogorov's 0-1 Law . . . . .	23
1.3	Conditional Probability & Expectation . . . . .	24
1.3.1	Motivation . . . . .	24
1.3.2	Definition of conditional expectation . . . . .	26
1.3.3	Properties of conditional expectation . . . . .	29
1.3.4	Conditioning on Random Variables . . . . .	30
<b>2</b>	<b>Convergence</b>	<b>34</b>
2.1	Weak Law of Large Numbers . . . . .	34
2.2	Strong Law of Large Numbers . . . . .	36
2.2.1	Almost Sure Convergence . . . . .	36
2.2.2	Borel-Cantelli Lemmas . . . . .	38
2.2.3	SLLN with with finite 4-th moment . . . . .	41
2.2.4	Kolmogorov's SLLN . . . . .	43
2.2.5	Connection to Ergodic Theorem . . . . .	49
<b>3</b>	<b>Central Limit Theorem</b>	<b>51</b>
3.1	Convergence in Distribution . . . . .	51
3.1.1	Definitions and Skorohod's theorem . . . . .	51
3.1.2	Characterization of Weak Convergence . . . . .	53

3.2	Characteristic functions . . . . .	55
3.3	Central Limit Theorem. . . . .	56
3.3.1	Introduction. . . . .	56
3.3.2	Triangular Arrays. . . . .	57
3.3.3	The Lindeberg Condition and Some Consequences . . . . .	58
3.3.4	The Lyapounov Condition. . . . .	58
3.3.5	Preliminaries to the proof of Lyapounov's Theorem . . . . .	59
3.3.6	Proof of Lyapounov's Theorem . . . . .	60
3.3.7	Proof of Lindeberg's Central Limit Theorem . . . . .	62
<b>4</b>	<b>Markov Chains</b>	<b>65</b>
4.1	Basic Definitions . . . . .	65
4.1.1	Transition Matrix . . . . .	65
4.1.2	Communicating classes and irreducible Markov chains . . . . .	67
4.1.3	Invariant distribution . . . . .	68
4.1.4	Time reversal . . . . .	69
4.1.5	Markov Chains for Sampling . . . . .	72
4.2	Hitting times and absorption probabilities . . . . .	75
4.3	Recurrence and transience . . . . .	79
4.4	More about invariant distributions . . . . .	84
4.4.1	The existence of an invariant distribution . . . . .	84
4.4.2	Convergence. . . . .	89
4.4.3	Ergodic theorem . . . . .	90
4.4.4	Another example of the Markov Chain Monte Carlo (MCMC) algorithm . . . . .	93
<b>5</b>	<b>Martingales</b>	<b>96</b>
5.1	Filtrations and Stopping Times. . . . .	96
5.2	Definition of Martingales . . . . .	98
5.3	Stopping times and martingales: Examples . . . . .	101
5.4	Martingale convergence theorem . . . . .	104
5.5	Uniformly integrable martingales . . . . .	106
5.6	Regular stopping times. . . . .	108
5.7	Applications of Martingales . . . . .	110
<b>6</b>	<b>Uniform Spanning Trees</b>	<b>112</b>
6.1	General Results . . . . .	112
6.1.1	What is a Uniform Spanning Tree (UST)? . . . . .	112
6.1.2	Wilson's algorithm for UST generation . . . . .	113
6.1.3	USTs, hitting probabilities, and potentials . . . . .	117
6.1.4	Voltages, Currents, and Projections . . . . .	124
6.1.5	Proof of the Burton-Pemantle theorem . . . . .	129
6.2	Square Lattice . . . . .	132
6.3	Bijection with Domino Tilings . . . . .	135
6.4	Connection with Eulerian circuits . . . . .	137

<b>7 Galton-Watson Trees</b>	<b>139</b>
7.1 Galton – Watson process . . . . .	139
7.2 GW process with immigration . . . . .	145
7.3 Size-biased GW trees . . . . .	147
7.4 Supercritical case and a proof of the Kesten-Stigum Theorem (Thm. 7.1.5) . . . . .	150
7.5 Critical Case . . . . .	154
7.6 Simply generated trees . . . . .	159
7.6.1 Definition and main properties . . . . .	159
7.6.2 The main convergence theorem about simply generated trees 163	
7.6.3 Further Examples . . . . .	166
7.6.4 How to generate simply generated random trees? . . . . .	166
7.6.5 Sampling from the multinomial distribution . . . . .	170
<b>8 Random planar maps</b>	<b>172</b>
8.1 Planar maps and Quadrangulations . . . . .	172
8.1.1 Quadrangulations and well-labelled trees . . . . .	174
8.1.2 Embedded trees . . . . .	179
8.1.3 Blossom trees . . . . .	182
<b>A Various Useful Facts</b>	<b>188</b>
A.1 Proof of Caratheodory theorem . . . . .	188
A.2 Function spaces . . . . .	191
A.3 Convergence of Functions and Integration . . . . .	192
A.4 Convergence in $L^1$ and uniform integrability . . . . .	193
A.5 Inequalities . . . . .	194
A.6 Change of Variable . . . . .	197
A.7 Types of Convergence of Random Variables . . . . .	198
A.8 SLLN with with finite 2-nd moment . . . . .	201
A.9 Devroye’s method for generation random variables . . . . .	203
A.10 Statistics of Random Structures . . . . .	205
A.10.1 Random permutations . . . . .	205
A.10.2 Random set partitions . . . . .	206

# Chapter 1

## Foundations



Using Frank Drake's famous equation, Betty calculates the probability of finding intelligent life on a Saturday night.

Figure 1.1: Applications of probability theory

### 1.1 Probability spaces, measures and $\sigma$ -fields

#### 1.1.1 Intuition

The *probability space* is a triple  $(\Omega, \mathcal{F}, \mathbb{P})$ , where the *sample space*  $\Omega$  is a set

of outcomes  $\omega$ ,  $\mathcal{F}$  is a  $\sigma$ -algebra of the events, and  $\mathbb{P}$  is the probability measure.

Intuitively, the set the *sample space*  $\Omega$  represent all possible outcomes  $\omega$  of an experiment. However, we observe the outcomes imperfectly and the *events* are the subsets  $A \subset \Omega$  for which it is possible to say definitely that either “the outcome is in  $A$ ” or “the outcome is not in  $A$ .”

When the experiment has a finite number of outcomes, every events is a finite union of some disjoint elementary events  $A_i$  which form a partition of  $\Omega$ .

Probability is a measure on  $\mathcal{F}$ , that is, it is a function that maps events to non-negative numbers, probabilities. In a case of a finite  $\Omega$ , the probability of every event  $A$  is the sum of probabilities of the elementary blocks of the partition  $\mathcal{F}$  which are subsets of  $A$ .

The  $\sigma$ -algebras are needed to generalize partitions to the case of infinite and possibly uncountable  $\Omega$ .

### 1.1.2 Probability Space

**Definition 1.1.1.** A class  $\mathcal{F}$  of subsets of a space  $\Omega$  is called an *algebra* (or a *field*) if it contains  $\Omega$  itself and is closed under complements and finite unions. That is,

1.  $\Omega \in \mathcal{F}$
2.  $A \in \mathcal{F}$  implies  $A^c \in \mathcal{F}$
3.  $A, B \in \mathcal{F}$  implies  $A \cup B \in \mathcal{F}$

Note that by DeMorgan’s law, given that  $\mathcal{F}$  is closed under complement,  $\mathcal{F}$  is closed under unions if and only if  $\mathcal{F}$  is closed under intersections. Therefore,  $A, B \in \mathcal{F}$  implies  $A \cup B \in \mathcal{F}$  in the above definition can be replaced with  $A, B \in \mathcal{F}$  implies  $A \cap B \in \mathcal{F}$ .

*Example 1.1.2.* Finite unions of all intervals in  $[0,1]$ .

**Definition 1.1.3.** A class  $\mathcal{F}$  of subsets of  $\Omega$  is a  $\sigma$ -*algebra* if it is an algebra and if it is closed under the formation of countable unions. That is,

1.  $\mathcal{F}$  is an algebra.
2.  $A_1, A_2, \dots \in \mathcal{F}$  implies  $A_1 \cup A_2 \cup \dots \in \mathcal{F}$ .

(This object is also often called a  $\sigma$ -*field*.)

An algebra is closed under finite set-theoretic operations whereas a  $\sigma$ -algebra is closed under countable set-theoretic operations.

Usually in a problem dealing with probabilities, one starts with a small class of subsets  $\mathcal{A}$ , for example, with the class of subintervals of  $[0, 1]$ . It is possible that when we perform countable operations on such a class  $\mathcal{A}$  of sets, we might end up operating on sets outside the class  $\mathcal{A}$ .

The  $\sigma$ -*algebra generated by*  $\mathcal{A}$  is denoted by  $\sigma(\mathcal{A})$  and defined as the intersection of all the  $\sigma$ -algebras containing  $\mathcal{A}$ . One can check that this intersection is indeed a  $\sigma$ -algebra. It is clear that it is the smallest  $\sigma$ -algebra containing  $\mathcal{A}$ .

*Example 1.1.4.* The class of all finite unions of intervals  $(a, b]$  in  $\Omega : (0, 1]$  is an algebra but not  $\sigma$  - algebra.

*Example 1.1.5.* The class of all subsets of  $[0, 1]$  is a  $\sigma$  - algebra.

*Example 1.1.6.* If  $\mathcal{A}$  is the class of subintervals  $(a, b]$  of  $\Omega = (0, 1]$ , then the sigma algebra generated by  $\mathcal{A}$  is denoted by  $\mathcal{B}$  and is called the Borel  $\sigma$ -algebra. Its elements are called the *Borel* sets of the unit interval.

Fact: There exist sets which are not Borel.

**Definition 1.1.7.** A set function<sup>1</sup>  $\mu$  on a  $\sigma$ -algebra  $\mathcal{F}$  is a *probability measure* if it satisfies the following conditions:

1.  $0 \leq \mu(A) \leq 1$  for  $A \in \mathcal{F}$ .
2.  $\mu(\emptyset) = 0, \mu(\Omega) = 1$ .
3. If  $A_i \in \mathcal{F}$  is a countable sequence of disjoint sets, then  $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$ .

If we relax assumption that  $\mu(\Omega) = 1$  and require only that  $0 \leq \mu(A)$  for all  $A \in \mathcal{F}$ , the function  $\mu$  is called a measure. A probability measure is often denoted  $\mathbb{P}$ .

If  $\mathcal{F}$  is a  $\sigma$ -algebra on  $\Omega$ , then the pair  $(\Omega, \mathcal{F})$  is called a *measurable space*, and the sets in  $\mathcal{F}$  are called *measurable*. If, in addition,  $\mathbb{P}$  is a probability measure on  $\mathcal{F}$ , then the triple  $(\Omega, \mathcal{F}, \mathbb{P})$  is called a *probability measure space* or simply a *probability space*.

The countable additivity of the probability measure gives rise to the following properties that are stated in a theorem.

**Theorem 1.1.8.** Let  $\mathbb{P}$  be a probability measure on a  $\sigma$ -algebra  $\mathcal{F}$ .

1. *Continuity from below:* If  $A_n$  and  $A$  lie in  $\mathcal{F}$  and  $A_n \uparrow A$ , then  $\mathbb{P}(A_n) \uparrow \mathbb{P}(A)$ .
2. *Continuity from above:* If  $A_n$  and  $A$  lie in  $\mathcal{F}$  and  $A_n \downarrow A$ , then  $\mathbb{P}(A_n) \downarrow \mathbb{P}(A)$ .
3. *Countable subadditivity:* If  $A_1, A_2, \dots$  and  $\bigcup_{k=1}^{\infty} A_k$  lie in  $\mathcal{F}$ , then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

*Example 1.1.9* (Invariant measure on  $S^1$ ).

Suppose that  $\Omega$  is the unit circle and  $\mathcal{F}$  is the class of all subsets of  $\Omega$ . We claim that there is no probability measure on  $\mathcal{F}$ , *invariant* with respect to all rotations.

---

<sup>1</sup>A set function is a real-valued function defined on some class of subsets of  $\Omega$ .

Indeed, let  $S^1 = \mathbb{R}/\mathbb{Z}$  and let two points  $x, y \in S^1$  be equivalent if  $x - y \in \mathbb{Q}$ . This is a valid equivalence relationship, and by Axiom of Choice we can form a set  $A$  by taking exactly one representative in each equivalence class. Define  $A_q = A + q \pmod{1}$  for every rational  $q$ .

Then every point in  $S^1$  belongs to one of the equivalence classes hence it equals to  $a + q$  for some rational  $q$  and some  $a \in A$ . That is, this point is in one of  $A + q$ . It follows that  $S^1$  is a countable union of  $A_q$ . In addition,  $A^q$  are disjoint. Otherwise,  $A$  would contain two different representatives from a single equivalence class. Finally, all  $A_q$  should have the same measure by invariance. However, then either measure of  $A_q$  is 0 and then the measure of  $S^1$  is 0, or the measure of  $A_q$  is positive and then the measure of  $S^1$  is infinite, which contradicts the definition of the invariant measure.

*Example 1.1.10* (Banach - Tarski paradox).

If  $\Omega$  is  $S^2$  and  $\mathcal{F}$  is the class of all subsets of  $\Omega$  then there is no *finitely-additive* finite measure on  $\Omega$ , invariant with respect to all rotations.

*Example 1.1.11* (Lebesgue measure).

Let  $\Omega = \mathbb{R}$  and  $\mathcal{B}$  is the Borel *sigma*-algebra. Then we can define the set-function  $\mu((a, b]) = b - a$  for  $a \leq b$ . Lebesgue showed that this function can be extended to a measure on all sets in the Borel  $\sigma$ -algebra  $\mathcal{B}$ . This measure is called the Lebesgue measure. The probability space on  $[0, 1]$  can then be defined by the restriction of  $\mu$  to the subsets of  $[0, 1]$ .

*Example 1.1.12* (Lebesgue-Stieltjes measure).

More generally one can define the Stieltjes measure on  $(\mathbb{R}, \mathcal{B})$  by using a function  $F$  with the following properties:

- (i)  $F$  is nondecreasing.
- (ii)  $F$  is right continuous, i.e.  $\lim_{y \downarrow x} F(y) = F(x)$ .

**Theorem 1.1.13.** *Associated with each Stieltjes measure function  $F$  there is a unique measure  $\mu$  on  $(\mathbb{R}, \mathcal{B})$  with  $\mu((a, b]) = F(b) - F(a)$ .*

When  $F(x) = x$  the resulting measure is the Lebesgue measure.

The proof of this result is non-trivial even in the case of the Lebesgue measure and is based on the Caratéodory Theorem formulated below. We skip it.

The Lebesgue measure can be extended to an even larger class of sets than Borel sets by adding all sets that are contained in Borel sets of measure zero and assigning measure zero to them. By passing to a minimal containing  $\sigma$ -algebra, one obtains a  $\sigma$ -algebra of *Lebesgue-measurable* sets. It turns out that this  $\sigma$ -algebra is larger than the Borel  $\sigma$ -algebra, and that the Lebesgue measure can be extended to this larger algebra.

### 1.1.3 The Caratéodory Theorem

The Caratéodory Theorem gives the conditions which guarantee that a measure can be extended from an algebra  $\mathcal{A}$  to the generated  $\sigma$ -algebra  $\sigma(\mathcal{A})$ . It is valid not only for probability measures but also for general  $\sigma$ -finite measures.



(The measure  $\mu$  is called  $\sigma$ -finite if there exists a sequence of sets  $A_n \in \mathcal{A}$  so that  $\mu(A_n) < \infty$  and  $\bigcup_{n=1}^{\infty} A_n = \Omega$ . For example, any probability measure is finite and therefore  $\sigma$ -finite. The Lebesgue measure on  $\mathbb{R}$  is not finite, but it is  $\sigma$ -finite.)

It turns out that if  $\mathcal{A}$  is an algebra, then the key condition for the existence of the measure on the  $\sigma$ -algebra  $\sigma(\mathcal{A})$  is the  $\sigma$ -additivity of the measure  $\mu$  on the algebra  $\mathcal{A}$ . However, since we cannot expect that a countable union of sets in  $\mathcal{A}$  is again a set in  $\mathcal{A}$ , we say that a measure  $\mu$  is  $\sigma$ -additive on an algebra  $\mathcal{A}$  if for every sequence of sets  $A_n \in \mathcal{A}$  such that  $\bigcap_n A_n = \emptyset$ , it is true that  $\lim_{n \rightarrow \infty} \mu(A_n) = 0$ .

**Theorem 1.1.14** (Carathéodory). *Let  $\mu$  be a  $\sigma$ -finite,  $\sigma$ -additive measure on an algebra  $\mathcal{A}$ . Then  $\mu$  has a unique extension to a  $\sigma$ -additive measure on  $\sigma(\mathcal{A})$ , the smallest  $\sigma$ -algebra containing  $\mathcal{A}$ .*

The proof of the Carathéodory Theorem is in Appendix. This theorem can be used to justify the existence of the Lebesgue-Stieltjes measure.

Indeed, in order to prove that the Lebesgue measure can be extended to all Borel sets, we need to show that it is countably additive on the algebra  $\mathcal{B}_0$  generated by intervals  $(a, b]$ . This algebra consists of finite unions of disjoint intervals. The following proof of countable additivity is taken from Billingsley “Probability and Measure”.

For shortness, let us use  $|I|$  to denote the length of interval  $I$ . That is,  $|(a, b]| = (b - a)$ .

**Theorem 1.1.15.** *Consider an interval  $I = (a, b]$  and collection of intervals  $I_k = (a_k, b_k)$ , which is either finite or countably infinite.*

1. *Suppose  $I_k$  are disjoint and  $\bigcap_k I_k \subset I$ . Then,  $\sum |I_k| \leq |I|$ .*
2. *Suppose  $I \subset \bigcup_k I_k$ . Then  $|I| \leq \sum |I_k|$ .*
3. *Suppose  $I_k$  are disjoint and  $\bigcap_k I_k = I$ . Then,  $\sum |I_k| = |I|$ .*

*Proof.* Obviously, (3) follows from (1) and (2).

Proof of (1). If the collection is finite, then we prove the statement by induction. Consider  $n$  intervals and suppose they are ordered in such a way that  $a_1 \leq \dots \leq a_n$ . Then  $\bigcup_{k=1}^{n-1} I_k \in (a, a_n]$  and so  $\sum_{k=1}^{n-1} |I_k| \leq a_n - a$  by the induction hypothesis. Consequently,  $\sum_{k=1}^n |I_k| \leq (a_n - a) + (b_n - a_n) < b - a$ .

If the collection is infinite, then  $\sum_{k=1}^n |I_k| \leq |I|$  for every  $n$  by the finite case, and the infinite case follows by passing to the limit.

Proof of (2). If the collection is finite, then we again proceed by induction. Consider the case of  $n$  intervals ordered as before, and suppose that  $a_n < b \leq b_n$ . (Otherwise,  $I$  is covered by  $n - 1$  interval and the induction hypothesis immediately applies.) In addition, assume that  $a < a_n$ , or the result is obvious. Then note that the interval  $(a, a_n]$  must be covered by  $\bigcup_{k=1}^{n-1} I_k$ . Moreover, since the intervals  $I_k$  are closed on the right, the interval  $(a, a_n]$  is covered and we can apply the induction hypothesis,  $\sum_{k=1}^{n-1} |I_k| \geq a_n - a$ . Hence,  $\sum_{k=1}^n |I_k| \geq (a_n - a) + (b_n - a_n) \geq b - a$ .

For the infinite collection, we choose an arbitrary  $\varepsilon > 0$  and consider the closed interval  $[a + \varepsilon, b]$  and open intervals  $(a_k, b_k + \varepsilon 2^{-k})$ . This is a compact interval and hence every infinite open cover of the interval contains a finite sub-cover. We can apply a finite case of (2) to this sub-cover, and find that  $b - a - \varepsilon \leq \sum_{k=1}^{\infty} (|I_k| + \varepsilon 2^{-k}) = \sum_{k=1}^{\infty} |I_k| + \varepsilon$ . Taking the limit  $\varepsilon \rightarrow 0$  on both sides gives the desired inequality.  $\square$

In the algebra  $\mathcal{B}_0$ , every set  $A$  is a union of the finite number of intervals,  $A = \cup_{k=1}^n I_k$  and the Lebesgue measure is defined as

$$|A| = \sum_{k=1}^n |I_k|.$$

**Theorem 1.1.16** (Lebesgue measure is countably-additive). *The Lebesgue measure is a (countable-additive) probability measure on the algebra  $\mathcal{B}_0$ .*

*Proof.* Suppose that  $A = \cup_{k=1}^{\infty} A_k$  where  $A$  and  $A_k$  are in  $\mathcal{B}_0$  and  $A_k$  are disjoint. Then  $A$  and  $A_k$  are disjoint unions of a finite number of intervals  $A = \cup_{i=1}^n I_i$ , and  $A_k = \cup_{j=1}^{n_k} J_j^{(k)}$ . Then, we consider every of the intervals  $I_i$  separately and note that

$$I_i = I_i \cap \bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} \bigcup_{j=1}^{n_k} (I_i \cap J_j^{(k)}).$$

Applying Theorem 1.1.15, we have

$$|A| = \sum_{i=1}^n |I_i| = \sum_{i=1}^n \sum_{k=1}^{\infty} \sum_{j=1}^{n_k} |I_i \cap J_j^{(k)}|$$

Since  $J_j^{(k)} \subset A$ , it is covered by  $\cup_{i=1}^n (I_i \cap J_j^{(k)})$ , and therefore  $\sum_{i=1}^n |I_i \cap J_j^{(k)}| = |J_j^{(k)}|$ . Hence, changing the order of summation above, we get,

$$|A| = \sum_{k=1}^{\infty} \sum_{j=1}^{n_k} |J_j^{(k)}| = \sum_{k=1}^{\infty} |A_k|.$$

$\square$

*Ex. 1.1.17.* Let us try to construct a measure on the set of rationals in  $[0, 1]$ ,  $\mathbb{Q} \subset [0, 1]$ , which would be similar to Lebesgue measure. Consider the function  $\mathbb{P}([a, b]) = b - a$  for all rational  $0 \leq a \leq b \leq 1$ . Can it be extended to an additive finite measure on the algebra generated by intervals in  $\mathbb{Q}$ ?

It is impossible to extend this measure to a countably additive measure on the  $\sigma$ -algebra generated by intervals in  $\mathbb{Q}$ . Indeed,  $\mathbb{P}(\{q\}) = 0$  for any set  $\{q\}$  containing a single rational  $q$  while  $\mathbb{P}(\mathbb{Q}) = \mathbb{P}(\cup_{q \in \mathbb{Q}} q) = 1$ . Which condition of the Carathéodory theorem is not satisfied and why?

**Theorem 1.1.18.** *The Lebesgue-Stieltjes measure is a (countable-additive) probability measure on the algebra  $\mathcal{B}_0$ .*

Sketch of the proof: Theorems 1.1.15 and 1.1.16 show that the Lebesgue measure is countably-additive. If we want to prove the Lebesgue-Stieltjes measure is countably-additive, the proofs go almost verbatim except at one point.

Recall claim (2) in Theorem 1.1.15: “Suppose  $I \subset \cup_k I_k$ . Then  $|I| \leq \sum |I_k|$ .”

For a finite collection  $I_k$  the proof is the same. For an infinite collection, we slightly adapt the proof. Let  $I = (a, b]$ . We choose an arbitrary  $\varepsilon > 0$  and consider the closed interval  $[a + \varepsilon/2, b]$ , the interval  $(a + \varepsilon, b]$  and open intervals  $(a_k, b_k + \delta_k)$ , such that  $F(b_k + \delta_k) - F(b_k) < \varepsilon 2^{-k}$ . This can be done by right-continuity of  $F(x)$ .

The interval  $[a + \varepsilon/2, b]$  is a compact interval and hence every infinite open cover of the interval contains a finite sub-cover. This sub-cover also covers  $(a + \varepsilon, b]$ . We can apply a finite case of (2) to this sub-cover chosen from intervals  $(a_k, b_k + \delta_k)$ , and find that  $F(b) - F(a + \varepsilon) \leq \sum_{k=1}^{\infty} (|I_k| + \varepsilon 2^{-k}) = \sum_{k=1}^{\infty} |I_k| + \varepsilon$ . Taking the limit  $\varepsilon \rightarrow 0$  on both sides and using again the right continuity of  $F(x)$  gives the desired inequality.

#### 1.1.4 Product measures and Kolmogorov’s extension theorem



We now introduce product spaces and product  $\sigma$ -algebras. Given  $(\Omega_i, \mathcal{F}_i)$  measurable sets indexed by  $i \in I$ , let  $\Omega = \prod_i \Omega_i$  the space of sequences  $\omega = (\omega_1, \omega_2, \dots)$ .

The product  $\sigma$ -algebra  $\mathcal{F}$  on  $\Omega$  is the  $\sigma$ -algebra generated by the algebra of the *cylinder sets*  $A_{i_1} \times \dots \times A_{i_n}$  where  $i_1 < \dots < i_n$  are arbitrary finite subsets of index set  $I$ .

*Example 1.1.19.* We can define a  $\sigma$ -field on  $\mathbb{R}^d$  as a product of Borel  $\sigma$ -fields on  $\mathbb{R}$ . It can be thought of as an algebra generated by all rectangles  $I_1 \times \dots \times I_d$ . It turns out that it coincide with the Borel  $\sigma$ -field on  $\mathbb{R}^d$ , that is, with the  $\sigma$ -field generated by all open sets of  $\mathbb{R}^d$ . (This is an easy exercise.)

We can also define the probability measure on the product space if the probability measures  $\mathbb{P}_i$  on spaces  $(\Omega_i, \mathcal{F}_i)$  are given. We start by assigning each cylinder measure  $\mathbb{P}(A_{i_1}) \times \dots \times \mathbb{P}(A_{i_n})$  where  $i_1 < \dots < i_n$ . We can check that this measure is  $\sigma$ -additive on the algebra of cylinder sets and then use the Caratheodory theorem to show that the measure can be extended to a  $\sigma$ -additive measure on  $\mathcal{F}$ . The proof of these statements is not trivial even for a finite number of measure spaces and we are skipping some details here.

The construction for a finite collection of  $\Omega = \mathbb{R}$  gives the Lebesgue measure on  $\mathbb{R}^d$ .

For infinite collections there are additional difficulties, which can be overcome for the case of Borel sets on  $\Omega = \mathbb{R}$ . One formulation of this result is the Kolmogorov’s extension theorem. Here is a variant of this theorem.

Let  $\{\mathbb{P}_n\}$  be a family of measures on  $\mathcal{B}(\mathbb{R}^n)$ , where  $\mathcal{B}(\mathbb{R}^n)$  is the Borel  $\sigma$ -field on  $\mathbb{R}^n$  and call this family of measures *consistent* if  $\mathbb{P}_{n+1}(B \times \mathbb{R}) = \mathbb{P}_n(B)$ .

**Theorem 1.1.20** (Kolmogorov). *Let  $\mathbb{P}_n$ ,  $n = 1, 2, \dots$  be a consistent family of probability measures on  $\mathcal{B}(\mathbb{R}^n)$ . Then there exists a unique probability measure on  $(\mathbb{R}^\infty, \mathcal{F} = \prod \mathcal{B}(\mathbb{R}))$  that agrees with  $\mathbb{P}_n$  on all cylinder sets that depend on the first  $n$  coordinates.*

*Proof.* The requirement of consistency guarantees that if we can take sufficiently large number of indices to include all cylinder sets in a finite union or intersection. In particular, this ensures that we can properly define a finitely-additive measure on the algebra of cylinder sets.

In order to show countable additivity, we can instead show that the measure is continuous at the empty set. That is, if we have a decreasing sequence of cylinder sets  $\overline{B}_n$  such that  $\overline{B}_n \rightarrow \emptyset$ , then  $\mathbb{P}(\overline{B}_n) \rightarrow 0$ . Suppose on the contrary that  $\lim \mathbb{P}(\overline{B}_n) = \delta > 0$ . Without loss of generality we can assume that  $\overline{B}_n$  is supported on the first  $n$  indices. We write  $B_n$  to denote the corresponding set in  $\mathbb{R}^n$ .

The probability measures  $\mathbb{P}_n$  on  $\mathcal{B}(\mathbb{R}^n)$  have the property that for every  $B_n$  and a given  $\delta > 0$ , we can find a compact set  $A_n$  which approximates  $B_n$  well, that is,  $A_n \subset B_n$  and  $\mathbb{P}_n(B_n \setminus A_n) \leq \delta/2^{n+1}$ . We can extend this property from  $\mathbb{R}^n$  to cylinder sets supported on the first  $n$  coordinates in  $\mathbb{R}^\infty$ . Let  $\overline{A}_n$  denote the cylinder sets corresponding to  $A_n$ .

Then define  $\overline{C}_n = \cap_{k=1}^n \overline{A}_k$ . Note that  $\overline{C}_n$  is also supported on the first  $n$  coordinates and every  $C_n$  (restriction of  $\overline{C}_n$  to the first  $n$  coordinates) is compact. (This can be proved by induction.)

The benefit of  $\overline{C}_n$  over  $\overline{A}_n$  is that it is a *decreasing family* of cylinder sets. However, it still approximates  $\overline{B}_n$  well,

$$\begin{aligned} \mathbb{P}(\overline{B}_n \setminus \overline{C}_n) &\leq \sum_{k=1}^n \mathbb{P}(\overline{B}_n \setminus \overline{A}_k) \\ &\leq \sum_{k=1}^n \mathbb{P}(\overline{B}_k \setminus \overline{A}_k) \leq \delta/2. \end{aligned}$$

The first inequality holds because  $\overline{B}_n \setminus \overline{C}_n = \cup(\overline{B}_k \setminus \overline{A}_k)$ , and the inequality in the second line holds because  $\overline{B}_k$  is a decreasing family of sets.

Since by assumption  $\mathbb{P}(\overline{B}_n) \geq \delta$ , hence  $\mathbb{P}(\overline{C}_n) > \delta/2$ . In addition note that  $\overline{C}_n \subset \overline{B}_n$  and therefore  $\overline{C}_n \rightarrow 0$ . The advantage of decreasing sets  $\overline{C}_n$  over  $\overline{B}_n$  is that the restriction of sets  $C_n$  to the first  $n$  coordinates are compact.

Now choose a point  $x^{(n)}$  in each of the cylinder sets  $\overline{C}_n$ . This gives a sequence of points  $x^{(n)} \in \mathbb{R}^\infty$ . Choose a subsequence  $(n_1)$  such that the first coordinate of  $x^{(n_1)}$  converges to a limit  $x_1^*$ . This is possible because  $\overline{C}_n$  is decreasing, hence  $x^{(n)}$  are all in  $\overline{C}_1$  and restriction of  $\overline{C}_1$  to the first coordinate,  $C_1$ , is compact. Then choose a subsequence  $(n_2)$  of the sequence  $(n_1)$  such that the second coordinate of  $x^{(n_2)}$  converged to a limit  $x_2^*$ . This is possible for a similar reason. Proceed further and consider the resulting sequence  $x^* = (x_1^*, x_2^*, \dots)$ .

By construction,  $(x_1^*, \dots, x_n^*) \in C_n$  and therefore  $x^* \in \overline{C_n}$  for every  $n$ . Hence  $x^* \in \bigcap C_n = \emptyset$ , which gives the desired contradiction.  $\square$

Similar to the previous theorem, the assumption of countable additivity in the Caratheodory theorem can be checked for infinite products of probability measures on finite sets, and on  $[0, 1]$ . However, it should be noted that it is not automatically satisfied for arbitrary  $(\Omega_i, \mathcal{F}_i)$ . Usually one requires additionally that  $(\Omega_i, \mathcal{F}_i)$  is a topological space with Borel-sigma algebra  $\mathcal{F}$  and that every  $\Omega_i$  is either compact or satisfy another appropriate condition which might involve measures  $\mathbb{P}_i$ .

For example, it is enough if for every  $\epsilon > 0$  there exists a compact set  $K_i$  with the measure  $\mathbb{P}_i(K_i) > 1 - \epsilon$ .

### 1.1.5 Absolute continuity and singularity of measures

The same measure space  $(\Omega, \mathcal{F})$  can have several different probability measures. A measure  $\mu$  is called *absolutely continuous* with respect to measure  $\nu$ , denoted  $\mu \ll \nu$  if for each  $A \in \mathcal{F}$ ,  $\nu(A) = 0$  implies that  $\mu(A) = 0$ .

For example, consider the case when  $\nu$  is the Lebesgue measure on  $\mathbb{R}$  and  $\mu$  is the Lebesgue-Stieltjes measure corresponding to function  $F(x)$ . Then if  $F(x)$  is differentiable then one can show that  $\mu$  is absolutely continuous with respect to  $\nu$  and, if we assume that we know how to integrate with respect to Lebesgue measure, then one can show that  $\mu(A) = \int_A F'(x)\nu(dx)$ .

In contrast, suppose that  $F(x)$  is the step function:  $F(x) = 0$  for  $x < a$  and  $F(x) = 1$  for  $x \geq a$ . Then, the measure  $\mu$  is not absolutely continuous with respect to Lebesgue measure, because  $\mu(\{a\}) = 1$ , ( $\mu$  has an atom at  $a$ , and the Lebesgue measure of a point is zero).

Two measures  $\mu$  and  $\nu$  are mutually singular, denoted by  $\mu \perp \nu$ , if we can find two *disjoint* sets  $S_\mu$  and  $S_\nu$ , (supports of  $\mu$  and  $\nu$ ) such that  $\mu(\Omega \setminus S_\mu) = 0$  and  $\nu(\Omega \setminus S_\nu) = 0$ .

For example, the atomic measure  $\mu$  above is singular with respect to the Lebesgue measure  $\nu$  since we can choose  $S_\mu = \{a\}$ ,  $S_\nu = \mathbb{R} \setminus \{a\}$ .

In general, a Lebesgue-Stieltjes measure is singular with respect to the Lebesgue measure if  $F'(x) = 0$  everywhere except on a set of measure zero.

An interesting fact about Lebesgue-Stieltjes measures is that they can be singular even if the function  $F(x)$  is continuous. An example is given by the Cantor staircase function which we will consider later.

An abstract form of the theorem about the representation of the absolutely-continuous Lebesgue-Stieltjes measure through its derivative is the Radon-Nikodým Theorem.

A real valued function  $f$  on a measure space  $(\Omega, \mathcal{F})$  is called measurable if for any Borel set  $A$ ,  $f^{-1}(A) \in \mathcal{F}$ . (In fact it is enough to require that  $f^{-1}(A) \in \mathcal{F}$  for all open sets  $A$ .)

**Theorem 1.1.21** (Radon-Nikodým). *If  $\mu$  and  $\nu$  are two  $\sigma$ -finite measures on  $(\Omega, \mathcal{F})$ , such that  $\mu \ll \nu$  then there exists a non-negative measurable  $f$ , called a*

density, such that

$$\mu(A) = \int_A f d\nu,$$

for all  $A \in \mathcal{F}$ . For two such densities,  $f$  and  $g$ , it is true that  $\nu(\{\omega : f(\omega) \neq g(\omega)\}) = 0$ .

The density is called the *Radon-Nikodým derivative* of  $\mu$  with respect to  $\nu$  and often denoted  $d\mu/d\nu$ .

## 1.2 Random Variables

### 1.2.1 Random variables as measurable functions

Let  $(\Omega, \mathcal{F})$  and  $(S, \mathcal{S})$  be two measurable spaces. A map  $X : \Omega \rightarrow S$  is *measurable* or a *random variable* (concisely denoted r.v.) if the inverse image of every measurable set is measurable.

$$X^{-1}(A) \equiv \{\omega : X(\omega) \in A\} \in \mathcal{F} \text{ for all } A \in \mathcal{S}$$

An *indicator function* on an event  $F \in \mathcal{F}$  is an example of a random variable (measurable function) where  $S = \{0, 1\}$  and  $\mathcal{S} = \{\emptyset, \{0\}, \{1\}, S\}$  is the collection of all subsets of  $S$ . The indicator function is defined as

$$\mathbb{1}_F(\omega) = \begin{cases} 1 & \text{if } \omega \in F \\ 0 & \text{if } \omega \notin F \end{cases}$$

A map from a topological space to another topological space is called *Borel measurable* (or  $\mathcal{B}$ -measurable) if it is measurable with respect to the Borel  $\sigma$ -algebras on these spaces. The continuous maps are obviously  $\mathcal{B}$ -measurable. However the class of  $\mathcal{B}$ -measurable functions is significantly larger, since the pre-images of the open sets are not required to be open, as in the case of continuous functions, but only required to be Borel sets.

A map from  $\mathbb{R}^d$  to a topological space  $S$  is called *Lebesgue measurable* if the pre-images of Borel sets are Lebesgue measurable.

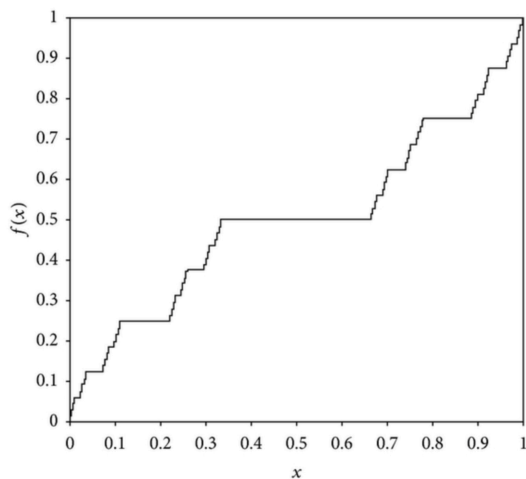
The point of the definition of the Borel and Lebesgue measurable functions is to ensure that the events  $\{\omega : f(\omega) \in A\}$  have a well-defined probability for sufficiently nice sets of elements in  $S$ , that is, for all Borel sets.

**Theorem 1.2.1.** *If maps  $X_1 : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_2, \mathcal{F}_2)$  and  $X_2 : (\Omega_2, \mathcal{F}_2) \rightarrow (\Omega_3, \mathcal{F}_3)$  are measurable, then their composition  $X_2 \circ X_1 : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_3, \mathcal{F}_3)$  is also measurable.*

In particular, the composition of Borel measurable functions is Borel measurable. However, the composition of Lebesgue measurable functions is not necessarily Lebesgue measurable.

**Definition 1.2.2.** The *Cantor set*  $C \subset [0, 1]$  is defined by removing  $(1/3, 2/3)$  from  $[0, 1]$  and then iteratively removing the middle third of each interval that remains.

*Ex.* 1.2.3. The Cantor set  $C$  is closed. The Lebesgue measure of  $C$  is 0.



*Example* 1.2.4. Define the function  $F$  by setting  $F(x) = 0$  for  $x \leq 0$ ,  $F(x) = 1$  for  $x \geq 1$ ,  $F(x) = 1/2$  for  $x \in [1/3, 2/3]$ , then  $F(x) = 1/4$  for  $x \in [1/9, 2/9]$ ,  $F(x) = 3/4$  for  $x \in [7/9, 8/9]$ , and so on.

It can be checked that  $F(x)$  is a non-decreasing continuous function, which is called the *Cantor-Lebesgue function* or the *Cantor staircase*.

Let  $f: [0, 1] \rightarrow [0, 1]$  be the Cantor-Lebesgue function restricted to the interval  $[0, 1]$ . This is a monotonic and continuous function, and the image  $f(C)$  of the Cantor set  $C$  is all of  $[0, 1]$ . Define  $g(x) = x + f(x)$ . Then  $g: [0, 1] \rightarrow [0, 2]$  is a strictly monotonic and continuous map, so its inverse  $h = g^{-1}$  is continuous, too.

Observe that  $g(C)$  has measure one in  $[0, 2]$ : this is because  $f$  is constant on every interval in the complement of  $C$ , so  $g$  maps such an interval to an interval of the same length. It follows that there is a non-Lebesgue measurable subset  $A$  of  $g(C)$ . (This is by Vitali's theorem: a subset of  $\mathbb{R}$  is a Lebesgue null set if and only if all its subsets are Lebesgue measurable. So if all subsets of  $g(C)$  are Lebesgue measurable, then  $g(C)$  has measure null, contradiction.)

Put  $B = g^{-1}(A) \subset C$ . Then  $B$  is a Lebesgue measurable set as a subset of the Lebesgue null set  $C$ , so the characteristic function  $1_B$  of  $B$  is Lebesgue measurable.

The function  $k = 1_B \circ h$  is the composition of the Lebesgue measurable function  $1_B$  and the continuous function  $h$ , but  $k$  is not Lebesgue measurable, since  $k^{-1}(1) = (1_B \circ h)^{-1}(1) = h^{-1}(B) = g(B) = A$ .

## 1.2.2 Generation of $\sigma$ -algebras

Let  $(X_i, i \in I)$  be a family of mappings of a set  $\Omega$  into measurable spaces  $(S_i, \mathcal{S}_i)$ ,  $i \in I$ . Here,  $I \neq \varnothing$  is an arbitrary index set (i.e., possibly uncountable).

For every  $X_i$  we can consider  $X_i^{-1}(\mathcal{S}_i)$ , the collection of inverse images for all sets in  $\mathcal{S}_i$ . One can check that these collections are  $\sigma$ -algebras.

Then the smallest  $\sigma$ -algebra generated by  $X_i^{-1}(\mathcal{S}_i)$  is called the  $\sigma$ -algebra generated by  $(X_i, i \in I)$  and denoted by  $\sigma(X_i, i \in I)$ . We can also define it as the smallest  $\sigma$ -algebra on  $\Omega$  with respect to which each  $X_i$  is measurable.

Here is a theorem that illustrates how these definitions can be used to prove measurability of functions of several variables.

**Theorem 1.2.5.** *Suppose  $(\Omega, \mathcal{F})$ ,  $(S_i, \mathcal{S}_i)$ , and  $(T, \mathcal{T})$  are measurable spaces,  $X_i: \Omega \rightarrow S_i$ ,  $i = 1, \dots, n$ , are measurable maps, and  $f$  is a measurable map from  $(S_1, \mathcal{S}_1) \times \dots \times (S_n, \mathcal{S}_n)$  to  $(T, \mathcal{T})$ . Then  $f(X_1, X_2, \dots, X_n)$  is a measurable map from  $(\Omega, \mathcal{F})$  to  $(T, \mathcal{T})$ .*

*Proof.* In view of Theorem 1.2.1, it suffices to prove that the map  $X: \Omega \rightarrow S \times \dots \times S$ , defined as  $\omega \rightarrow (X_1(\omega), \dots, X_n(\omega))$ , is measurable. To do this, we observe that the  $\sigma$ -algebra on the product space  $S \times \dots \times S$  is generated by products  $A_1 \times \dots \times A_n$ , where  $A_1, \dots, A_n \in \mathcal{S}$ . Then,

$$\{X \in A_1 \times \dots \times A_n\} = \bigcap_{i=1}^n \{X_i \in A_i\} \in \mathcal{F},$$

and therefore  $X$  is measurable. □

## 1.2.3 Distributions

If  $X$  is a real r.v. defined on some probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , then  $X$  induces a probability measure on  $(\mathbb{R}, \mathcal{B})$  called its *distribution measure*. By definition,  $\mu(B) = \mathbb{P}(X \in B) = \mathbb{P}(X^{-1}(B))$ , as a function of Borel sets  $B$  of  $\mathbb{R}$ . To show that  $\mu$  is a probability measure one needs to check countable-additivity. However, it is simply inherited from the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . Namely for disjoint  $B_i$ 's,

$$\begin{aligned} \mu(\cup_i B_i) &= \mathbb{P}[X^{-1}(\cup_i B_i)] \\ &= \mathbb{P}(\cup_i X^{-1}(B_i)) = \sum_i \mathbb{P}(X^{-1}(B_i)) = \sum_i \mu(B_i) \end{aligned}$$

The distribution of a r.v.  $X$  can be described by its *cumulative distribution function* (cdf),  $F(x) = \mathbb{P}(X \leq x)$ .



**Theorem 1.2.6.** A cdf  $F$  of every probability measure on  $\mathbb{R}$  has the following properties:

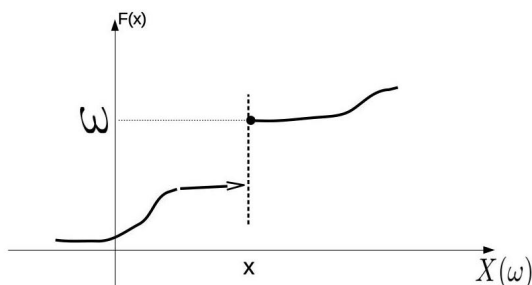
1.  $F$  is a non-decreasing function of  $x$ .
2.  $\lim_{x \rightarrow \infty} F(x) = 1$  and  $\lim_{x \rightarrow -\infty} F(x) = 0$
3.  $F$  is right continuous, i.e.,  $\lim_{y \downarrow x} F(y) = F(x)$

*Proof.* Refer to Theorem 1.1. in Durrett on page 4. □

**Theorem 1.2.7.** If  $F$  satisfies the properties of Theorem 1.2.6, then it is the distribution function of a random variable that takes values in  $\mathbb{R}$ .

This theorem is one of simplest examples when a random variable is constructed from a distribution function.

In addition, this theorem gives (in a sense) a new proof that the Lebesgue-Stieltjes measure on  $\mathbb{R}$  build from a right-continuous non-decreasing function is countably-additive. Now this Lebesgue-Stieltjes measure arises as the distribution measure of the random variable constructed in the theorem. (More precisely, the theorem implies that a right-continuous non-decreasing function  $F(x)$ , with some-additional constraints on its behavior on infinity, defines a (Lebesgue-Stieltjes) countably-additive probability measure on  $\mathbb{R}$ .) The downside is that this proof is more complicated than the proof we described above.



**Figure 1.2:** Construction of a r.v. with a given CDF

*Proof.* We are following Durrett's book here. Let  $F : \mathbb{R} \rightarrow [0, 1]$  have properties 1, 2, 3 in Theorem 1.2.6. We will construct a random variable defined on  $(\Omega = (0, 1], \mathcal{F}, \mathbb{P}) = ((0, 1], \mathcal{B}((0, 1]), \mathbb{P})$ , where  $\mathbb{P}$  denotes the Lebesgue measure, and show that it has the distribution function  $F$ .

The idea of the following proof is that for every  $\omega \in (0, 1]$ , we trying to define  $X(\omega)$  as the inverse of the function  $F(x)$ . This is not always possible so we do it in a sophisticated way.

Note that the set  $\{y : F(y) < \omega\}$  is always an open interval of the form  $(-\infty, \alpha)$ . It is open because otherwise we would have a sequence  $\{\alpha_k\}$  with

$\alpha_k \downarrow \alpha$  such that  $F(\alpha_k) \geq \omega$  and  $F(\alpha) < \omega$  and this would contradict right continuity.

We define  $X(\omega) = \alpha$ , or more formally.

$$X(\omega) = \sup\{y : F(y) < \omega\}.$$

One needs to check the measurability of this function, that is, the fact that the inverse image of every Borel set is Borel. However, we refer to Durrett for this proof.

Then, note that if we manage to show that the following sets are equal,

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

then the definitions of the distribution function and the Lebesgue measure imply that

$$F_X(x) := \mathbb{P}(\omega : X(\omega) \leq x) = \mathbb{P}(\omega : \omega \leq F(x)) = F(x).$$

To check the set equality above, observe that  $\omega \leq F(x)$  means tautologically that  $x$  is outside of the set  $\{y : F(y) < \omega\}$  which means that  $x \geq \alpha = X(\omega)$ .

On the other hand, if  $\omega > F(x)$ , then since  $F$  is right continuous, there is an  $\epsilon > 0$  so that it is still  $\omega > F(x + \epsilon)$  which means that  $x + \epsilon$  is in the set  $\{y : F(y) < \omega\}$ . It follows that  $x$  is strictly less than the supremum of this set,  $X(\omega)$ , that is  $X(\omega) \geq x + \epsilon > x$ . This completes the proof of the set equality stated above and therefore completes the proof of the theorem.  $\square$

Having proved the existence of a r.v.  $X$  with distribution function  $F$ , the uniqueness can be checked easily by the  $\pi - \lambda$  theorem.

*Example 1.2.8* (Singular measures).

Recall the Cantor staircase function  $F(x)$  from Example 1.2.4. Since this function is continuous and non-decreasing, it is clear that it is a valid distribution function.

From the definition, we see that  $dF/dx = 0$  for every  $x$  in the complement of  $C$ . As the Lebesgue measure of  $C$  is zero, we see that the derivative of  $F$  is zero except on a set of zero Lebesgue measure. Such distribution functions are called *Lebesgue singular distribution functions* and the corresponding measures are called *singular measures*.

In particular, there is no function  $f$  for which  $F(x) = \int_{-\infty}^x f(t) dt$  holds.

Even discrete distribution functions can be quite complex.

*Example 1.2.9* (Distribution function with a dense subset of discontinuities).

Let  $q_1, q_2, \dots$  be an enumeration of the rational numbers and set

$$F(x) = \sum_{i=1}^{\infty} 2^{-i} \mathbb{1}_{[q_i, \infty)}(x).$$

Clearly, such  $F$  is non-decreasing, with limits 0 and 1 as  $x \rightarrow -\infty$  and  $x \rightarrow \infty$ , respectively. It is not hard to check that  $F$  is also right continuous, hence a distribution function, whereas by construction  $F$  is discontinuous at each rational number.

## 1.2.4 Expected value



Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space.

**Definition 1.2.10.** Let  $X : \Omega \rightarrow \mathbb{R}$  be a  $\mathcal{B}$ -measurable random variable. The *expected value* of  $X$  is defined by

$$\mathbb{E}(X) := \int_{\Omega} X d\mathbb{P} = \int_{\Omega} X(\omega) \mathbb{P}(d\omega) \quad (1.1)$$

The integral is defined as in Lebesgue integration, whenever  $\int_{\Omega} |X| d\mathbb{P} < \infty$ .

**Theorem 1.2.11** (Existence of the integral for nonnegative r.r.v.). *Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space. There is a unique functional  $\mathbb{E} : X \mapsto \mathbb{E}(X) \in [0, \infty]$  such that*

$$\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A), \quad \forall A \in \mathcal{F} \quad (1.2)$$

$$\mathbb{E}(cX) = c\mathbb{E}(X), \quad \forall c \geq 0, X \geq 0 \quad (1.3)$$

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y), \quad \forall X, Y \geq 0 \quad (1.4)$$

$$X \leq Y \Rightarrow \mathbb{E}(X) \leq \mathbb{E}(Y) \quad (1.5)$$

$$X_n \uparrow X \Rightarrow \mathbb{E}(X_n) \uparrow \mathbb{E}(X) \quad (1.6)$$

*Sketch of Proof.* From these desired properties, we see immediately how to define  $\mathbb{E}(X)$ . The procedure is well known from Lebesgue integration. First extend  $\mathbb{E}$  from indicators to simple r.v.'s by linearity, then to positive r.v.'s by continuity from below, and finally check that everything is consistent.

Step 1: Simple random variables

Check that if  $X = \sum_{i=1}^n c_i \mathbb{1}_{A_i}$  is a simple random variable, then

$$\mathbb{E}(X) = \sum_{i=1}^n c_i \mathbb{P}(A_i) \quad (1.7)$$

works. Verify that  $\mathbb{E}$  is well defined, etc.

Step 2: Nonnegative random variables

Now use (1.6) to extend  $\mathbb{E}$  for general  $X \geq 0$ . We know that there exists an increasing sequence  $X_n$  of simple r.v. with  $X_n \uparrow X$ . Now see that  $\mathbb{E}(X_n) \uparrow$  (by monotonicity of  $\mathbb{E}$ ). Define

$$\mathbb{E}(X) = \lim_{n \rightarrow \infty} \mathbb{E}(X_n) \quad (1.8)$$

Verify again that  $\mathbb{E}(X)$  is well defined.

**Remark:** Note that  $\mathbb{E}(X) = +\infty$  is possible even if  $\mathbb{P}(X < \infty) = 1$ . As an example look at  $G$  which is a geometric r.v., i.e.  $\mathbb{P}(G = g) = 2^{-g}, \forall g = 1, 2, 3, \dots$ . Note that  $\mathbb{P}(G < \infty) = 1$ , but  $\mathbb{E}(2^G) = \sum_{i=1}^{\infty} 2^g 2^{-g} = \infty$ .

Step 3: Signed random variables

Write  $X$  as  $X = X^+ - X^-$ , where  $X^+ := \max(X, 0)$  and  $X^- := -\min(X, 0)$ . Define  $\mathbb{E}(X)$  as follows

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-) \quad (1.9)$$

provided this expression is not  $\infty - \infty$ . Such  $X$  are *quasi-integrable*.  $X$  is *integrable* if  $\mathbb{E}(|X|) < \infty$ .

□

### 1.2.5 Independence

Two sub-algebras  $\mathcal{F}_1$  and  $\mathcal{F}_2$  of the  $\sigma$ -algebra  $\mathcal{F}$  in a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  are independent if for every  $A_1 \in \mathcal{F}_1$  and  $A_2 \in \mathcal{F}_2$ ,  $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$ .

Random variables  $X_1$  and  $X_2$  on the same probability space, are independent if the  $\sigma$ -algebras that they generate are independent. That means that for every Borel sets  $B_1$  and  $B_2$   $\mathbb{P}(X_1 \in B_1, X_2 \in B_2) = \mathbb{P}(X_1 \in B_1)\mathbb{P}(X_2 \in B_2)$ .

In fact it is usually sufficient to check it for a subclass of Borel sets, for example for intervals  $(-\infty, a]$ .

We can extend this definition to a collection of  $n \geq 2$  random variables and even to infinite collections of random variables. In particular, for an infinite sequence of real-valued random variables,  $X_1, X_2, \dots$ , the cumulative distribution function is defined as

$$F_{X_i}(x) := P(X_i \leq x).$$

The random variables  $X_1, X_2, \dots$ , are called *independent*, if

$$\mathbb{P}\left(\bigcap_{i=1}^n \{X_i \leq x_i\}\right) = \prod_{i=1}^n F_{X_i}(x_i) \quad (1.10)$$

for all  $n \geq 1$  and all choices of  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .

*Example 1.2.12.* Consider

$$F_i(x) = \int_{-\infty}^x f_i(y) dy$$

where  $f_i$  is the density of  $X_i$ . Then, the joint law of independent random variables  $X_1, \dots, X_n$  on  $\mathbb{R}^n$  has the density


$$h(x) = \prod_{i=1}^n f_i(x_i)$$

with respect to Lebesgue measure  $dx_1 dx_2 \dots dx_n$ .

**Theorem 1.2.13.** *If  $X$  and  $Y$  are independent and  $\mathbb{E}(|X|) < \infty$ ,  $\mathbb{E}(|Y|) < \infty$  then  $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ .*

This can be proved by using the Fubini theorem.

### 1.2.6 Kolmogorov's 0-1 Law

 Suppose that  $X_1, X_2, \dots$  are independent random variables (not necessarily real valued). Let  $\mathcal{F}'_n = \sigma(X_n, X_{n+1}, \dots)$  be the future after time  $n$ , that is, the smallest  $\sigma$ -field with respect to which all the  $X_m$ ,  $m \geq n$ , are measurable. Let  $\mathcal{T} = \bigcap_n \mathcal{F}'_n$  be the “remote future”, or the *tail*  $\sigma$ -field.

*Example 1.2.14.*  $\{\omega : S_n(\omega) \text{ converges}\} \in \mathcal{T}$ .

**Theorem 1.2.15** (Kolmogorov's 0-1 Law). *If  $X_1, X_2, \dots$  are independent and  $A \in \mathcal{T}$  then  $\mathbb{P}(A) = 0$  or  $1$ .*

We will only sketch the proof. For details, see textbooks.

*Proof.* The idea is to show that  $A$  is independent of itself, that is,  $\mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$ , so  $\mathbb{P}(A) = \mathbb{P}(A)^2$ , and hence  $\mathbb{P}(A) = 0$  or  $1$ . We will prove this in two steps:

(a)  $A \in \sigma(X_1, \dots, X_k)$  and  $B \in \sigma(X_{k+1}, X_{k+2}, \dots)$  are independent.

Proof of (a): If  $B \in \sigma(X_{k+1}, \dots, X_{k+j})$  for some  $j$ , this is more or less evident. Passing to the limit  $j \rightarrow \infty$  needs a justification which can be found in a textbook.

(b)  $A \in \sigma(X_1, X_2, \dots)$  and  $B \in \mathcal{T}$  are independent.

Proof of (b): Since  $\mathcal{T} \subset \sigma(X_{k+1}, X_{k+2}, \dots)$ , if  $A \in \sigma(X_1, \dots, X_k)$  for some  $k$ , this follows from (a). Passing to the limit  $k \rightarrow \infty$  needs a justification that can be found in a textbook.

Since  $\mathcal{T} \subset \sigma(X_1, X_2, \dots)$ , (b) implies that  $A \in \mathcal{T}$  is independent of itself and the theorem follows. □

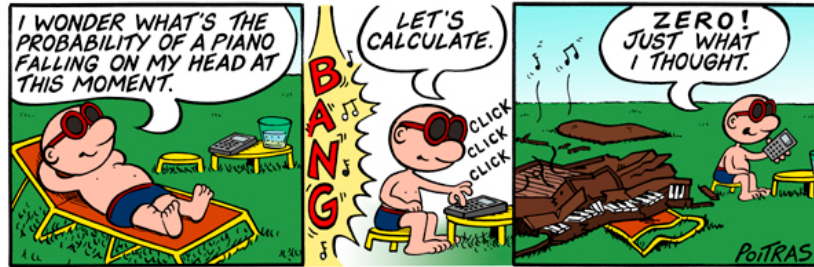


Figure 1.3: Everything is conditional

## 1.3 Conditional Probability & Expectation

### 1.3.1 Motivation

Calculating conditional probabilities might lead to paradoxes if we are not careful. This is illustrated by the following example.

*Example 1.3.1.* Suppose the random variables  $X$  and  $Y$  have the joint density function

$$f_{XY}(x, y) = \begin{cases} 4xy, & \text{if } x \in [0, 1], y \in [0, 1], \\ 0, & \text{otherwise.} \end{cases}$$

Find the conditional density of  $X$  given that  $Y = X$ .

As we will see later, the ambiguity here is created by the condition  $Y = X$ . Consider two different approaches to this problem.

**Solution 1.**

Let  $U = X$  and  $V = Y - X$ . Our goal is to find the conditional density of  $U$  given that  $V = 0$ .

First, we find the joint density of  $U$  and  $V$ . We have

$$f_{UV}(u, v) = f_{XY}(x, y)|J|^{-1},$$

where  $J$  is the jacobian of the transformation,

$$J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

Therefore,  $f_{UV}(u, v) = 4u(u + v)$ , for  $0 < u < 1$ ,  $-u < v < -u + 1$ .

The marginal density of  $V$  for  $-1 < v < 0$  is

$$\begin{aligned} f_V(v) &= 4 \int_{-v}^1 (u^2 + uv) du \\ &= \frac{2}{3} (1 + v)^2 (2 - v). \end{aligned}$$

For  $0 < v < 1$ ,

$$\begin{aligned} f_V(v) &= 4 \int_0^{1-v} (u^2 + uv) du \\ &= \frac{2}{3}(v-1)^2(2+v). \end{aligned}$$

Using either probability density we have  $f_V(0) = 4/3$ . Therefore,

$$f_{U|V}(u|v=0) = \frac{f_{UV}(u,0)}{f_V(0)} = 3u^2 = 3x^2.$$

Hence,

$$f(X|Y=X) = 3x^2.$$

**Solution 2.** Take  $U = X$  and  $W = Y/X$ . Now, the goal is to find  $f_{U|W}(u, w=1)$ .

We calculate the Jacobian of the transformation as  $J = 1/x$  and  $f_{UW}(u, w) = f_{XY}(x, y)|J|^{-1} = 4u^3w$  for  $0 < u < 1$  and  $0 < w < 1/u$ . The marginal density of  $W$  is  $f_W(w) = w$  for  $0 < w < 1$  and  $f_W(w) = 1/w^3$  for  $w > 1$ . In both cases,  $f_W(1) = 1$ .

Therefore,

$$df_{U|W}(u|w=1) = \frac{f_{UW}(u,1)}{f_W(1)} = 4u^3 = 4x^3.$$

Hence,

$$f(X|Y=X) = 4x^3,$$

which is clearly different from our previous answer.


How can we explain this paradox?

The reason behind it is that the set  $\{Y = X\}$  has zero measure and therefore we cannot apply the usual formula

$$f_X(x|B)dx = \frac{f(x, B)dx}{P(B)}.$$

In order to make sense, we have to approximate the set  $\{Y = X\}$  by measurable sets with positive probability. Now, the main problem is that we are not given a  $\sigma$ -algebra  $\mathcal{G}$  from which we can take the approximating sets. We can choose  $\mathcal{G}_1 = \sigma(Y - X)$  or  $\mathcal{G}_2 = \sigma(Y/X)$ , which consist of the unions of the level sets of the functions  $Y - X$  and  $Y/X$ , respectively. Then these two approximations are different and it turns out that the conditional densities with respect to these two  $\sigma$ -algebras are different too.

### 1.3.2 Definition of conditional expectation

 We present the definition of conditional expectation due to Kolmogorov (1933).

**Definition 1.3.2.** Given the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , some sub- $\sigma$ -field  $\mathcal{G} \subset \mathcal{F}$ , and a random variable  $X \in \mathcal{L}^1(\mathcal{F})$  (meaning that  $X$  is  $\mathcal{F}$ -measurable and  $\mathbb{E}|X| < \infty$ ), the conditional expectation of  $X$  given  $\mathcal{G}$  is the (almost surely unique) random variable  $\hat{X}$  such that

- i.  $\hat{X} \in \mathcal{L}^1(\mathcal{G})$  that is,  $\hat{X}$  is  $\mathcal{G}$ -measurable; and
- ii.  $\mathbb{E}(\hat{X}\mathbb{1}_G) = \mathbb{E}(X\mathbb{1}_G)$  for all  $G \in \mathcal{G}$ : that is,  $\hat{X}$  integrates like  $X$  over all  $\mathcal{G}$ -sets.

The random variable  $\hat{X}$  is denoted by  $\mathbb{E}(X|\mathcal{G})$ .

The conditional expectation can also be defined for non-negative random variables even if they have infinite expectation.

**Theorem 1.3.3.** *The conditional expectation  $\mathbb{E}(X|\mathcal{G})$  is unique up to almost sure equivalence.*

*Proof.* Suppose that two random variables  $\hat{X}_1$  and  $\hat{X}_2$  are candidates for the conditional expectation  $\mathbb{E}(X|\mathcal{G})$ . Let  $Y := \hat{X}_1 - \hat{X}_2$ . So we have  $Y \in L^1(\mathcal{G})$  and  $\mathbb{E}(Y\mathbb{1}_G) = 0 \forall G \in \mathcal{G}$ .

In particular, if we take  $G = \{Y > \varepsilon\}$ , then  $\mathbb{E}(Y\mathbb{1}_{Y>\varepsilon}) = 0$ . Since  $\mathbb{E}(Y\mathbb{1}_{Y>\varepsilon}) \geq \varepsilon\mathbb{P}(Y > \varepsilon)$ , we conclude that  $\mathbb{P}(Y > \varepsilon) = 0$  for every  $\varepsilon > 0$ .

Interchanging the roles of  $X_1$  and  $X_2$ , we have  $\mathbb{P}(Y < -\varepsilon) = 0$ . And since  $\varepsilon$  is arbitrary,  $\mathbb{P}(Y = 0) = 1$ .  $\square$

**Theorem 1.3.4.** *The conditional expectation  $\mathbb{E}(X|\mathcal{G})$  exists for all  $\mathcal{F}$  measurable functions with finite expectation.*

The conditional expectation also exists for non-negative  $\mathcal{F}$ -measurable functions even if their expectation is infinite, but in this case this function can sometime take value  $+\infty$ .

We are not going to prove this theorem from the basic principles but rather point out that it is a consequence of results from either the measure theory or the theory of Hilbert spaces.

**Measure theory proof.**

*Proof of the existence of conditional expectation via Radon-Nikodym Theorem.* Assume first that  $X \geq 0$  and define the measure  $Q$  on  $[\Omega, \mathcal{G}]$ , so that for every  $C \in \mathcal{G}$ ,

$$Q(C) = \int_C X dP = \mathbb{E}(X\mathbb{1}_C)$$



. This is a finite measure because  $\mathbb{E}|X| < \infty$ . Note that  $Q$  is absolutely continuous with respect to  $P$ , if we consider  $P$  as a measure on  $(\Omega, \mathcal{G})$ . Indeed if  $P(C) = 0$ , then we can approximate  $X$  by bounded functions  $X_n$ , and use the dominated convergence theorem to find that

$$Q(C) = \int_C X dP = \lim_{n \rightarrow \infty} \int_C X_n dP = 0.$$

Then the Radon-Nikodym theorem is applicable and it implies the existence of the density  $\hat{X} = dQ/dP$  which has all the properties of the conditional expectation  $\hat{X} = \mathbb{E}(X|\mathcal{G})$ .

For general  $X$  we can employ  $\mathbb{E}(X^+|\mathcal{G}) - \mathbb{E}(X^-|\mathcal{G})$ . □

[See Durrett for a full proof]

**Hilbert space method.**

This gives a nice geometric picture for the case when  $Y \in \mathcal{L}^2(\Omega, \mathbb{P})$ . The proof is based on the following two results from the theory of Hilbert spaces.

**Lemma 1.3.5.** *Every nonempty, closed, convex set  $E$  in a Hilbert space  $H$  contains a unique element of smallest norm.*

**Lemma 1.3.6** (Existence of Projections in Hilbert Space). *Given a closed subspace  $K$  of a Hilbert space  $H$  and element  $x \in H$ , there exists a decomposition  $x = y + z$  where  $y \in K$  and  $z \in K^\perp$  (the orthogonal complement).*

(See Rudin 87 (p.79) for a full discussion of Lemma 1.3.5.)

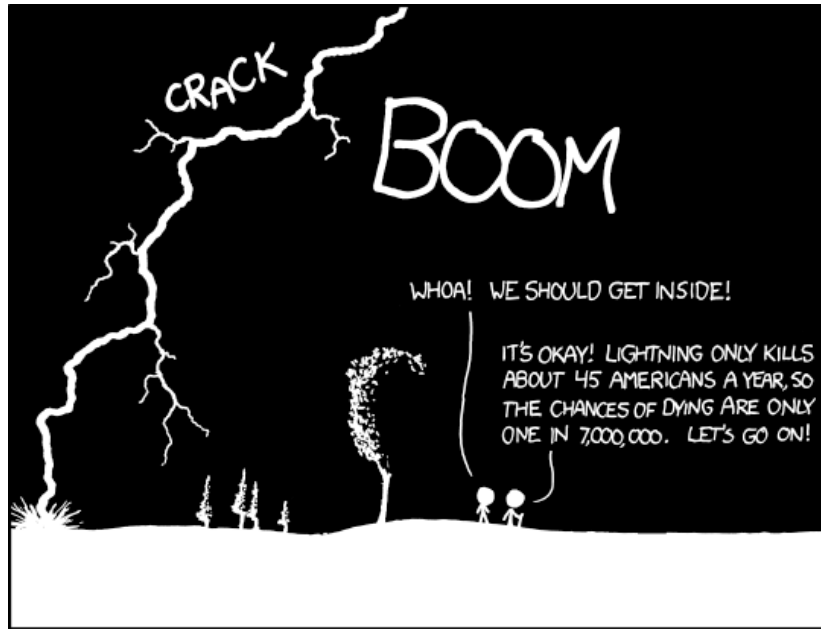
*Existence of conditional expectation via Hilbert space projections.* Suppose  $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ . Requirement (ii) demands that for all  $X \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ ,

$$\mathbb{E}\left((Y - \mathbb{E}(Y|\mathcal{G}))X\right) = 0$$

which has the geometric interpretation of requiring  $Y - \mathbb{E}(Y|\mathcal{G})$  to be orthogonal to the subspace  $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ . Requirement (i) says that  $\mathbb{E}(Y|\mathcal{G}) \in \mathcal{L}^2(\mathcal{G})$ .

So  $\mathbb{E}(Y|\mathcal{G})$  is the orthogonal projection of  $Y$  onto the closed subspace  $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ . The lemma above shows that such a projection is well defined.

The rest of the proof, which we omit, is concerned with extending the result from  $\mathcal{L}^2$ -spaces (the random variables with finite 2nd moment) to  $\mathcal{L}^1$  spaces. □



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Figure 1.4: More about conditional probability

### 1.3.3 Properties of conditional expectation

We record some basic properties of  $\mathbb{E}(\cdot|\mathcal{G})$  as an operator,  $X \mapsto \mathbb{E}(X|\mathcal{G})$ :

1. Positivity:  $Y \geq 0 \Rightarrow \mathbb{E}(Y|\mathcal{G}) \geq 0$
2. Linearity:  $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$
3.  $\mathbb{E}(\cdot|\mathcal{G})$  is a projection:  $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})$
4.  $\mathbb{E}(\cdot|\mathcal{G})$  is continuous with norm 1 in  $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$  spaces for  $p \geq 1$ :

$$\|\mathbb{E}(X|\mathcal{G})\|_p \leq \|X\|_p$$

and

$$X_n \xrightarrow{2} X \text{ implies } \mathbb{E}(X_n|\mathcal{G}) \xrightarrow{2} \mathbb{E}(X|\mathcal{G})$$

5. Tower property. If  $\mathcal{H} \subset \mathcal{G}$ , then:  $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ .
6. If  $Y \in \mathcal{G}$  and  $\mathbb{E}|XY| < \infty$ , then  $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(X|\mathcal{G})Y$ .

7.  $\mathbb{E}(\cdot|\mathcal{G})$  respects monotone convergence:

$$0 \leq X_n \uparrow X \text{ implies } \mathbb{E}(X_n|\mathcal{G}) \uparrow \mathbb{E}(X|\mathcal{G})$$

8. If  $\varphi$  is convex and  $\mathbb{E}|\varphi(X)| < \infty$  then a conditional form of Jensen's inequality holds:  $\varphi(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}(\varphi(X)|\mathcal{G})$ .


9.  $\mathbb{E}(\cdot|\mathcal{G})$  is an orthogonal projection in  $\mathcal{L}^2$ :  $\mathbb{E}((X-\hat{X})Z) = 0$  for  $\hat{X} = \mathbb{E}(X|\mathcal{G})$  and all  $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ .

10. *Repeated Conditioning*: for  $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \dots$ ,  $\mathcal{G}_\infty = \sigma(\cup \mathcal{G}_i)$  and  $X \in \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P})$  with  $p \geq 1$ ,

$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{a.s.} \mathbb{E}(X|\mathcal{G}_\infty)$$

$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{p} \mathbb{E}(X|\mathcal{G}_\infty)$$

### 1.3.4 Conditioning on Random Variables

eturning to the example in Section 1.3.1, note that if we use the  $\sigma$ -algebra generated by the event  $\{Y = X\}$ , then the conditional expectation is not defined uniquely. The reason is that this event has zero probability and the uniqueness holds only up to sets of measure zero.

However, we were able to calculate the conditional density after we reformulated the problem and used a richer  $\sigma$ -algebra. This  $\sigma$ -algebra allowed us to approximate the sets with zero probability. In this section we talk more about the concept of conditional densities and, more generally, about the concept of conditional distributions.

If  $X$  and  $Y$  are two random variables then we use notation  $\mathbb{E}(Y|X)$  to denote  $\mathbb{E}(Y|\mathcal{G})$  where  $\mathcal{G}$  is the  $\sigma$ -algebra generated by  $X$ .

It is very useful to be able to write formulas like

$$\mathbb{E}(Y|X = x) = \int y f(y|x) dy,$$

where  $f(y|x)$  denoted the conditional density. It would be useful to define this concept with more rigour.

**Definition 1.3.7.** Let  $(S, \mathcal{S})$  and  $(T, \mathcal{T})$  be two measure spaces. A *Markov kernel* from  $(S, \mathcal{S})$  to  $(T, \mathcal{T})$  is a collection of probability measures  $P_s$  on  $(T, \mathcal{T})$  indexed by a parameter  $s \in S$ . It is supposed to satisfy the measurability condition: for each  $A \in \mathcal{T}$ ,  $s \mapsto P_s(A)$  is measurable relative to  $\mathcal{S}$ .

*Example 1.3.8.* In statistics literature,  $\theta \in \Theta$  is a parameter and  $\mathbb{P}_\theta$  is a family of probability measures on a measurable space of "data":  $(X, \mathcal{X})$ .

**Definition 1.3.9.** Let  $X : \Omega \mapsto S$  and  $Y : \Omega \mapsto T$  be two random variables on  $(\Omega, \mathcal{F}, \mathbb{P})$ . Then a *regular conditional distribution* for  $Y$  given  $X$  is a Markov kernel  $P_x$  such that the joint distribution of  $(X, Y)$  has the property

$$\mathbb{P}(X \in A, Y \in B) = \int_A P_x(B) \mathbb{P}(X \in dx)$$

The regular conditional distributions allows us to calculate the conditional expectations in a straightforward way:

$$\mathbb{E}(Y|X = x) = \int y P_x(dy).$$

The regular conditional distributions usually but not always exist. We will not go in details of this.

*Example 1.3.10* (If joint density exists).

If  $X, Y$  have density  $f(x, y)$  relative to  $dx, dy$  for some reference measures  $dx$  on  $(S, \mathcal{S})$  and  $dy$  on  $(T, \mathcal{T})$ , then

$$P_x(dy) = \frac{f(x, y)}{f_X(x)} dy,$$

where  $f_X$  is the marginal density of  $X$ .

*Example 1.3.11* (Discrete conditional probability distribution).

A point  $(X, Y)$  is picked proportional to length measure on the perimeter of an equilateral triangle with vertices  $(0, 0)$ ,  $(1, 0)$  and  $(\frac{1}{2}, \frac{\sqrt{3}}{2})$ . What is the conditional distribution of  $Y$  given  $X = x$ ?

First, observe that  $X$  is uniform on  $[0, 1]$  and  $Y \sim \frac{1}{3} \cdot \delta_0 + \frac{2}{3} \cdot \mathcal{U}[0, \frac{\sqrt{3}}{2}]$ .

For  $0 \leq x \leq \frac{1}{2}$  the only possible values of  $Y$  consistent with  $X = x$  are  $y = 0$  or  $y = \sqrt{3}x$ . Thus

$$\mathbb{P}(Y > 0 | 0 \leq x \leq \frac{1}{2}) = \frac{\mathbb{P}(Y > 0, 0 \leq x \leq \frac{1}{2})}{\mathbb{P}(0 \leq x \leq \frac{1}{2})} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

More generally, choose  $\varepsilon > 0$  and  $0 < x < 1/2 - \varepsilon$ . Then, it is not difficult to convince yourself by calculating lengths of intervals that

$$\mathbb{P}(Y > 0 | x < X < x + \varepsilon) = \frac{\mathbb{P}(Y > 0, x < X < x + \varepsilon)}{\mathbb{P}(x < X < x + \varepsilon)} = \frac{2}{3}$$

and therefore

$$\mathbb{P}(Y = 0 | x < X < x + \varepsilon) = \frac{1}{3}$$

By taking the limit  $\varepsilon \rightarrow 0$ , for  $0 \leq x \leq \frac{1}{2}$ , the conditional distribution is

$$P_x = \frac{1}{3} \delta_0 + \frac{2}{3} \delta_{\sqrt{3}x},$$

where  $\delta_a$  denotes the atomic probability measure concentrated at  $a$ . A similar calculation can also be done for  $\frac{1}{2} \leq x \leq 1$ .

The conditional densities and distributions can also be helpful to calculate the marginal densities. If the marginal density of  $X$  and the conditional density of  $Y$  given  $X$  are known, then the marginal density of  $Y$  can be calculated as

$$f_Y(y)dy = \int f_{Y|X}(y|x)f_X(x)dx.$$




"I wish we hadn't learned probability 'cause I don't think our odds are good."

Figure 1.5: Bleak future for probabilists

# Chapter 2

## Convergence

### 2.1 Weak Law of Large Numbers

The Weak Law of Large Numbers is a statement about sums of independent random variables. Before we state the WLLN, it is necessary to define convergence in probability.

**Definition 2.1.1.** Given a sequence of r.v.'s  $Y_n$  defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , say  $Y_n$  converges in probability to  $Y$ ,  $Y_n \xrightarrow{\mathbb{P}} Y$ , if  $Y$  is a r.v. on  $(\Omega, \mathcal{F})$ , and for all  $\epsilon > 0$ ,

$$\lim_{m \rightarrow \infty} \mathbb{P}(|Y_n - Y| > \epsilon) = 0.$$

**Theorem 2.1.2** (Weak Law of Large Numbers). *Let  $X, X_1, X_2, \dots$  be a sequence of i.i.d. random variables with  $E|X| < \infty$  and define  $S_n = X_1 + X_2 + \dots + X_n$ . Then*

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} EX.$$

*Proof.* In this proof, we employ the common strategy of first proving the result under an  $L^2$  condition (i.e. assuming that the second moment is finite), and then using truncation to get rid of the extraneous moment condition.

First, we assume  $EX^2 < \infty$ . By independence of  $X_i$ ,

$$\text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\text{Var}(X)}{n}.$$

By Chebychev's inequality,  $\forall \epsilon > 0$ ,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - EX\right| > \epsilon\right) \leq \frac{1}{\epsilon^2} \text{Var}\left(\frac{S_n}{n}\right) = \frac{\text{Var}(X)}{n\epsilon^2} \rightarrow 0.$$

Thus,  $\frac{S_n}{n} \xrightarrow{\mathbb{P}} EX$  under the finite second moment condition. To transition from  $L^2$  to  $L^1$ , we use truncation. For  $0 < x < \infty$  let

$$\begin{aligned} X_{xk} &= X_k \mathbb{1}_{(|X_k| \leq x)} \\ Y_{xk} &= X_k \mathbb{1}_{(|X_k| > x)} \end{aligned}$$

Then, we have  $X_k = X_{xk} + Y_{xk}$  and

$$\begin{aligned} \frac{S_n}{n} &= \frac{1}{n} \sum_{k=1}^n X_{xk} + \frac{1}{n} \sum_{k=1}^n Y_{xk} \\ &= U_{xn} + V_{xn} \end{aligned}$$

By monotonicity of expectation, we have

$$E \left| \frac{1}{n} \sum_{k=1}^n Y_{xk} \right| \leq \frac{1}{n} \sum_{k=1}^n E|Y_{xk}| = E(|X| \mathbb{1}_{(|X| > x)}) \rightarrow 0, \quad x \rightarrow \infty,$$

where the last convergence can be shown by using the dominated convergence theorem.

Fix  $1 > \epsilon > 0$  and choose  $x$  such that

$$E(|X| \mathbb{1}_{(|X| > x)}) = E|Y_{x1}| < \epsilon^2.$$

Let  $\mu_x = E(X_{x1})$  and  $\mu = E(X)$ . Then, we also have

$$|\mu_x - \mu| \leq E|Y_{x1}| < \epsilon^2 < \epsilon.$$

Let  $B_n = \{|U_{xn} - \mu_x| > \epsilon\}$  and  $C_n = \{|V_{xn}| > \epsilon\}$ . Noting that  $E(X_{xk}^2) \leq x^2 < \infty$ , we can apply the Weak Law of Large Numbers to  $U_{xn}$ . Thus, we choose  $N > 0$  such that  $\forall n > N$ ,

$$\mathbb{P}(B_n) = \mathbb{P}(|U_{xn} - \mu_x| > \epsilon) < \epsilon.$$

Now, by Markov's inequality, we also have

$$\mathbb{P}(C_n) = \mathbb{P}(|V_{xn}| > \epsilon) \leq \frac{E|V_{xn}|}{\epsilon} \leq \frac{E|Y_{x1}|}{\epsilon} \leq \epsilon$$

But on  $B_n^c \cap C_n^c = (B_n \cup C_n)^c$ , we have  $|U_{xn} - \mu_x| \leq \epsilon$  and  $|V_{xn}| \leq \epsilon$ , and therefore

$$\left| \frac{S_n}{n} - \mu \right| \leq |U_{xn} - \mu_x| + |V_{xn}| + |\mu_x - \mu| \leq 2\epsilon + \epsilon^2 \leq 3\epsilon.$$

Thus,  $\forall n > N$ ,

$$\mathbb{P} \left( \left| \frac{S_n}{n} - EX \right| > 3\epsilon \right) \leq \mathbb{P}(B_n \cup C_n) \leq 2\epsilon.$$

□

## 2.2 Strong Law of Large Numbers

### 2.2.1 Almost Sure Convergence

Here our goal is to show that the law of large number holds with respect to a stronger concept of convergence, the almost sure convergence of random variables.

First, we show some preliminary results that show that it makes sense to say that a sequence of random variables converges with probability 1.

This can be done in several ways. If we are considering the real-valued random variables, then we can use the fact that  $\mathbb{R}$  is linearly ordered and the convergence of a sequence  $x_n$  is equivalent to the statement that the  $\limsup x_n$  and  $\liminf x_n$  are equal.

**Theorem 2.2.1.** *If  $X_1, X_2, \dots$  are measurable real-valued functions on  $(\Omega, \mathcal{F})$ , then  $\inf X_n, \sup X_n, \liminf X_n$ , and  $\limsup X_n$  are also measurable.*

*Proof.* We only need to prove that the pre-images of the sets  $(-\infty, a)$  are measurable. First,

$$\{\inf X_n < a\} = \bigcup_n \{X_n < a\},$$

which is a measurable set.

Similarly,

$$\{\liminf X_n < a\} = \bigcap_m \bigcup_{n \geq m} \{X_n < a\},$$

which is also measurable.

The argument for  $\sup X_n$  and  $\limsup X_n$  proceed in the same way.  $\square$

Difference of measurable functions is measurable. Hence  $\limsup_{n \rightarrow \infty} X_n - \liminf_{n \rightarrow \infty} X_n$  is a random variable. Hence,

$$\Omega_0 \equiv \left\{ \omega : \lim_{n \rightarrow \infty} X_n \text{ exists} \right\} = \left\{ \omega : \limsup_{n \rightarrow \infty} X_n - \liminf_{n \rightarrow \infty} X_n = 0 \right\}$$

is a measurable set and its probability is well defined.

**Definition 2.2.2.** If  $X_n(\omega)$  converges for almost all  $\omega$ , i.e.,  $\mathbb{P}(\Omega_0) = 1$ , we say that  $X_n$  converges *almost surely* (a.s.) to  $X = \liminf_{n \rightarrow \infty} X_n$ .

It is also said that  $X_n(\omega)$  converges *almost everywhere* (a.e.) or *with probability 1*. The notation is  $X_n \xrightarrow{a.s.} X$ .

Note that in general we consider two random variables as equivalent (“equal in a wide sense”), if they differ only on the set of measure 0. So we could take  $\limsup_{n \rightarrow \infty} X_n$  instead of  $\liminf_{n \rightarrow \infty} X_n$  in the definition of the almost sure convergence.

The second way is more general and works for random variables  $X_n$  that take values in a metric space  $S$ . In this case we start with the distribution measures



$\mu_{X_1}$  on  $S$ ,  $\mu_{X_1, X_2}$  on  $S^2 = S \times S$ , and generally,  $\mu_{X_1, \dots, X_n}$  on  $S^n = S \times \dots \times S$ . By the Kolmogorov extension theorem, one can define a measure  $\mu_\infty$  on  $S^\infty$ , the space of infinite sequences. Recall that a sequence  $x = (x_1, x_2, \dots)$  is convergent to  $a$  if it belongs to the set

$$A = \bigcap_{k=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{i \geq n} \left\{ x : d(x_i, a) < \frac{1}{k} \right\}$$

So we say that a sequence of random variables  $X_n$  *almost surely converges* to a limit  $a$  if  $\mu_\infty(A) = 1$ .

Question: What is the difference between almost sure convergence and convergence in probability? Is any of them implies the other one?

It is a theorem that almost sure convergence implies convergence in probability. (See Appendix for a proof.) However the converse is not true, and there are sequences of random variables that converge in probability but not almost surely.

*Example 2.2.3* (Moving blip).

On  $\Omega = [0, 1]$  with Borel  $\sigma$ -algebra and Lebesgue measure, define  $X_n(\omega) = \mathbb{1}_{(x_n, x_{n+1})}(\omega)$  where the interval  $(x_n, x_{n+1})$  is wrapped around the unit circle. Let  $x_n$  be any sequence such that  $x_n$  grows monotonically  $\rightarrow \infty$  and  $x_{n+1} - x_n \rightarrow 0$  (e.g.  $x_n = 1 + \frac{1}{2} + \dots + \frac{1}{n} \pmod{1}$ , or  $x_n = \log n \pmod{1}$ ). Then  $\mathbb{P}(|X_n| > \epsilon) = x_{n+1} - x_n \rightarrow 0$  for any fixed  $\epsilon > 0$  and therefore  $X_n \xrightarrow{\mathbb{P}} 0$ , but for any  $\omega$  and arbitrary  $N$  we can find  $n > N$  so that  $X_n(\omega) = 1$ . Hence for all  $\omega$  the sequence  $X_n$  does not converge to 0 (or any other value). In particular,  $X_n$  does not converge to 0 almost surely.

## 2.2.2 Borel-Cantelli Lemmas

Our target will be to prove that  $S_n/n$  converges to  $\mathbb{E}X$  not only in probability but also almost surely, so we are going to develop tools for establishing almost sure convergence.

First, note that the convergence of a sequence  $x_n$  to a limit  $x$  means that for every open interval  $I$  around  $x$ , the sequence  $x_n$  will be outside the interval for only a finite number of times. Hence, we are interested in calculating the probability that an event  $X_n \in I$  happens only a finite number of times. This calculation can be done by using the Borel - Cantelli lemmas.

Let us introduce some notation. Let  $A_n$  be a sequence of events. Then we can define a new event  $\{A_n \text{ i.o.}\}$ ,

$$\begin{aligned} \{A_n \text{ i.o.}\} &= \limsup A_n = \lim_{m \rightarrow \infty} \bigcup_{n \geq m} A_n \\ &= \bigcap_m \bigcup_{n \geq m} A_n. \end{aligned}$$

In words,  $\{A_n \text{ i.o.}\}$  consists of outcomes  $\omega$  that are in infinitely many  $A_n$ , that is, they repeat *infinitely often*.

Similarly, we define  $\{A_n \text{ ev.}\}$  as the set of outcomes  $\omega$ , which are in all  $A_n$  except for a finitely many  $n$ . Here “ev.” stands for eventually.

Formally,

$$\begin{aligned} \{A_n \text{ ev.}\} &= \liminf A_n = \lim_{m \rightarrow \infty} \bigcap_{n \geq m} A_n \\ &= \bigcup_m \bigcap_{n \geq m} A_n. \end{aligned}$$

Of course,  $\{A_n \text{ ev.}\} \subset \{A_n \text{ i.o.}\}$ , and

$$\mathbb{P}(A_n \text{ ev.}) \leq \mathbb{P}(A_n \text{ i.o.})$$

Recall that for real valued random variables  $X_n$  and  $X$ ,

$$\begin{aligned} \{X_n \rightarrow X\} &= \{\omega : X_n(\omega) \rightarrow X(\omega)\} \\ &= \{\omega : \forall \epsilon > 0, |X_n(\omega) - X(\omega)| \leq \epsilon \text{ eventually}\} \end{aligned}$$

Thus,

$$\begin{aligned} X_n \xrightarrow{a.s.} X &\Leftrightarrow \mathbb{P}\{\omega : \forall \epsilon > 0, |X_n(\omega) - X(\omega)| \leq \epsilon \text{ eventually}\} = 1 \\ &\Leftrightarrow \forall \epsilon > 0, \mathbb{P}(|X_n - X| \leq \epsilon \text{ ev.}) = 1 \\ &\Leftrightarrow \forall \epsilon > 0, \mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = 0 \end{aligned}$$

(The if and only if statement in the second line of this display holds because the event in the first line is the intersection of the events in the second line, and the intersection can be made countable by taking only rational  $\epsilon$ .)

Let the event  $A_n := \{|X_n - X| > \epsilon\}$ . Then, we are motivated to find useful conditions for  $\mathbb{P}(A_n \text{ i.o.}) = 0$ .

Recall that  $\{A_n \text{ i.o.}\} = \bigcap_n \bigcup_{m \geq n} A_m$ .

**Theorem 2.2.4** (Borel-Cantelli Lemmas). *Let  $(\Omega, F, \mathbb{P})$  be a probability space and let  $(A_n)$  be a sequence of events in  $F$ . Then,*

1. *If  $\sum_n \mathbb{P}(A_n) < \infty$ , then  $\mathbb{P}(A_n \text{ i.o.}) = 0$ .*
2. *If  $\sum_n \mathbb{P}(A_n) = \infty$  and  $A_n$  are independent, then  $\mathbb{P}(A_n \text{ i.o.}) = 1$ .*

[In the following we will mostly use BCL(1), however BCL(2) is also sometimes useful and it should be noted that there are some substitutes for independence in BCL(2).]

*Proof.* (Of BCL 1)

$$\begin{aligned} \mathbb{P}(A_n \text{ i.o.}) &= \lim_{m \rightarrow \infty} \mathbb{P}(\bigcup_{n \geq m} A_n) \\ &\leq \lim_{m \rightarrow \infty} \sum_{n \geq m} \mathbb{P}(A_n) = 0 \quad \text{since } \sum_{i=1}^{\infty} \mathbb{P}(A_n) < \infty. \end{aligned}$$

□

*Proof.* (Of BCL II) Assume that  $\sum \mathbb{P}(A_n) = \infty$  and the  $A_n$ 's are independent. We will show that  $\mathbb{P}(A_n^c \text{ ev.}) = 0$ , which implies that  $\mathbb{P}(A_n \text{ i.o.}) = 1$ .

$$\mathbb{P}(A_n^c \text{ ev.}) = \lim_{n \rightarrow \infty} \mathbb{P}(\cap_{m \geq n} A_m^c) = \lim_{n \rightarrow \infty} \prod_{m \geq n} \mathbb{P}(A_m^c) \quad (2.1)$$

$$= \lim_{n \rightarrow \infty} \prod_{m \geq n} (1 - \mathbb{P}(A_m)) \leq \lim_{n \rightarrow \infty} \prod_{m \geq n} \exp(-\mathbb{P}(A_m)) \quad (2.2)$$

$$= \lim_{n \rightarrow \infty} \exp\left(-\sum_{m \geq n} \mathbb{P}(A_m)\right) = 0$$

since  $\sum_{m \geq n} \mathbb{P}(A_m) = \infty$ , for every  $n$ .

For (2.1), we used the following fact (due to the independence of  $A_n$ ):

$$\mathbb{P}(\cap_{m \geq n} A_m^c) = \lim_{N \rightarrow \infty} \mathbb{P}(\cap_{n \leq m \leq N} A_m^c) = \lim_{N \rightarrow \infty} \prod_{n \leq m \leq N} \mathbb{P}(A_m^c) = \prod_{n \leq m} \mathbb{P}(A_m^c).$$

For (2.2),  $1 - x \leq \exp(-x)$  was used.  $\square$

Here is an example that the assumption of independence in BCL(2) is essential. Consider the space  $\Omega = (0, 1]$  with the  $\sigma$ -algebra  $\mathcal{B}$  of Borel subsets, and the Lebesgue measure as  $\mathbb{P}$ . The events  $A_n = (0, 1/n] \in \mathcal{B}$ . Then,  $\mathbb{P}(A_n) = 1/n$ ,  $\sum \mathbb{P}(A_n) = \infty$ , but  $\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}(\emptyset) = 0$ .

*Example 2.2.5.* Consider random walk in  $\mathbb{Z}^d$ , where  $S_0 = 0$ ,  $S_n = X_1 + \dots + X_n$ , for  $n = 1, \dots$ , and  $X_i$  are i.i.d. variables in  $\mathbb{Z}^d$ . In the simplest case, each  $X_i$  has uniform distribution on  $2^d$  possible  $\pm$  vectors. For example, if  $d = 3$ , we have  $2^3 = 8$  neighbors

$$\left\{ \begin{array}{c} (+1, +1, +1) \\ \vdots \\ (-1, -1, -1) \end{array} \right\}.$$

Note that each coordinate of  $S_n$  does a simple coin-tossing walk independently. It is true that

$$\mathbb{P}(S_n = 0 \text{ i.o.}) = \begin{cases} 1 & \text{if } d = 1 \text{ or } 2 \quad (\text{recurrent}), \\ 0 & \text{if } d \geq 3 \quad (\text{transient}), \end{cases} \quad (2.3)$$

and we can prove the transiency by using the Borel-Cantelli lemma.

*Proof of Transience for  $d \geq 3$ .* Let us start with  $d = 1$ , then  $\mathbb{P}(S_{2n+1} = 0) = 0$  and

$$\mathbb{P}(S_{2n} = 0) = \mathbb{P}(n \text{ “+1” and } n \text{ “-1” among } X_n) \quad (2.4)$$

$$= \binom{2n}{n} 2^{-2n} \quad (2.5)$$

$$\sim \frac{c}{\sqrt{n}} \text{ as } n \rightarrow \infty. \quad (2.6)$$

where we used the independence of  $X_n$  and the Stirling formula to approximate the binomial coefficient.

Since the coordinates of  $S_{2n}$  are independent for a fixed  $n$ , we have more generally

$$\sum_{n=0}^{\infty} \mathbb{P}(S_n = 0) \approx \sum_{n=0}^{\infty} \left( \frac{c}{\sqrt{n}} \right)^d = \begin{cases} \infty & d = 1, 2 \\ A_d < \infty & d = 3, 4, \dots \end{cases} \quad (2.7)$$

Thus, for  $d \geq 3$ ,  $\sum_n \mathbb{P}(S_{2n} = 0) < \infty$ , and (BC I) implies transiency.  $\square$

We cannot use this method to prove recurrency for  $d \leq 2$ , since the events  $S_{2n} = 0$  are dependent. The recurrency in this case is proved by other methods, which allow one to handle the events that are dependent in a certain controlled way. Usually, this is done by tools from the theory of Markov chains.

### 2.2.3 SLLN with with finite 4-th moment



The following is a version of the Law of Large Numbers, in which the convergence holds almost surely.

**Theorem 2.2.6.** *If  $X_1, \dots, X_n, \dots$ , is a sequence of independent identically distributed random variables with  $\mathbb{E}|X_i|^4 = C < \infty$ , then*

$$\lim_{n \rightarrow \infty} \frac{S_n}{n} = \lim_{n \rightarrow \infty} \frac{X_1 + \dots + X_n}{n} = \mathbb{E}(X_1)$$

with probability 1.

*Proof.* We can assume without loss of generality that  $\mathbb{E}[X_i] = 0$ . Otherwise, just take  $Y_i = X_i - \mathbb{E}[X_i]$ .

A simple calculation shows

$$\mathbb{E}[(S_n)^4] = n\mathbb{E}[(X_1)^4] + 3n(n-1)\mathbb{E}[(X_1)^2]^2 \leq nC + 3n^2\sigma^4,$$

and by applying a Chebychev type inequality using fourth moments,

$$\mathbb{P}\left[\frac{|S_n|}{n} \geq \delta\right] = \mathbb{P}[|S_n|^4 \geq (n\delta)^4] \leq \frac{nC + 3n^2\sigma^4}{n^4\delta^4}.$$

Hence,

$$\sum_{n=1}^{\infty} \mathbb{P}\left[\frac{|S_n|}{n} \geq \delta\right] < \infty,$$

and we can now apply the Borel-Cantelli Lemma (BC I) to conclude that these events will happen only finitely many times.  $\square$

The weak convergence of  $S_n/n$  holds whenever i.i.d  $X_n$  have finite mean. The result about strong convergence of  $S_n/n$  in Theorem 2.2.6 assumes that the fourth moment of  $X_i$  exists. Can this assumption be weakened?

The method we used relied on the existence of the fourth moment. If we will try to repeat it with the second moment instead of the fourth, then we find that

$$\mathbb{P}\left[\frac{|S_n|}{n} \geq \delta\right] = \mathbb{P}[|S_n|^2 \geq (n\delta)^2] \leq \frac{C}{n\delta^2},$$

and the sum of the probabilities does not converge.

On the other hand, if we consider partial sums  $S_{n^2}$  instead of  $S_n$ , we would find that

$$\mathbb{P}\left[\frac{|S_{n^2}|}{n^2} \geq \delta\right] = \mathbb{P}[|S_{n^2}|^2 \geq (n^2\delta)^2] \leq \frac{C}{n^2\delta^2}.$$

Hence the subsequence  $S_{n^2}/n^2$  converges almost surely to 0.

After some reflection, one can notice that it is possible to extract the almost surely convergent subsequence from any given subsequence of  $S_n/n$ .

This gives some hope that it is possible to prove the Strong Law of Large Numbers using the following theorem.

**Theorem 2.2.7.** *Let  $y_n$ ,  $n = 1, 2, \dots$ , be a sequence of elements of a topological space. If every subsequence  $y_m$ ,  $\{m\} = A \subset \mathbb{N}$  has a further subsequence  $y_l$ ,  $\{l\} \subset A$  that converges to  $y$ , then  $y_n \rightarrow y$ .*

*Proof.* If  $y_n \not\rightarrow y$  then there is an open set  $G$  containing  $y$  and a subsequence  $y_{n_m}$  with  $y_{n_m} \notin G$  for all  $m$ . But then clearly no subsequence of  $y_{n_m}$  converges to  $y$ .  $\square$

Unfortunately, this method does not work since one can show the following result.

**Lemma 2.2.8.**  $X_n \xrightarrow{\mathbb{P}} X$  if and only if for every subsequence  $X_{n_m}$  there exists a further subsequence  $X_{n_{m_k}}$  that converges almost surely to  $X$ .

We will prove only forward direction.

*Proof of forward direction:* If  $X_n \xrightarrow{\mathbb{P}} X$  then there exists  $\epsilon_k \downarrow 0$  such that  $\sum_k \mathbb{P}(|X_{n_k} - X| > \epsilon_k) < \infty$ . For example, we can take  $\epsilon_k = 1/k$  and choose  $n_k$  so that  $\mathbb{P}(|X_{n_k} - X| > 1/k) \leq 1/2^k$ . Then,  $\sum_k \mathbb{P}(|X_{n_k} - X| > \epsilon_k) < \infty$ , and by **BCL I** we can conclude that  $X_{n_k} \rightarrow X$  a.s.  $\square$

This result implies that almost sure convergence of random variables does not come from a topology on the space of random variables.


Indeed, if it were, then Lemma 2.2.8 and Theorem 2.2.7 would jointly imply that every sequence, which converges in probability, also converges almost surely. However, we know that this is not true.

To summarize, the Theorem 2.2.7 does not seem to be helpful for proving Strong Law of Large numbers under weaker conditions. In fact, Kolmogorov

has found an ingenious way to use the convergence of subsequences to prove the Strong Law of Large Numbers under the assumption of the finite second moment. This argument can be found in Appendix.

However, there is another, even more powerful method which shows that the existence of the first moment is sufficient for the validity of strong law of large numbers. This method is due to Kolmogorov and Khinchin.

### 2.2.4 Kolmogorov's SLLN

 Can one prove the SLLN under the same assumptions as in the WLLN, – identically distributed random variables  $X_i$  with finite mean? Surprisingly, the answer is “yes”. The proof introduces a new important idea that the convergence of averages is closely related to the convergence of certain series. However, even with this idea in mind, the proof works as a bit of magic.

**Theorem 2.2.9** (Kolmogorov's SLLN). *Let  $X_1, X_2, \dots$  be i.i.d. with  $\mathbb{E}|X_i| < \infty$ ,  $\mathbb{E}X_i = 0$ , and let  $S_n = X_1 + \dots + X_n$ . Then  $S_n/n \rightarrow 0$  with probability 1 as  $n \rightarrow \infty$ .*

The plan of the proof is as follows: by truncation and centering we will introduce a new sequence of random variables  $\hat{Y}_i$  which are independent although not identically distributed. These new random variables will have zero mean and the following properties:

1.  $n^{-1} \sum_{i=1}^n \hat{Y}_i$  converges a.s. if and only if  $n^{-1} \sum_{i=1}^n X_i$  converges a.s.;
2. If these two series converge they converge to the same limit, and
3. The series  $\sum_{k=1}^{\infty} \text{Var} \hat{Y}_k / k^2 < \infty$ .

Then we apply another of the great Kolmogorov's theorems, that say that the convergence of series  $\sum_{k=1}^{\infty} \text{Var} \hat{Y}_k / k^2$  implies the almost sure convergence of the series  $\sum_{k=1}^{\infty} \hat{Y}_k / k$ . Then we will use a classical lemma by Kronecker that says the the convergences of these series implies that the sequence  $n^{-1} \sum_{k=1}^n \hat{Y}_k \rightarrow 0$ , and we are done.

We will need two lemmas that relate convergence of averages to convergence of sequences and series.

**Lemma 2.2.10** (Toeplitz). *Suppose that a sequence  $\{x_n\}$  converges to  $x$ . Then,*

$$\frac{x_1 + \dots + x_n}{n} \rightarrow x.$$

*Proof.* For an  $\varepsilon > 0$ , let  $n_0$  be such that  $|x_j - x| < \varepsilon/2$  for all  $j > n_0$ . Then, choose  $n_1 > n_0$  such that  $\frac{1}{n_1} \sum_{j=1}^{n_0} |x_j - x| < \varepsilon/2$ . Then, for  $n > n_1$ ,

$$\left| \frac{1}{n} \sum_{j=1}^n x_j - x \right| \leq \frac{1}{n_1} \sum_{j=1}^{n_0} |x_j - x| + \frac{1}{n} \sum_{j=n_0+1}^n |x_j - x| \leq \varepsilon.$$

□

**Lemma 2.2.11** (Kronecker). *Suppose that  $\{y_j\}$  is a sequence of numbers such that  $\sum(y_j/j)$  converges. Then,*

$$\frac{y_1 + y_2 + \dots + y_n}{n} \rightarrow 0.$$

*Proof.* Let  $S_n = \sum_{j=1}^n (y_j/j)$  for  $n > 0$  and  $S_0 = 0$ . Then

$$\sum_{j=1}^n y_j = \sum_{j=1}^n j(S_j - S_{j-1}) = nS_n - \sum_{j=1}^n S_{j-1}.$$

Hence,

$$\frac{1}{n} \sum_{j=1}^n y_j = S_n - \frac{1}{n} \sum_{j=1}^n S_{j-1}.$$

By assumption,  $S_n$  converges to a number  $x$ . Hence, by the Toeplitz lemma,  $\frac{1}{n} \sum_{j=1}^n S_{j-1} \rightarrow x$ , and  $\frac{1}{n} \sum_{j=1}^n y_j \rightarrow 0$ .  $\square$

*Proof of Theorem 2.2.9.* We define truncated random variables

$$Y_n = \begin{cases} X_n, & \text{if } |X_n| \leq n, \\ 0, & \text{if } |X_n| \geq n, \end{cases}$$

and their centered variants,  $\hat{Y}_n = Y_n - \mathbb{E}Y_n$ . and also define

$$\begin{aligned} a_n &= \mathbb{P}[X_n \neq Y_n], \\ b_n &= \mathbb{E}[Y_n], \\ c_n &= \mathbb{V}\text{ar}(Y_n). \end{aligned}$$

First we note that

$$\sum_n a_n = \sum_n \mathbb{P}[|X_1| \geq n] \leq \mathbb{E}|X_1| < \infty.$$

(The inequality in the middle is an easy exercise.) By the Borel-Cantelli lemma (BC1), this implies that  $\mathbb{P}(X_n \neq Y_n \text{ i.o.}) = 0$ . In particular this means that  $\frac{S_n}{n} \rightarrow 0$  a.s. if and only if  $(Y_1 + \dots + Y_n)/n \rightarrow 0$  a.s.

Then, the biases of the truncated variables go to zero.

$$\lim_{n \rightarrow \infty} b_n = 0,$$

because

$$\begin{aligned} |\mathbb{E}Y_n| &= |\mathbb{E}(Y_n - X_n)| \leq \int_{|X_n| > n} |X_n(\omega)| d\mathbb{P}(\omega) \\ &= \int_{|X_1| > n} |X_1(\omega)| d\mathbb{P}(\omega) \rightarrow 0, \text{ as } n \rightarrow \infty. \end{aligned}$$

By the Toeplitz lemma, this implies that  $\frac{1}{n} \sum_{j=1}^n b_j \rightarrow 0$  and therefore,  $(Y_1 + \dots + Y_n)/n \rightarrow 0$  a.s. if and only if  $((Y_1 - b_1) + \dots + (Y_n - b_n))/n \rightarrow 0$  a.s.

Finally, observe that the variances of the truncated variables,  $c_j = \text{Var}(Y_j)$ , cannot grow too fast,

$$\sum_n \frac{c_n}{n^2} < \infty.$$

Indeed, let  $\alpha$  denote the common distribution of  $X_i$ . Then,

$$\begin{aligned} \sum_n \frac{c_n}{n^2} &\leq \sum_n \frac{\mathbb{E}Y_n^2}{n^2} = \sum_n \int_{|x| \leq n} \frac{x^2}{n^2} d\alpha \\ &= \int x^2 \left( \sum_{n \geq x} \frac{1}{n^2} \right) d\alpha \leq C \int |x| d\alpha < \infty. \end{aligned}$$

In the last inequality, we used the fact that

$$\sum_{n \geq x} \frac{1}{n^2} \leq \frac{C}{x}$$

for all  $x \geq 1$ , for a suitable choice of  $C$ .

By the Kronecker lemma, the convergence  $((Y_1 - b_1) + \dots + (Y_n - b_n))/n \rightarrow 0$  would follow from the convergence of the series  $\sum_{n=1}^{\infty} (Y_n - b_n)/n$ . We conclude the proof by noting that the condition  $\sum_n \frac{\text{Var}Y_n}{n^2} < \infty$  implies the almost sure convergence of  $\sum_{n=1}^{\infty} (Y_n - b_n)/n$  by Theorem 2.2.12, which we prove below.  $\square$

**Theorem 2.2.12** (Kolmogorov and Khinchin). *Let  $X_n$  be a sequence of independent random variables and  $\mathbb{E}X_n = 0$ . Then if  $\sum_n \mathbb{E}X_n^2 < \infty$ , then the series  $\sum_n X_n$  converges with probability 1.*

The proof of this important theorem is based on the Kolmogorov's maximal inequality. Let  $S_k = \sum_{j=1}^k X_j$  and define

$$T_n(\omega) = \sup_{1 \leq k \leq n} |S_k(\omega)| = \sup_{1 \leq k \leq n} \left| \sum_{j=1}^k X_j(\omega) \right|.$$

**Theorem 2.2.13** (Kolmogorov's Inequality). *Assume that  $\mathbb{E}X_i = 0$  and  $\text{Var}(X_i) = \sigma_i^2 \leq \infty$  and let  $s_n^2 = \sum_{j=1}^n \sigma_j^2$ . Then*

$$\mathbb{P}\{T_n \geq l\} \leq \frac{s_n^2}{l^2}.$$

The important point here is that the estimate depends only on  $s_n^2$  and not on the number of summands. In fact the Chebyshev bound on  $S_n$  is

$$\mathbb{P}\{|S_n| \geq l\} \leq \frac{s_n^2}{l^2}$$

and therefore the supremum over  $k$  does not cost anything.



*Proof.* Let us define the events

$$E_k = \{|S_1| < l, \dots, |S_{k-1}| < l, |S_k| \geq l\}.$$

Then  $\{T_n \geq l\}$  is a disjoint union of  $E_k$ . If we use the independence of  $S_n - S_k$  and  $S_k \mathbb{1}_{E_k}$  that only depends on  $X_1, \dots, X_k$ , then we can write

$$\begin{aligned} \mathbb{P}\{E_k\} &\leq \frac{1}{l^2} \int_{E_k} S_k^2 dP \\ &\leq \frac{1}{l^2} \int_{E_k} \left( S_k^2 + (S_n - S_k)^2 \right) dP \\ &= \frac{1}{l^2} \int_{E_k} \left( S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2 \right) dP \\ &= \frac{1}{l^2} \int_{E_k} S_n^2 dP \end{aligned}$$

Summing over  $k$  from 1 to  $n$ ,

$$\mathbb{P}\{T_n \geq l\} \leq \frac{1}{l^2} \int_{T_n \geq l} S_n^2 dP \leq \frac{s_n^2}{l^2}.$$

□

The proof of Theorem 2.2.12 is based on the fact that the sequence  $\{S_n = X_1 + \dots + X_n\}$  converges to  $S$  if and only if

$$\mathbb{P}\left\{ \sup_{p,q \geq n} |S_p - S_q| \geq \varepsilon \right\} \rightarrow 0,$$

as  $n \rightarrow \infty$  for every  $\varepsilon > 0$ .

Recall that a sequence of numbers  $\{\xi_n\}$  is called Cauchy (or fundamental) if  $M_n := \sup_{p,q \geq n} |\xi_p - \xi_q| \rightarrow 0$  as  $n \rightarrow \infty$ , and that it is a fact of real analysis that a sequence is convergent if and only if it is Cauchy.

**Lemma 2.2.14.** *A random sequence  $\xi_n$  is Cauchy with probability 1 if and only if for all  $\varepsilon > 0$ ,  $\mathbb{P}[M_n \geq \varepsilon] \rightarrow 0$  as  $n \rightarrow \infty$ .*

*Proof.* Let  $B_{k,l}^\varepsilon = \{\omega : |\xi_k - \xi_l| \geq \varepsilon\}$ , and

$$B^\varepsilon = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n, l \geq n} B_{k,l}^\varepsilon.$$

Then  $\xi_n(\omega)$  is not fundamental if and only if  $\omega \in \cup_{\varepsilon > 0} B^\varepsilon$ . By continuity of probability measure,  $\mathbb{P}(\cup_{\varepsilon > 0} B^\varepsilon) = 0$  if and only if  $\mathbb{P}(B^\varepsilon) = 0$  for all  $\varepsilon > 0$ . This holds if and only if for all  $\varepsilon \geq 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left[ \bigcup_{k \geq n, l \geq n} B_{k,l}^\varepsilon \right] = 0,$$

and this exactly the condition that  $\mathbb{P}[M_n \geq \varepsilon] \rightarrow 0$  as  $n \rightarrow \infty$ . □

*Proof of Theorem 2.2.12.* We want to show that  $S_n = X_1 + \dots + X_n$  is Cauchy a.s. By the previous lemma it is enough to show that  $\mathbb{P}(M_n > \epsilon) \rightarrow 0$  for all  $\epsilon > 0$ , where  $M_n = \sup_{p,q \geq n} |S_p - S_q|$ .

Let  $M_n^* := \sup_{p \geq n} |S_p - S_n|$ . By the triangle inequality,

$$|S_p - S_q| \leq |S_p - S_n| + |S_q - S_n| \Rightarrow M_n^* \leq M_n \leq 2M_n^*,$$

so it is sufficient to show that  $M_n^* \xrightarrow{P} 0$ .

For all  $\epsilon > 0$ ,

$$\begin{aligned} \mathbb{P}\left(\sup_{p \geq n} |S_p - S_n| > \epsilon\right) &= \lim_{N \rightarrow \infty} \mathbb{P}\left(\max_{n \leq p \leq N} |S_p - S_n| > \epsilon\right) \\ &\leq \lim_{N \rightarrow \infty} \sum_{i=n+1}^N \frac{\sigma_i^2}{\epsilon^2} = \sum_{i=n+1}^{\infty} \frac{\sigma_i^2}{\epsilon^2} \end{aligned}$$

where we applied Kolmogorov's inequality in the second step. Since by assumption,  $\sum_{i=1}^{\infty} \sigma_i^2 < \infty$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\sup_{p \geq n} |S_p - S_n| > \epsilon\right) = 0,$$

the sequence  $S_n$  is Cauchy with probability 1, and therefore  $S_n$  converges almost surely.  $\square$

*Remark:* Just orthogonality rather than independence of the  $X_i$ s is not enough to get an a.s. limit. Counterexamples are hard. According to classical results of Rademacher-Menchoff, for orthogonal  $X_i$  the condition

$$\sum_i (\log^2 i) \sigma_i^2 < \infty$$

is enough for a.s. convergence of  $S_n$ , whereas if  $b_i \uparrow$  with  $b_i = o(\log^2 i)$  there exist orthogonal  $X_i$  such that  $\sum_i b_i \sigma_i^2 < \infty$  and  $S_n$  diverges almost surely.

When we have some additional information on the moments of the random variables  $X_n$ , the strong law can be improved by using Kolmogorov's method.

**Theorem 2.2.15.** *Let  $X_1, X_2, \dots$  be i.i.d. with  $\mathbb{E}X_i = 0$ ,  $\mathbb{E}X_i^2 = \sigma^2 < \infty$ , and let  $S_n = X_1 + \dots + X_n$ . If  $\epsilon > 0$ , then*

$$\frac{S_n}{n^{1/2}(\log n)^{1/2+\epsilon}} \rightarrow 0$$

with probability 1 as  $n \rightarrow \infty$ .

*Proof.* Let  $a_n = n^{1/2}(\log n)^{1/2+\epsilon}$  for  $n \geq 2$  and  $a_1 > 0$ . Then,

$$\sum_n \mathbb{V}\text{ar}(X_n/a_n) = \sigma^2 \left( \frac{1}{a_1^2} + \sum_{n \geq 2} \frac{1}{n(\log n)^{1+2\epsilon}} \right) < \infty.$$

So by Theorem 2.2.12,  $\sum_n X_n/a_n$  converges with probability 1, and an application of the Kronecker Lemma delivers the result.  $\square$

## 2.2.5 Connection to Ergodic Theorem

We can think about the SLLN as a consequence of an ergodic theorem.

Recall that if  $(M, \mathcal{B}, \mu)$  is a probability space, then  $T : M \rightarrow M$  is a *measure-preserving* transformation of  $M$ , if it is measurable and  $\mu(T^{-1}A) = \mu(A)$  for any  $A \in \mathcal{B}$ . Transformation  $T$  is called *ergodic* if every invariant subset have measure 0 or 1. The Birkhoff ergodic theorem says that for any measurable function  $f$  and ergodic  $T$ , we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \int f(x) \mu(dx)$$

for  $\mu$ -almost all  $x$ . This has a striking resemblance to Kolmogorov's SLLN. In fact, it turns out that one can derive the SLLN as a consequence of the ergodic theorem by defining a suitable space  $M$ .

If we have a sequence of i.i.d. real-valued random variables  $X_i$  defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ , then can define a product space  $M = \prod_{i=1}^{\infty} \mathbb{R}$  with the product  $\sigma$ -algebra  $\widehat{\mathcal{B}}$ , and the measure  $\mu$ , which is a product of the distribution measures  $\mu_i$  of the random variables  $X_i$ . We define the functions  $Y_i : M \rightarrow \mathbb{R}$  that map  $(\widehat{\omega}_1, \widehat{\omega}_2, \dots)$  to  $\widehat{\omega}_i$  and observe that  $X_i$  and  $Y_i$  have the same distribution:

$$\mu(Y_1 \in A_1, \dots, Y_n \in A_n) = \mathbb{P}(X_1 \in A_1, \dots, X_n \in A_n).$$

We also have a shift operator  $T$  on  $M$  that maps  $(\omega_1, \omega_2, \omega_3, \dots)$  to  $(\omega_2, \omega_3, \omega_4, \dots)$ . It is clear that this operator is measure-preserving in the sense that  $\mu(T^{-1}(A)) = \mu(A)$ , for any measurable set  $A \in \widehat{\mathcal{F}}$ .

**Lemma 2.2.16.** *Transformation  $T$  is ergodic.*

This requires a proof and the ideas of the proof are similar to the ideas behind the proof of the Kolmogorov's zero-one theorem. What is used is that  $\mu$  is a product measure generated by cylinder sets. See a book on ergodic theory for a proof.

In addition,  $Y_i(\widehat{\omega}) = \widehat{\omega}_i = Y_1(T^{i-1}\widehat{\omega})$ .

So by Birkhoff's ergodic theorem, it follows that

$$\frac{1}{n} \sum_{i=1}^n Y_i(\widehat{\omega}) = \frac{1}{n} \sum_{k=0}^{n-1} Y_1(T^k \widehat{\omega}) \rightarrow \widehat{\mathbb{E}} Y_1 = \mathbb{E} X_1,$$

$\mu$  almost surely. It remains to get back from measure  $\mu$  and functions  $Y_i$  to random variables  $X_i$  and measure  $\mathbb{P}$ .


Let  $C \subset M$  be the set of sequences  $(x_i) \in M$  such the  $\frac{1}{n} \sum_{i=1}^n x_i$  converges to  $\mathbb{E}(X_1)$ . Then the assertion  $\frac{1}{n} \sum_{i=1}^n X_i$  converges to  $\mathbb{E}(X_1)$  almost surely means that  $\mathbb{P}(\omega, (X_k(\omega)) \in C) = 1$ . This is equivalent to the statement that  $\mu(\widehat{\omega}, (Y_k(\widehat{\omega})) \in C) = 1$  because  $X_k$  and  $Y_k$  have the same distribution measure  $\mu$  on  $M$ . And as we have just seen,  $\mu(\widehat{\omega}, (Y_k(\widehat{\omega})) \in C) = 1$  because of Birkhoff's ergodic theorem.



## Chapter 3

# Central Limit Theorem

### 3.1 Convergence in Distribution

The strong law of large numbers, the ergodic theorem, and similar results state that a certain property holds almost surely or with probability 1. One problem is that for uncountably infinite spaces  $\Omega$ , they are not verifiable experimentally: it is not possible to observe an infinite sequence.

In addition, if the convergence holds for almost every point in an interval, it will not tell us if it holds on a specific point, or at all rational points, or at all algebraic points. All these point sets have the Lebesgue measure of zero.

So, from the practical point of view we might be more interested in learning something about the distribution of random sums with many terms rather than in proving that almost always these sums converge to zero.

In this chapter we will define the convergence in distribution, learn how to prove this convergence by using appropriate test function, and derive a prototypical limit theorem, the Central Limit Theorem for random sums.

#### 3.1.1 Definitions and Skorohod's theorem

**Definition 3.1.1.** Let  $X_n$  be a sequence of real-valued random variables and  $F_{X_n}(x)$  be their distribution functions,  $F_{X_n}(x) = \mathbb{P}(X_n \leq x)$ . Then  $X_n$  converges in distribution to  $X$ ,  $X_n \xrightarrow{d} X$  if  $F_{X_n}(x) \rightarrow F_X(x)$  for all  $x$  at which  $F_X(x)$  is continuous.

We call this type of convergence of random variables *convergence in distribution* or *weak convergence*.

*Note.* This is really a notion of convergence of distribution measures of  $X_n$  rather than of convergence of random variables  $X_n$  themselves. The random variables  $X_n$  can even be defined on different spaces.

The convergence almost surely and convergence in probability imply the convergence in distribution (exercise). In order to talk about reverse implication

we need to make sure that random variables are defined on the same probability space.

Let us write  $Y \stackrel{d}{=} X$  to denote that random variables  $Y$  and  $X$  have the same distribution.

**Theorem 3.1.2** (Skorokhod).  $X_n \xrightarrow{d} X \iff$  there exists a probability space with random variables  $Y_n$  and  $Y$ , such that  $Y_n \stackrel{d}{=} X_n$ ,  $Y \stackrel{d}{=} X$  and  $Y_n \xrightarrow{a.s.} Y$ .

*Proof.* Let  $\Omega = [0, 1]$ ,  $\mathcal{F}$  is the Borel sets and  $\mathbb{P}$  is the Lebesgue measure. Define  $Y_n$  and  $Y$  as follows. Let

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x), \\ F(x) &= \mathbb{P}(X \leq x), \end{aligned}$$

and define

$$\begin{aligned} Y_n &= F_n^{-1}(\omega) = \sup\{y : F_n(y) < \omega\}, \\ Y &= F^{-1}(\omega) = \sup\{y : F(y) < \omega\}. \end{aligned}$$

See the book for the details of the proof. □

The following result is an application of the above.

**Theorem 3.1.3.**  $X_n \xrightarrow{d} X$  if and only if for every bounded continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)].$$

*Proof.* If  $X_n \xrightarrow{d} X$ , then by using Skorokhod's theorem we can assume that  $X_n \xrightarrow{a.s.} X$ . Then by continuity of  $f$ , we have  $f(X_n) \xrightarrow{a.s.} f(X)$ , and also it is easy to check that the sequence  $f(X_n)$  is almost surely bounded. Therefore, we can take expectations and use the bounded convergence theorem.

In the opposite direction, see Resnick. □

Using this property we can give a different definition of the convergence in distribution, which applies to a more general class of random variables.

**Definition 3.1.4.** Let  $S$  be a metric space, and  $\mathcal{B}$  be the Borel  $\sigma$ -field on  $S$ . Let  $P_1, P_2, \dots$  be a sequence of probability measures on  $(S, \mathcal{B})$ . Say  $P_n$  converges in distribution to  $P$  or  $P_n \xrightarrow{d} P$  for some probability measure  $P$  on  $(S, \mathcal{B})$ , if

$$\int f dP_n \rightarrow \int f dP$$


for every bounded continuous function  $f : S \rightarrow \mathbb{R}$ .

This type of convergence is also often called *weak convergence* or *weak\* convergence*. The limits are unique because one can show that if

$$\int f dP = \int f dQ \text{ for all bounded continuous } f,$$

then  $P(A) = Q(A)$  for all  $A \in \mathcal{B}$ .

### 3.1.2 Characterization of Weak Convergence

**Theorem 3.1.5.**  Let  $P_n, n = 1, 2, \dots$ , and  $P$  be probability measures on  $\mathbb{R}$ . The following are equivalent:

1.  $\int f dP_n \rightarrow \int f dP$  for all bounded continuous  $f$ ;
2. Same for all infinitely differentiable functions with all bounded derivatives  $C^\infty$ ;
3.  $P_n(-\infty, x] \rightarrow P(-\infty, x]$  for all  $x$  at which  $x \rightarrow P(-\infty, x]$  is continuous;
4. Condition 1 for all  $f$  such that  $f$  is bounded and continuous except on a set of  $P$  measure 0; and
5. Condition 3 with  $P_n(C)$  for all  $C$  closed, or with  $P_n(O)$  for all  $O$  open.

*Proof.* 1  $\Rightarrow$  3:

Define  $f_{u,v}$  by

$$f_{u,v}(x) = \begin{cases} 1 & \text{if } x \leq u \\ 0 & \text{if } x \geq v \\ \text{linear} & \text{if } u \leq x \leq v. \end{cases}$$

For  $\epsilon > 0$ ,

$$f_{x-\epsilon, x} \leq \mathbb{1}(-\infty, x] \leq f_{x, x+\epsilon}.$$

Write  $Pf$  for  $\int f dP$ . So if  $P_n \rightarrow P$ , then

$$P_n f_{x-\epsilon, x} \leq P_n(-\infty, x] \leq P_n f_{x, x+\epsilon}.$$

Let  $n \rightarrow \infty$ ,

$$Pf_{x-\epsilon, x} \leq \liminf_{n \rightarrow \infty} P_n(-\infty, x] \leq \limsup_{n \rightarrow \infty} P_n(-\infty, x] \leq Pf_{x, x+\epsilon}$$

and

$$P(-\infty, x - \epsilon] \leq Pf_{x-\epsilon, x} \leq P(-\infty, x] \leq Pf_{x, x+\epsilon} \leq P(-\infty, x + \epsilon].$$

Now assume  $y \rightarrow P(-\infty, y]$  is continuous at  $y = x$ . Let  $\epsilon \rightarrow 0$ , we see that by taking  $\epsilon$  sufficiently small we can make  $Pf_{x-\epsilon, x}$  and  $Pf_{x, x+\epsilon}$  as close as we like to  $P(-\infty, x]$ . Then we can conclude

$$\liminf_{n \rightarrow \infty} P_n(-\infty, x] = \limsup_{n \rightarrow \infty} P_n(-\infty, x] = P(-\infty, x].$$

3  $\Rightarrow$  1:

To show this we use another approximation. 3 gives

$$P_n f \rightarrow P f \text{ for } f = \mathbb{1}(-\infty, x] \text{ where } x \text{ is a continuity point of the distribution of } P. \quad (3.1)$$

First observe that the set of continuity points of  $P$  is dense in  $\mathbb{R}$ , as there are only countably many jumps of  $x \rightarrow P(-\infty, x]$ . Second, note that we can extend (3.1) from indicators to finite linear combinations of such indicators, i.e. to step functions.

Now, let  $f$  be continuous and bounded in magnitude by  $M$ . Choose some target  $\epsilon > 0$  and choose  $B$  so that  $B$  and  $-B$  are both continuity points of the limit distribution  $x \rightarrow P(-\infty, x]$  and  $P(-B, B]^c < \epsilon$ . Note that there exists  $n(\epsilon)$  such that  $P_n(-B, B]^c < 2\epsilon$  for all  $n \geq n(\epsilon)$ .

Next, choose a step function  $s$  so that

$$|s(x) - f(x)| \leq \epsilon$$

for all  $x \in (-B, B]$  and  $s = 0$  outside  $(-B, B]$  (this can be done by uniform continuity of  $f$  on  $[-B, B]$ ).

Also

$$|P_n f - P_n s| \leq 2\epsilon M + \epsilon \quad (3.2)$$

for  $n \geq n(\epsilon)$ . (Note that our  $s$  depends on  $\epsilon$ .)

Choose  $n$  even larger so that  $|P_n s - P s| \leq \epsilon$ . Thus, by the triangle inequality,


$$|P_n f - P s| \leq 2\epsilon M + 2\epsilon.$$

We also have (3.2) for  $n \rightarrow \infty$ , so we can replace  $P_n$  by  $P$  and put it all together to get

$$|P_n f - P f| \leq 4\epsilon M + 4\epsilon$$

for all sufficiently large  $n$ . □

## 3.2 Characteristic functions

 In the previous section we have seen that a sequence of measures  $\mu_n$  converges weakly if for every test function from a suitable family (continuous functions, indicators of  $(-\infty, x]$  and so on), the integrals against  $\mu_n$  converge to the integral against the limit measure  $\mu$ . A very useful family of test functions is given by functions  $f_t(x) = e^{itx}$ , where  $t$  is a parameter.

**Definition 3.2.1.** If  $\mu$  is a probability measure on the real line  $\mathbb{R}$ , then its *characteristic function* is defined by

$$\varphi(t) = \int e^{itx} d\mu.$$



It is easy to see that this is a bounded complex-valued function  $|\varphi(t)| < 1$ , well-defined for every probability measure  $\mu$ . The fact that the characteristic function is well-defined for any  $\mu$  is the main benefit of it over the moment-generating function  $m_X(T) := \mathbb{E}(e^{tX})$ .

One of the most important properties of characteristic functions is that it can be well approximated uniformly by its Taylor series.

**Theorem 3.2.2.** *Let  $X$  be a random variable with characteristic function  $\varphi(t)$ . If  $\mathbb{E}|X|^n < \infty$  for some  $n \geq 1$ , then (i) the  $r$ -th derivative  $\varphi^{(r)}(t)$  exists everywhere for every  $r \leq n$ , (ii)  $\varphi^{(r)}(0) = i^r \mathbb{E}X^r$ , and (iii)*

$$\varphi(t) = \sum_{r=0}^n n \frac{(it)^r}{r!} \mathbb{E}X^r + \frac{(it)^n}{n!} \varepsilon_n(t),$$


where  $|\varepsilon(t)| \leq 3\mathbb{E}|X|^n$ , and  $\varepsilon_n(t) \rightarrow 0$ , as  $t \rightarrow 0$ .

The basis for applications of characteristic functions is the following theorem.

**Theorem 3.2.3** (Levy - Cramer Continuity Theorem). *Let  $\varphi_n(t)$  be the characteristic functions of measures  $\mu_n$  on the real line. If for every real  $t$ , the sequence  $\varphi_n(t)$  converges to  $\varphi(t)$ , which is a characteristic function of a measure  $\mu$ , then  $\mu_n$  converge weakly to  $\mu$ .*

## 3.3 Central Limit Theorem

### 3.3.1 Introduction

**Theorem 3.3.1** (CLT).  Let  $X_1, X_2, \dots$  be i.i.d. with  $\mathbb{E}(X_n) = \mu$ ,  $\text{Var}(x_n) = \sigma^2 < \infty$ . If  $S_n = X_1 + \dots + X_n$ , then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1)$$

This is a particular case of a more general theorem: the Lindeberg - Feller CLT.

There are two approaches to the proof of Central Limit Theorems. One is to show the convergence of the characteristic functions.

The second approach, due to Lindeberg, uses only the continuously differentiable test functions.

Let us prove here Theorem 3.3.1 by using characteristic functions.

*Proof.* Let  $\varphi(t) = \mathbb{E}e^{it(X_1 - \mu)}$ . Then, by independence,

$$\varphi_n(t) := \mathbb{E} \exp\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) = \left[\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

By Theorem 3.2.2,

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

for  $t \rightarrow 0$ . Hence, for a fixed  $t$  and  $n \rightarrow \infty$ ,

$$\varphi_n(t) = \left[ 1 - \frac{\sigma^2 t^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right) \right]^n \rightarrow e^{-t^2/2}.$$

Hence by Theorem 3.2.3, the distribution measure of  $\frac{S_n - n\mu}{\sigma\sqrt{n}}$  converges to the measure of the standard Gaussian random variable.  $\square$

### 3.3.2 Triangular Arrays

**I**f we study the sums of random variables which are independent but not necessarily identically distributed, then a language of triangular arrays is useful. Therefore, throughout this section we shall study the sequence of sums

$$S_i = \sum_j X_{ij}$$

obtained by summing the rows of a *triangular array* of random variables

$$\begin{array}{cccc} X_{11}, & X_{12}, & \dots, & X_{1n_1} \\ X_{21}, & X_{22}, & \dots, & X_{2n_2} \\ X_{31}, & X_{32}, & \dots, & X_{3n_3} \\ \vdots & \vdots & \vdots & \vdots \end{array}$$

(In the formula for  $S_i$ ,  $i$  ranges over  $\{1, 2, \dots\}$ , and  $j$  ranges over  $\{1, 2, \dots, n_i\}$ .)


It will be assumed throughout that the triangular arrays we consider satisfy *Three Triangular Array Conditions*:

1. For each  $i$ , the  $n_i$  random variables  $X_{i1}, X_{i2}, \dots, X_{in_i}$  in the  $i$ th row are mutually independent.
2.  $\mathbb{E}(X_{ij}) = 0$  for all  $i, j$ , and
3.  $\sum_j \mathbb{E}X_{ij}^2 = 1$  for all  $i$ .

We have some remarks for these conditions:

- It is *not* assumed that random variables in each row are identically distributed.
- It is *not* assumed that different rows are independent. In fact, they are not in a common application of triangular arrays to the study of sums  $S_n = X_1 + X_2 + \dots + X_n$ .
- It will usually be the case that  $n_i \rightarrow \infty$  as  $i \rightarrow \infty$ .

### 3.3.3 The Lindeberg Condition and Some Consequences

**Theorem 3.3.2** (Lindeberg's Theorem).  Suppose that in addition to the Triangular Array Conditions, a triangular array satisfies Lindeberg's condition:

$$\forall \epsilon > 0, \lim_{i \rightarrow \infty} \sum_{j=1}^{n_i} \mathbb{E}[X_{ij}^2 \mathbb{1}(|X_{ij}| > \epsilon)] = 0 \quad (3.3)$$

Then  $S_i \xrightarrow{d} \mathcal{N}(0, 1)$ .

The Lindeberg condition makes precise in what sense the random variables must be relatively negligible with respect to the sum for the CLT to hold. It says that for arbitrarily small fixed  $\epsilon > 0$ , the terms with absolute value greater than  $\epsilon$  contribute negligibly small to the total variance in a row  $i$ , as  $i$  increases.

Another natural condition is as follows:

$$\forall \epsilon > 0, \lim_{i \rightarrow \infty} \max_j \mathbb{P}(|X_{ij}| > \epsilon) = 0. \quad (3.4)$$


An array with property (3.4) is said to be *uniformly asymptotically negligible (UAN)*. One can show that this condition is implied by the Lindeberg's condition, but not vice-versa.

A converse to Lindeberg's Theorem is as follows:

**Theorem 3.3.3** (Feller's Theorem). *If a triangular array satisfies the Triangular Array Conditions and is UAN, then  $S_i \xrightarrow{d} \mathcal{N}(0, 1)$  (if and) only if Lindeberg's condition (3.3) holds.*

*Proof.* See Billingsley, Theorem 27.4, or Kallenberg, 5.12. □

### 3.3.4 The Lyapounov Condition

 A condition stronger (but often easier to check) than Lindeberg's is the *Lyapounov condition*:

$$\exists \delta > 0 \text{ such that } \lim_{i \rightarrow \infty} \sum_j \mathbb{E}|X_{ij}|^{2+\delta} = 0 \quad (3.5)$$

**Lemma 3.3.4.** *Lyapounov's condition implies Lindeberg's condition.*

*Proof.* Fix any  $\epsilon, \delta > 0$ . For any random variable  $|X| > \epsilon$ , we have

$$X^2 = \frac{|X|^{2+\delta}}{|X|^\delta} \leq \frac{|X|^{2+\delta}}{\epsilon^\delta}$$

Thus for any random variable  $X$  we have

$$\mathbb{E}[X^2 \mathbb{1}(|X| > \epsilon)] \leq \frac{\mathbb{E}|X|^{2+\delta}}{\epsilon^\delta}$$

Take  $X = X_{ij}$  to be the elements of our triangular array, and take  $\delta$  to be the value from Lyapounov's condition. Then we can sum over  $j$  on the RHS and take the limit as  $i \rightarrow \infty$  on both sides to get the Lindeberg's condition.  $\square$

**Theorem 3.3.5** (Lyapounov's Theorem). *If a triangular array satisfies the Triangular Array Conditions and the Lyapounov condition (3.5), then  $S_i \xrightarrow{d} \mathcal{N}(0, 1)$ .*

This follows from Lindeberg's Theorem, but we prove it with  $\delta = 1$  below.

### 3.3.5 Preliminaries to the proof of Lyapounov's Theorem

 We prove the Lyapounov's CLT by using the Lindeberg method and we need two preliminary facts. First:

**Lemma 3.3.6.** *If  $X \sim \mathcal{N}(0, \sigma^2)$ ,  $Y \sim \mathcal{N}(0, \tau^2)$  are independent, then  $X + Y \sim \mathcal{N}(0, \sigma^2 + \tau^2)$ .*

*Proof.* Either

1. use the formula for the convolution of densities, or
2. use characteristic or moment generating functions, or
3. use the radial symmetry of the joint density function of i.i.d.  $\mathcal{N}(0, \sigma^2 + \tau^2)$  random variables  $U$  and  $V$  to argue that  $U \sin \theta + V \cos \theta \sim \mathcal{N}(0, \sigma^2 + \tau^2)$ .

Take  $\sin(\theta) = \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right)^{1/2}$ .

To see how rotational invariance is unique to the normal distribution, see Kallenberg 13.2.


$\square$

Second:

**Lemma 3.3.7.**  *$S_i \xrightarrow{d} Z$  if and only if  $\lim_{i \rightarrow \infty} \mathbb{E}f(S_x) = \mathbb{E}f(Z)$  for all  $f \in \mathbf{C}_b^3(\mathbb{R})$ , the set of functions from  $\mathbb{R}$  to  $\mathbb{R}$  with three bounded, continuous derivatives.*

*Proof.* See Durrett, Theorem 2.2, and use that  $\mathbf{C}_b^3(\mathbb{R})$  is dense in  $\mathbf{C}_b(\mathbb{R})$ .  $\square$

### 3.3.6 Proof of Lyapounov's Theorem

 This proof illustrates the general idea of the proof of Lindeberg's theorem, and avoids a few tricky details which will be dealt with later.

*Proof.* With  $n$  fixed, let  $X_1, X_2, \dots, X_n$  be independent random variables, not necessarily identically distributed. Suppose  $\mathbb{E}X_j = 0$  and let  $\sigma_j^2 = \mathbb{E}(X_j^2) < \infty$ . Then for  $S = \sum_{j=1}^n X_j$  we have  $\sigma^2 := \text{Var}S = \sum_{j=1}^n \sigma_j^2$ . Note:

1. If  $\forall j, X_j \sim \mathcal{N}(0, \sigma_j^2)$ , then  $S \sim \mathcal{N}(0, \sigma^2)$  by Lemma 10.5.
2. Given independent random variables  $X_1, X_2, \dots, X_n$  with arbitrary distributions, we can always construct a new sequence  $Z_1, Z_2, \dots, Z_n$  of *normal* random variables with matching means and variances so that all of  $Z_i$  and  $X_i$  are mutually independent. This may involve changing the basic probability space, but that does not matter because the distribution of  $S$  is determined by the joint distribution of  $(X_1, X_2, \dots, X_n)$ , which remains the same.

Let

$$\begin{aligned} S &:= S_0 := X_1 + X_2 + X_3 + \dots + X_n, \\ S_1 &:= Z_1 + X_2 + X_3 + \dots + X_n, \\ S_2 &:= Z_1 + Z_2 + X_3 + \dots + X_n, \\ &\vdots \qquad \qquad \qquad \vdots \qquad \qquad \qquad \vdots \\ T := S_n &:= Z_1 + Z_2 + Z_3 + \dots + Z_n, \end{aligned}$$

We want to show that  $S$  is "close" in distribution to  $T$ , i.e., that  $\mathbb{E}f(S)$  is close to  $\mathbb{E}f(T)$  for all  $f \in \mathbf{C}_b^3(\mathbb{R})$  with uniform bound  $K$  on  $f$  and its first three derivatives:  $|f^{(i)}|, i = 1, 2, 3$ .

By the triangle inequality,

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \leq \sum_{j=1}^n |\mathbb{E}f(S_j) - \mathbb{E}f(S_{j-1})|. \quad (3.6)$$

Let  $R_j$  be the sum of the common terms in  $S_{j-1}$  and  $S_j$ . Then  $S_{j-1} = R_j + X_j$  and  $S_j = R_j + Z_j$ . Note that by construction,  $R_j$  and  $X_j$  are independent, as are  $R_j$  and  $Z_j$ .

We need to compare  $\mathbb{E}f(R_j + X_j)$  and  $\mathbb{E}f(R_j + Z_j)$ . By the Taylor series expansion up to the third term,

$$f(R_j + X_j) = f(R_j) + X_j f^{(1)}(R_j) + \frac{X_j^2}{2!} f^{(2)}(R_j) + \frac{X_j^3}{3!} f^{(3)}(\alpha_j),$$

$$f(R_j + Z_j) = f(R_j) + Z_j f^{(1)}(R_j) + \frac{Z_j^2}{2!} f^{(2)}(R_j) + \frac{Z_j^3}{3!} f^{(3)}(\beta_j),$$

where  $\alpha_j$  is a point between  $R_j$  and  $R_j + X_j$  and  $\beta_j$  is a point between  $R_j$  and  $R_j + Z_j$ .

So, assuming that the  $X$ 's have a finite third moments, and noting that the  $Z$ 's do as well (see below), we can take expectations in each of these identities and subtract the resulting equations. Using independence and the fact that  $X$  and  $Z$  agree on their first and second moments, we see that everything below the third order cancels. Therefore,

$$|\mathbb{E}f(S_j) - \mathbb{E}f(S_{j-1})| = |\mathbb{E}f(R_j + X_j) - \mathbb{E}f(R_j + Z_j)| \quad (3.7)$$

$$= \left| \mathbb{E} \frac{X_j^3}{3!} f^{(3)}(\alpha_j) - \mathbb{E} \frac{Z_j^3}{3!} f^{(3)}(\beta_j) \right| \quad (3.8)$$

$$\leq \frac{K}{6} (\mathbb{E}|X_j|^3 + \mathbb{E}|Z_j|^3). \quad (3.9)$$

Let  $c$  be the third moment of a standard normal random variable. This is finite since,

$$c = 2 \int_0^\infty x^3 \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx = 2 \cdot \frac{2}{\sqrt{2\pi}} < \infty$$

Therefore,  $\mathbb{E}|Z_j|^3 = c\sigma_j^3$ .

Jensen's inequality implies that  $\|X\|_2 = (\mathbb{E}|X|^2)^{\frac{1}{2}} \leq (\mathbb{E}|X|^3)^{\frac{1}{3}} = \|X\|_3$ , so  $\sigma_j^3 \leq \mathbb{E}|X_j|^3$ , and therefore  $\mathbb{E}|Z_j|^3 = c\sigma_j^3 \leq c\mathbb{E}|X_j|^3$ , for each  $j$ .

Applying this to (3.9), we get

$$\frac{K}{6} (\mathbb{E}|X_j|^3 + \mathbb{E}|Z_j|^3) \leq \frac{K(1+c)}{6} \mathbb{E}|X_j|^3.$$

Now, from (3.6), we get

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \leq \frac{K(c+1)}{6} \sum_{j=1}^n \mathbb{E}|X_j|^3, \quad (3.10)$$


So far we have only considered one row of the array, but (3.10) is in fact true for every row with  $K$  and  $c$  unchanged and  $T$  having the same distribution. For each  $i$  we have,

$$|\mathbb{E}f(S_i) - \mathbb{E}f(T)| \leq \frac{K(c+1)}{6} \sum_{j=1}^{n_i} \mathbb{E}|X_{ij}|^3, \quad (3.11)$$

Now, assuming Lyapounov's condition holds for  $\delta = 1$ , the RHS of (3.11) goes to zero as  $i \rightarrow \infty$ .

By Lemma 10.6,  $S_i \xrightarrow{d} \mathcal{N}(0, 1)$  as  $i \rightarrow \infty$ . □

### 3.3.7 Proof of Lindeberg's Central Limit Theorem

 For Lyapounov's version of the CLT, we looked at a triangular array  $\{X_{ij}\}$  with  $\mathbb{E}X_{ij} = 0$ ,  $\mathbb{E}X_{ij}^2 = \sigma_{ij}^2$ ,  $\sum_{j=1}^{n_i} \sigma_{ij}^2 = 1$ . Taking  $S_i = X_{i1} + X_{i2} + \dots + X_{in_i}$ , we

saw that we could prove  $S_i \xrightarrow{d} \mathcal{N}(0, 1)$  assuming that  $\lim_{i \rightarrow \infty} \sum_{k=1}^{n_i} \mathbb{E}|X_{ij}|^3 = 0$ .

This is a condition on third moments - we would like to see if a weaker condition will suffice. We used third moments in a Taylor series expansion as follows:

$$f(R + X) = f(R) + Xf^{(1)}(R) + \frac{X^2}{2!}f^{(2)}(R) + \frac{X^3}{3!}f^{(3)}(\alpha), \quad (3.12)$$

where  $\alpha$  is a point between  $R$  and  $R + X$ .

Roughly, without the third moments assumption, the above expression is bad when  $X$  is large - although the first two moments exist, we might have  $\mathbb{E}|X|^3 = \infty$ . The idea now is to use the form in equation (3.12) when  $X$  is small and to make use of

$$f(R + X) = f(R) + Xf^{(1)}(R) + \frac{X^2}{2!}f^{(2)}(\gamma) \quad (3.13)$$

where  $\gamma$  is a point between  $R$  and  $R + X$ , when  $X$  is large.

Equating these expansions (3.12) and (3.13) for  $f(R + X)$ , we get an alternative form for the remainder in (3.12):

$$\frac{X^3}{6}f^{(3)}(\alpha) = \frac{X^2}{2}f^{(2)}(\gamma) - \frac{X^2}{2}f^{(2)}(R) \quad (3.14)$$

$$= \frac{X^2}{2}[f^{(2)}(\gamma) - f^{(2)}(R)]\mathbb{1}(|X| > \epsilon) \quad (3.15)$$

$$+ \frac{X^3}{6}f^{(3)}(\alpha)\mathbb{1}(|X| \leq \epsilon) \quad (3.16)$$

for  $\epsilon > 0$ . Thus, for  $f$  with  $|f^{(i)}| \leq K$  for  $i = 2, 3$ , we get

$$\left| \frac{X^3}{6}f^{(3)}(\alpha) \right| \leq KX^2\mathbb{1}(|X| > \epsilon) + \frac{K}{6}|X|^3\mathbb{1}(|X| \leq \epsilon) \quad (3.17)$$

$$\leq KX^2\mathbb{1}(|X| > \epsilon) + \frac{K}{6}\epsilon X^2, \quad (3.18)$$

an alternative to the upper bound  $\frac{K}{6}|X|^3$ , which we used in (3.9).

Now we return to the setup of section 10.5 and use our new result to get more refined bounds. From (3.6) and (3.8), we had

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \leq \sum_{j=1}^{n_j} \left| \mathbb{E} \frac{X_j^3}{6} f^{(3)}(\alpha_j) - \mathbb{E} \frac{Z_j^3}{6} f^{(3)}(\beta_j) \right|$$

Using (3.6), the new bound for  $X_j^3$  (3.18), the assumption that  $|f^{(3)}| < K$ , and since  $\mathbb{E}|Z_j|^3 = c\sigma_j^3$ , we get

$$|\mathbb{E}f(S) - \mathbb{E}f(T)| \leq \sum_{j=1}^n \left[ K\mathbb{E}X_j^2\mathbb{1}(|X_j| > \epsilon) + \frac{K}{6}\epsilon\mathbb{E}X_j^2 \right] + \sum_{j=1}^n \frac{K}{6}c\sigma_j^3 \quad (3.19)$$

$$= K \sum_{j=1}^n \mathbb{E}X_j^2\mathbb{1}(|X_j| > \epsilon) + \frac{K}{6}\epsilon\sigma^2 + \frac{cK}{6} \sum_{j=1}^n \sigma_j^3 \quad (3.20)$$

As  $i \rightarrow \infty$  (going down the rows of the triangular array), the first term goes to zero by the Lindeberg condition. The last term goes to zero since

$$\sum_{j=1}^{n(i)} \sigma_{ij}^3 \leq \left( \max_{1 \leq j \leq n(i)} \sigma_{ij} \right) \sum_{j=1}^{n(i)} \sigma_{ij}^2 = \sigma^2 \max_{1 \leq j \leq n(i)} \sigma_{ij},$$

which tends to zero by (??). Only  $\frac{K}{6} \epsilon \sigma^2$  remains, and letting  $\epsilon \rightarrow 0$  finishes the argument.



**Figure 3.1:** The first application of Markov chains was to the linguistic study of Pushkin's poem "Eugene Onegin". In the picture (by Ilya Repin), Eugene Onegin kills his friend Vladimir Lensky



# Chapter 4

## Markov Chains

### 4.1 Basic Definitions

#### 4.1.1 Transition Matrix

Let  $S$  be a countable set, and  $X_n, n \geq 0$ , be a sequence of random variables that take values in the state space  $S$ . (We will often identify  $S$  with a subset of integers.) We say that  $X_n$  is a *discrete-time Markov chain* with the initial probability distribution  $\lambda$  on  $S$ , and transition matrix  $P$  if

1.  $\mathbb{P}(X_0 = i) = \lambda_i$ ;
2.  $\mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, \dots, X_n = i_n) = P_{i_n, i_{n+1}}$ .

This definition allows us calculate the joint distributions. For every sequence of states,  $(i_0, \dots, i_n)$

$$\mathbb{P}(X_0 = i_0, \dots, X_n = i_n) = \lambda_{i_0} P_{i_0, i_1} P_{i_1, i_2} \dots P_{i_{n-1}, i_n}.$$

In particular if we sum over all  $i_0, \dots, i_{n-1}$ , we will find the marginal distribution of  $X_n$ ,

$$\mathbb{P}(X_n = j) = (\lambda P^n)_j.$$

Here  $P^n$  is the  $n$ -th power of the matrix  $P$ ,  $\lambda P^n$  denote the product of vector  $\lambda$  by matrix  $P^n$ , and  $(\lambda P^n)_j$  is the  $j$ -th component of this product.

It follows that

$$\mathbb{P}(X_n = j | X_0 = i) = (P^n)_{ij}.$$

We will often write the conditional probabilities  $\mathbb{P}(A | X_0 = i)$  as  $\mathbb{P}_i(A)$ , so, for example, the previous result is  $\mathbb{P}_i(X_n = j) = (P^n)_{ij}$ .

*Example 4.1.1* (Random walk on a graph).

Recall that a *graph*  $G = (V, E)$  is a set of vertices  $V$  and a set of edges  $E$ , which are simply a pair of vertices  $E \subset V \times V$ . We will usually assume that the graph is simple, that is, that there are no multiple edges (edges with the same endpoints) and that there are no loops, edges that have the same vertex as both endpoints. The edges  $(v_1, v_2)$  and  $(v_2, v_1)$  are not distinguished, so the graph is undirected. A degree of a vertex  $v$ , denoted  $d(v)$ , is the number of edges which are incident to  $v$ , that is, that have  $v$  as one of its endpoints.

Now we define a Markov chain which is called a *simple random walk* on  $G$ . The states are vertices and the transition probability  $P_{uv} = 1/d(u)$ . The interpretation is that if there is a particle at vertex  $u$ , it has equal probabilities move along each of the edges incident to  $u$ .

### Exercises

*Ex.* 4.1.2. Let  $X_0$  be a random variable with values in a countable set  $I$ . Let  $Y_1, Y_2, \dots$  be a sequence of independent random variables, uniformly distributed on  $[0, 1]$ . Suppose that we are given a function  $G: I \times [0, 1] \rightarrow I$  and define inductively  $X_{n+1} = G(X_n, Y_{n+1})$ .

Show that  $(X_n)_{n \geq 0}$  is a Markov chain and express its transition matrix  $P$  in terms of  $G$ . Can all Markov chains be realized in this way? How would you simulate a Markov chain using a computer?

*Ex.* 4.1.3. Suppose that  $Z_0, Z_1, \dots$  are i. i. d. random variables such that  $Z_i = 1$  with probability  $p$  and  $Z_i = 0$  with probability  $1 - p$ . Set  $S_0 = 0$ ,  $S_n = Z_1 + \dots + Z_n$ . In each of the following cases determine whether  $(X_n)_{n \geq 0}$  is a Markov chain.

$$\begin{aligned} X_n &= S_n, \\ X_n &= S_0 + \dots + S_n, \\ X_n &= (S_n, S_0 + \dots + S_n). \end{aligned}$$

In the cases where  $X_n$  is a Markov chain, find its state-space and transition matrix, and in the cases where it is not a Markov chain give an example where  $P(X_{n+1} = i | X_n = j, X_{n-1} = k)$  is not independent of  $k$ .

*Ex.* 4.1.4. Flip coins ( $X_n = 0$  or  $1$ ), where each coin flip has parameter  $p = 3/4$  if the last 3 outcomes are 1's, and  $p = 1/2$  otherwise. For example,  $\mathbb{P}(X_4 = 1 | X_1 = 0, X_2 = 1, X_3 = 1) = 1/2$ .

1. Is  $\{X_n\}$  a Markov chain?
2. Let  $Z_n = (X_n, X_{n+1}, X_{n+2})$ . Argue that  $\{Z_n\}$  is a Markov chain and give the transition matrix.

*Ex.* 4.1.5. Let  $X_n$  be a Markov chain with transition matrix  $P$ .

1. Find

$$\mathbb{P}(X_n = j | X_{n-1} = i, X_{n+1} = j).$$

2. Calculate this probability when  $X_n$  is a simple random walk (up one with probability  $p$  and down one with probability  $q$ ).

### 4.1.2 Communicating classes and irreducible Markov chains

A Markov chain can be thought of as a non-deterministic automaton, with fixed probabilities of transition from one state to another. In this cases, it is often useful to re-present the chain by a directed graph. Each vertex in this graph represent a state, and state  $i$  is connected by a directed edge to  $j$  if there is a non-zero probability of transition from  $i$  to  $j$ . One can think about the probability as a label on this graph.

In this way we obtain an identification of Markov chains with weighted directed graphs, and an evolution of a Markov chain can be represented by a random walk on the graph.

Sometimes, it is possible to divide the state space of the Markov chain in pieces, so that the behavior of Markov chain is easier to analyze by considering each piece individually. These pieces are called *communicating classes*. (In graph theory, they are called *strongly connected components*, and can be calculated by using *Kosaraju's algorithm*.)

Namely, we say that  $i$  leads to  $j$  and write  $i \rightarrow j$ , if

$$\mathbb{P}_i(X_n = j \text{ for some } n \geq 0) > 0.$$

We say that  $i$  communicates with  $j$  and write  $i \leftrightarrow j$ , if both  $i \rightarrow j$  and  $j \rightarrow i$ . It is easy to see that this is an equivalence relation and therefore it partitions the state space into equivalence classes, which we call *communicating classes*.

**Definition 4.1.6.** A Markov chain is called *irreducible* if it has only one communicating class.

We say that a class is *closed* if the conditions  $i \in C$  and  $i \rightarrow j$  imply that  $j \in C$ . That is, the chain cannot go from a closed class outside. A single state  $i$  is called *absorbing* if  $\{i\}$  is a closed class.

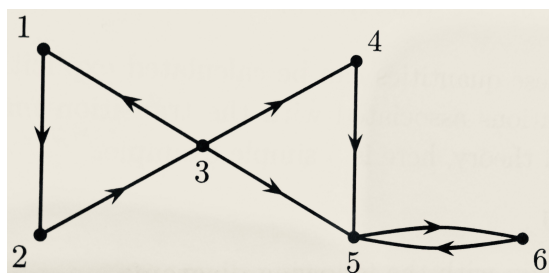


Figure 4.1

*Example 4.1.7.* Consider the chain in Figure 4.1. The arrows represent transitions with non-zero probabilities. For this chain, the classes are  $\{1, 2, 3\}$ ,  $\{4\}$ , and  $\{5, 6\}$ , and the only closed class is  $\{5, 6\}$ .

*Ex.* 4.1.8. Show that every transition matrix on a finite state-space has at least one closed communicating class. Find an example of a transition matrix with no closed communicating class.

### 4.1.3 Invariant distribution

A measure (or distribution)  $\lambda$  on a countable state space is a (non-zero) vector with non-negative entries. We call it a probability distribution if the sum of the entries is 1.

If  $P$  is the transition matrix of a Markov chain then a measure  $\pi$  is called *invariant* if

$$\pi P = \pi$$

The terms *equilibrium* or *stationary* measure are also used to mean the same.

The definition of the invariant distribution implies that if  $X_n$  is distributed according to  $\pi$  then  $X_{n+1}$  will also be distributed according to  $\pi$ .

Note that  $\sum_j P_{ij} = 1$  for all  $i$ , which means that matrix  $P$  has a *right* eigenvector with eigenvalue 1 that has all its entries equal to 1. From an algebraic viewpoint an invariant measure is a *left* eigenvector of the matrix  $P$  with eigenvalue 1. This gives us a practical method for computation of the invariant distribution if the state space is finite (and not too large).

In addition, from the properties of the eigenvectors, we can immediately conclude that for every finite Markov chain, the invariant distribution always exists, although it might be non-unique. It is also clear that we can normalize it so that the invariant distribution is actually a probability distribution.

For infinite chains the invariant distribution not necessarily exists even if do not require it to be a probability distribution. Here is an example.

Here is an example of a chain without an invariant measure.

*Example* 4.1.9. Consider the chain on  $\{0, 1, 2, \dots\}$  with the transition probabilities

$$p_{i,i+1} = p_i, \quad p_{i,0} = 1 - p_i =: q_i,$$

where  $p_i \in (0, 1)$ . We are going to show that for a suitable choice of  $q_i$ , the chain do not have a non-trivial invariant measure.

Let  $\pi$  be an invariant measure. Then, by definition,

$$\begin{aligned} \pi_0 &= \sum_{i=0}^{\infty} q_i \pi_i, \\ \pi_i &= p_{i-1} \pi_{i-1}, \text{ for } i \geq 1. \end{aligned}$$

The last equality implies that  $\pi_1 = p_0 \pi_0$ ,  $\pi_2 = p_1 p_0 \pi_0$ , and so on. Assume  $\pi_0 > 0$ , since otherwise the measure is trivial. Then,  $\pi_i < \pi_0$  for all  $i \geq 1$ . Hence, from the first equation, we get

$$\pi_0 < \pi_0 \sum_{i=0}^{\infty} q_i,$$

Take  $q_i = 2^{-i-1}$ , then  $\sum_{i=0}^{\infty} q_i = 1$  and  $\pi_0 < \pi_0$ , which is a contradiction.

We will consider the question of the invariant measure existence for infinite chains in a later section.

#### 4.1.4 Time reversal

It is an interesting question of whether one can infer a direction of time from the evolution of a Markov chain.

First of all, is a sequence  $X_n, X_{n-1}, \dots, X_0$ , a Markov chain?

We calculate

$$\mathbb{P}(X_k = i_k | X_{k+1} = i_{k+1}, \dots, X_n = i_n) = \frac{\mathbb{P}(X_k = i_k)}{\mathbb{P}(X_{k+1} = i_{k+1})} P_{i_k i_{k+1}}.$$

In general, the quantity on the left depends on  $k$  and is different for different choices of the initial distribution of  $X_0$  in the original chain. So, we have a Markov chain with changing transition probabilities.

However, if we assume that the original distribution was invariant, then  $X_n$  is distributed according to the invariant distribution  $\pi$ , and the reversed sequence of  $X_k$  is a Markov chain with constant transition probabilities, given by

$$\hat{P}_{ji} = \frac{\pi_i}{\pi_j} P_{ij}.$$

**Theorem 4.1.10.** *Let  $X_n$  be an irreducible Markov chain with transition matrix  $P$  and let  $X_0$  distributed according to the invariant distribution  $\pi$ . Let  $N > 0$  and set  $Y_n = X_{N-n}$  for  $n = 0, 1, \dots, N$ . Then  $Y_n$  is a Markov chain with the initial distribution  $\pi$  and the matrix  $\hat{P}$ , determined by equations  $\pi_j \hat{P}_{ji} = \pi_i P_{ij}$ .*

*Proof.* We have

$$\begin{aligned} \mathbb{P}(Y_0 = a, \dots, Y_N = z) &= \mathbb{P}(X_0 = z, \dots, X_N = a) \\ &= \pi_z P_{zy} P_{yx} \dots P_{ba} \\ &= \hat{P}_{yz} \pi_y P_{yx} \dots P_{ba} \\ &= \dots \\ &= \pi_a \hat{P}_{ab} \dots \hat{P}_{xy} \hat{P}_{yz}, \end{aligned}$$

where we applied the definition of  $\hat{P}$  several times. This equality implies that  $Y_i$  is a Markov chain with initial distribution  $\pi$  and the transition matrix  $\hat{P}$ .  $\square$

In particular, a Markov chain cannot be distinguished from its time-reversal if

$$\pi_j P_{ji} = \pi_i P_{ij}$$

These equations are called the *detailed balance equations*. If a chain satisfies these equations then it is called *reversible*. It turns out that reversible Markov chains are easier to understand than non-reversible chains.

In fact, it is redundant to require that the distribution in the detailed balance equations is invariant.

**Lemma 4.1.11.** *If distribution  $\lambda$  satisfy  $\lambda_j P_{ji} = \lambda_i P_{ij}$ , then  $\lambda$  is invariant.*

*Proof.*

$$(\lambda P)_i = \sum_j \lambda_j P_{ji} = \sum_j \lambda_i P_{ij} = \lambda_i.$$

□

This gives us a convenient tool for finding the invariant distribution of a chain.

*Example 4.1.12* (Random walk on a graph).

Consider a graph  $G$  with vertices  $i \in V$ . A *degree* (or *valence*) of a vertex  $i$  is the number of edges incident with  $i$ . A random walk on the graph  $G$  has the transition matrix  $P$  with entries  $P_{ij} = 1/d_i$  if  $(i, j)$  is an edge, and  $P_{ij} = 0$  otherwise. Here  $d_i$  denotes the degree of the vertex  $i$ . It is easy to check that  $P$  satisfies the detailed balance condition with  $\lambda_i = d_i$ . It follows that the random walk is reversible with the invariant measure  $\pi = d_i$ .

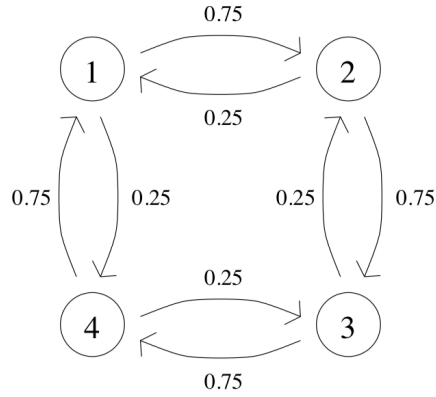
If the graph  $G$  is not regular, that is, if it has vertices of differing degrees, then this invariant measure is not uniform. Vertices with larger degree will be visited more often than vertices with smaller degree. What if we want to have at our disposal a Markov chain on the graph  $G$  that would have the same transitions, – from a vertex to their neighbors, – but that would have a uniform distribution on vertices?

In this case, we can use a *lazy random walk*. Namely, suppose  $d = \max\{d_1, \dots, d_{|V|}\}$  is the maximum vertex degree in the graph. Then we set  $P_{ij} = 1/d$  if  $j \neq i$ , and  $P_{ii} = 1 - d_i/d$ . In other words, if  $d_i < d$  then with positive probability the particle will stay at vertex  $i$  and wait for the next time period. It is easy to see from the detailed balance equation that the uniform distribution is invariant for this chain.

In Figure 4.2, a non-reversible chain is presented. It is clear from symmetry that the invariant distribution is uniform, but then the detailed balance equation is not satisfied:  $P_{ji} \neq P_{ij}$ .

### Exercises

*Ex. 4.1.13.* In each of the following cases determine whether the stochastic matrix  $P$  is reversible:



**Figure 4.2:** An example of a non-reversible chain: a random walk with a bias.

1.

$$\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix};$$

2.

$$\begin{bmatrix} 0 & p & 1-p \\ 1-p & 0 & p \\ p & 1-p & 0 \end{bmatrix};$$

3. The state space is  $\{0, 1, \dots, N\}$  and  $p_{ij} = 0$  if  $|j - i| \geq 2$ .

4. The state space is  $I = \{0, 1, 2, \dots\}$ ,  $p_{01} = 1$ , and  $p_{i,i+1} = p$ ,  $p_{i,i-1} = 1 - p$  for  $i \geq 1$ .

5.  $p_{ij} = p_{ji}$  for all  $i, j \in S$ .

*Ex.* 4.1.14. A Markov chain with the state space  $\{0, 1, 2\}$  has the transition probability matrix

$$P = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.6 & 0.2 & 0.2 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}.$$

After a long period of time, you observe the chain and see that it is in state 1. What is the conditional probability that the previous state was state 2? That is, find

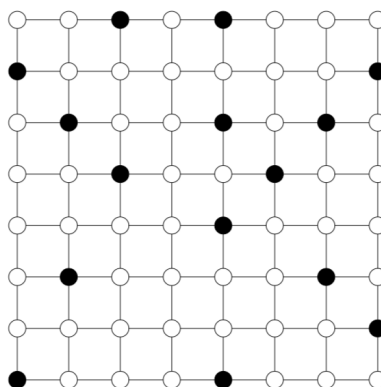
$$\lim_{n \rightarrow \infty} \mathbb{P}(X_{n-1} = 2 | X_n = 1).$$

### 4.1.5 Markov Chains for Sampling

An important application of Markov chains occurs in the situation when we want to sample from a specific distribution on a complicated state space. This sample will serve as a mean to understand the state space better. In this case, we often want to construct a Markov chain that will explore the state space and that will have this target distribution as its equilibrium distribution.

There is a family of algorithms invented for this specific task which are known under the general name of the Markov Chain Monte Carlo (MCMC) algorithm.

In fact there are many such algorithms for various problems and here we will only look at a single simple example.



**Figure 4.3:** An example of a feasible state on a the 8-by-8 lattice graph. Black and white circles represent 1's and 0's, respectively.

*Example 4.1.15* (The hard-core model).

Let  $G$  be a graph with vertex set  $V$  and edge set  $E$ . A *configuration* is an assignment of 0's and 1's to the vertices, and an assignment is *feasible* if no two 1's occupy adjacent vertices. We prescribe the same probability to each feasible configuration. (An example of a feasible configuration on a lattice graph is shown in Figure 4.3).

A possible question for such a model could be about the expected number of 1's in a random feasible configuration.

It is not clear how to answer this question theoretically, and in order to answer this question by using simulations we want to be able to sample from the uniform distribution on the space of all feasible configurations. For this purpose, we can use a suitably designed Markov chain  $X_n$  on the state space of configurations. Let  $X_n(v)$  denote the assignment that the configuration  $X_n$  prescribes to the vertex  $v$ . Then the chain has the following transition mechanism.



1. Pick a vertex  $v$  at random (uniformly). If any of the neighbors of  $v$  takes value 1 in  $X_n$ , repeat a random choice of  $v$ .
2. If all neighbors of  $v$  take value 0 in  $X_n$ , then toss a fair coin. If the coin comes up heads, then set  $X_{n+1}(v) = 1$ , otherwise let  $X_{n+1}(v) = 0$ .
3. For all  $w \neq v$ , set  $X_{n+1}(w) = X_n(w)$ .

We claim that its invariant distribution is uniform on the set of all feasible configurations. We prove it by showing that the chain is reversible with respect to the uniform distribution. Essentially this means that the probability to get from configuration  $\omega_1$  to configuration  $\omega_2$  is the same as the probability to get from  $\omega_2$  to  $\omega_1$ .

Indeed, if  $\omega_2 \neq \omega_1$ , then the probability of transition from a feasible  $\omega_1$  to  $\omega_2$  is non-zero if and only if the configurations are different at exactly one vertex  $v$ . Probability to choose this vertex is the same in both configurations. This vertex must be surrounded by zeros or one of these configurations is not feasible, so it is not in state space. And if this vertex is chosen then the probability of transition from  $\omega_1$  to  $\omega_2$  equals  $1/2$  and it is the same as the probability to go from  $\omega_2$  to  $\omega_1$ .

By results that will follow, the distribution of the chain converges to invariant. So, if we run the chain for a sufficiently large time  $T$ , then we will eventually get a state  $X_n$  which is close to a sample point from a uniform distribution. Repeating this procedure  $N$  times we will get a sample of  $N$  random assignments which will be approximately i.i.d and we can use them to estimate the expected number of 1's.

This algorithm is an example of “a Gibbs sampler”, or “Glauber dynamics”. The characteristic features of this type of algorithms is that they operate on states which can be represented as assignments  $s(v)$  on a certain set of vertices  $v \in V$ . The updates (transitions) of an assignment are done locally, on a single vertex  $v$  of the set  $V$ . The key point is that an assignment on  $v$  is set to a particular value with the probability equal to the conditional probability of this value given the values on other vertices  $w$ 's, and the conditional probability is calculated assuming the joint distribution  $\pi$ .

It is the fact that these conditional probabilities have a simple form that makes this method useful.

In order to see that this chain is reversible with the invariant distribution  $\pi$ , note that only possible transitions are between states that differ in only one  $v$ . In this case, we have

$$\begin{aligned}
\pi([s, \widehat{s}])P_{[s, \widehat{s}], [s', \widehat{s}]} &= \pi([s, \widehat{s}])\mathbb{P}([s', \widehat{s}] | \widehat{s}) \\
&= \pi([s, \widehat{s}]) \frac{\pi([s', \widehat{s}])}{\sum_t \pi([t, \widehat{s}])} \\
&= \pi([s', \widehat{s}])\mathbb{P}([s, \widehat{s}] | \widehat{s}) \\
&= \pi([s', \widehat{s}])P_{[s', \widehat{s}], [s, \widehat{s}]},
\end{aligned}$$

where  $s$  and  $s'$  denote the assignments on  $v$  in two neighboring configurations, and  $\widehat{s}$  denotes the assignment on vertices  $w$  different from  $v$ . Hence, the detailed balance condition holds and the chain is reversible.

The calculation is almost tautological, and it becomes useful only when the conditional probabilities are easy to calculate.

## 4.2 Hitting times and absorption probabilities



The *hitting time* of a subset  $A$  is the random variable

$$H^A = \inf\{n \geq 0 : X_n \in A\}.$$

The conditional expectation of  $H^A$  given that the chain starts at  $i$  is denoted by  $k_i^A$ , and the conditional probability that  $X_n$  ever hits  $A$  is called the *hitting probability* of  $A$  and denoted

$$h_i^A = \mathbb{P}_i(H^A < \infty).$$

**Theorem 4.2.1.** *The vector of hitting probabilities  $h^A$  is a minimal non-negative solution to the system of linear equations*

$$h_i^A = 1, \quad \text{for } i \in A, \quad (4.1)$$

$$h_i^A = \sum_j P_{ij} h_j^A, \quad \text{for } i \notin A. \quad (4.2)$$

Minimality means that for any other non-negative solution  $\overline{h^A}$ , it must be that  $\overline{h_i^A} > h_i^A$  for every  $i$ .

In words, the function  $h_i$  is  $P$ -harmonic at the states  $i$  which are outside of the set  $A$ , that is, the value of  $h$  at state  $i$  is a weighted average of values of  $h$  at the neighbors of  $i$ . The values of  $h$  at states  $i$  which are in  $A$  give a boundary condition  $h = 1$  for this harmonic function.

*Proof.* First, we show that the vector of hitting times must satisfy the equations (4.1) and (4.2). Obviously, it satisfies the first equation. For the second, we write

$$h_i^A = \sum_j \mathbb{P}_i(H^A < \infty | X_1 = j) P_{ij}$$

Since  $i \notin A$  the event  $H^A < \infty$  depend only on  $X_1, X_2, \dots$  and therefore by Markov property we have that  $\mathbb{P}_i(H^A < \infty | X_1 = j) = \mathbb{P}_j(H^A < \infty) = h_j^A$ . Therefore,

$$h_i^A = \sum_j P_{ij} h_j^A,$$

and the second equality is proved.

Next we establish that for every other solution  $x$ ,  $x_i \geq h_i^A$ . If  $i \in A$  then  $x_i = h_i^A = 1$ . If  $i \notin A$ , then

$$\begin{aligned} x_i &= \sum_j P_{ij} x_j = \sum_{j \in A} P_{ij} + \sum_{j \notin A} P_{ij} x_j \\ &= \mathbb{P}_i(X_1 \in A) + \sum_{j \notin A} P_{ij} x_j \\ &= \mathbb{P}_i(X_1 \in A) + \sum_{j \notin A} P_{ij} \left( \sum_{k \in A} P_{jk} + \sum_{k \notin A} P_{jk} x_k \right) \\ &= \mathbb{P}_i(X_1 \in A) + \mathbb{P}_i(X_1 \notin A, X_2 \in A) + \sum_{j \notin A} \sum_{k \notin A} P_{jk} x_k \end{aligned}$$

If we repeat this procedure for  $n$  steps, we find that

$$\begin{aligned} x_i &= \mathbb{P}_i(X_1 \in A) + \dots + \mathbb{P}_i(X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A) \\ &\quad + \sum_{j_1 \notin A} \dots \sum_{j_n \notin A} P_{j_1, j_2} \dots P_{j_{n-1}, j_n} x_{j_n} \\ &\geq \mathbb{P}_i(H^A \leq n). \end{aligned}$$

By taking the limit we find that  $x_i \geq \mathbb{P}_i(H^A < \infty) = h_i^A$ .  $\square$

**Theorem 4.2.2.** *The vector of mean hitting times  $k^A$  is the minimal non-negative solution to the system of linear equations*

$$k_i^A = 0, \quad \text{for } i \in A, \quad (4.3)$$

$$k_i^A = 1 + \sum_{j \notin A} P_{ij} k_j^A, \quad \text{for } i \notin A. \quad (4.4)$$

*Proof.* First, we show that  $k^A$  satisfies the equations (4.3) and (4.4). The first one is evident. If we assume that  $i \notin A$ , then

$$k_i^A = \mathbb{E}_i(H^A) = \sum_j \mathbb{E}_i(H_A | X_1 = j) \mathbb{P}_i(X_1 = j)$$

By the Markov property,  $\mathbb{E}_i(H_A | X_1 = j) = 1 + \mathbb{E}_j(H_A)$ , and therefore,

$$k_i^A = 1 + \sum_{j \notin A} P_{ij} k_j^A,$$

which shows that equation (4.4) also holds.

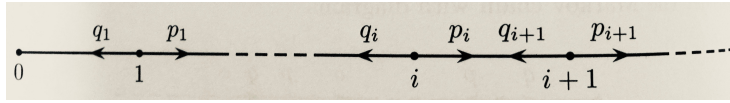
Suppose now that  $x$  is a solution of (4.3) and (4.4). Then, for  $i \in A$ ,  $x_i = k_i^A = 0$ . If  $i \notin A$ , then

$$\begin{aligned} x_i &= 1 + \sum_{j \notin A} p_{ij} x_j = 1 + \sum_{j \notin A} p_{ij} \left( 1 + \sum_{k \notin A} p_{jk} x_k \right) \\ &= \mathbb{P}_i(H^A \geq 1) + \mathbb{P}_i(H^A \geq 2) + \sum_{j \notin A} \sum_{k \notin A} p_{ij} p_{jk} x_k \end{aligned}$$

By repeating this procedure and taking the limit we find that

$$x_k \geq \sum_{n=1}^{\infty} \mathbb{P}_i(H^A \geq n) = \mathbb{E}_i(H^A) = k_i^A,$$

which shows that  $k_i^A$  is a minimal solution and completes the proof.  $\square$



**Figure 4.4:** The birth-and-death chain

*Example 4.2.3* (Birth-and-death chain). Consider the Markov chain in Figure 4.4. The state 0 is the absorbing state and we wish to calculate the absorption probability  $h_i = \mathbb{P}_i(\text{"hit 0"})$ .

We have  $h_0 = 1$  and  $h_i = p_i h_{i+1} + q_i h_{i-1}$  for  $i \geq 1$ . The latter equation implies that  $p_i(h_i - h_{i+1}) = q_i(h_{i-1} - h_i)$ , so if we introduce  $u_i = h_{i-1} - h_i$ , then we can write

$$u_{i+1} = \frac{q_i}{p_i} u_i = \frac{q_i q_{i-1} \cdots q_1}{p_i p_{i-1} \cdots p_1} u_1 = \gamma_i u_1,$$

where the last equality defines  $\gamma_i$ .

Since  $u_1 + \cdots + u_i = h_0 - h_i$ , so

$$h_i = 1 - u_1(\gamma_0 + \cdots + \gamma_{i-1}),$$

where  $\gamma_0 = 1$ .

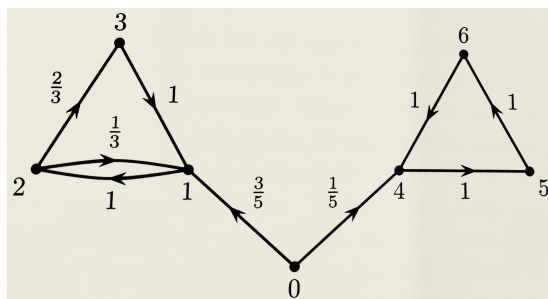
If  $\sum_{i=0}^{\infty} \gamma_i = \infty$ , then the restriction  $h_i \in [0, 1]$  forces  $u_1 = 0$  and  $h_i = 1$  for all  $i$ . (Extinction is inevitable.)

However, in the case when  $\sum_{i=0}^{\infty} \gamma_i = A < \infty$ , we can take a  $u_1$  in the region  $[0, A^{-1}]$ , and then the solution is

$$h_i = \frac{\sum_{j=i}^{\infty} \gamma_j}{\sum_{j=0}^{\infty} \gamma_j} < 1,$$

which shows that for every  $i$  there is a positive probability of survival.

### Exercises



**Figure 4.5:** An example of a discrete-time Markov Chain

*Ex. 4.2.4.* Consider the Markov chain shown in Figure 4.5. For this chain, show that

1. The probability of hitting 6, starting from 0, is  $1/4$ .
2. The probability of hitting 3, starting from 1, is 1.
3. It takes on average 3 steps to hit 3 starting from 1.

*Ex. 4.2.5.* Consider the Markov chain on states  $\{0, 1, 2, 3, 4\}$  whose transition matrix is

$$P = \begin{bmatrix} q & p & 0 & 0 & 0 \\ q & 0 & p & 0 & 0 \\ q & 0 & 0 & p & 0 \\ q & 0 & 0 & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

where  $p + q = 1$ . Determine the mean time to reach state 4 starting from state 0.

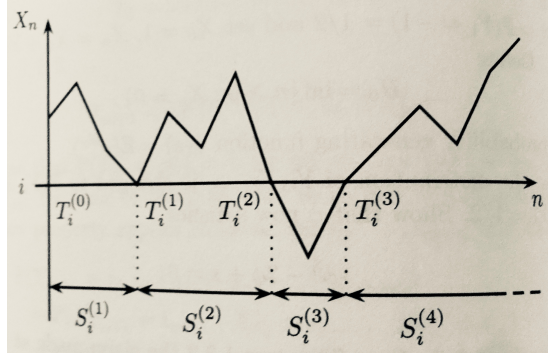
### 4.3 Recurrence and transience

Consider the event “ $X_n = i$  for infinitely many  $n$ ”. By a very general result due to Kolmogorov, this event can only have the probability 0 or 1. If the probability is 0 then we call state  $i$  *transient*. If it is 1, then we call it *recurrent*.

In order to define a useful criterion for whether a state is transient or recurrent, we define passage times  $T_i^{(r)}$ . They are defined inductively. Let  $T_i^{(0)} = 0$  and define the *first passage time* as

$$T_i = T_i^{(1)} = \inf n \geq 1: X_n = i,$$

where the infimum of the empty set is  $\infty$  by convention. This random variable is also called the *return probability*.



**Figure 4.6:** Passage Times

In general, the passage time number  $r + 1$  is defined as

$$T_i^{(r+1)} = \inf n \geq T_i^{(r)} + 1 : X_n = i.$$

The length of the  $r$ -th excursion is defined as  $S_i^{(r)} = T_i^{(r)} - T_i^{(r-1)}$  if  $T_i^{(r-1)} < \infty$ , and 0, otherwise. These definitions are illustrated in Figure 4.6

Define also the *number of visits* to  $i$  as

$$V_i = \sum_{n=0}^{\infty} \mathbb{1}_{X_n=i}.$$

The expectation of  $V_i$  is

$$\mathbb{E}_i V_i = \sum_{n=0}^{\infty} \mathbb{E}_i \mathbb{1}_{X_n=i} = \sum_{n=0}^{\infty} \mathbb{P}_i \{X_n = i\} = \sum_{n=0}^{\infty} P_{ii}^n$$

**Theorem 4.3.1.** *There can be only two situations:*

1. *State  $i$  is recurrent,  $\mathbb{P}\{T_i < \infty\} = 1$  and  $\sum_n P_{ii}^n = \infty$ ;*
2. *State  $i$  is transient,  $\mathbb{P}\{T_i < \infty\} < 1$  and  $\sum_n P_{ii}^n < \infty$ ;*

The proof of this theorem is based on two lemmas.

**Lemma 4.3.2.**  $S_i^{(r)}$  is independent of  $\{X_m, m \leq T_i^{(r-1)}\}$  conditional on  $\{T_i^{(r-1)} \leq \infty\}$ , and

$$\mathbb{P}[S_i^{(r)} = n | \{T_i^{(r-1)} \leq \infty\}] = \mathbb{P}_i(T_i = n).$$

This lemma is a consequence of the strong Markov property of Markov chains. Essentially, if  $\{T_i^{(r-1)} \leq \infty\}$ , then at time  $t = T_i^{(r-1)}$ , the process  $X$  is at  $X_t = i$ , and its future behavior is the same as that of the original process started from  $X_0 = i$ . In particular, the time before the first new return to  $i$ ,  $S_i^{(r)}$  has the same distribution as that of the first return  $T_i$ .

**Lemma 4.3.3.**

$$\mathbb{P}_i(V_i > r) = \left[ \mathbb{P}_i(T_i < \infty) \right]^r.$$

*Proof.* Note that the event  $\{V_i > r\}$  is the same as  $\{T_i^{(r)} < \infty\}$ . For  $r = 0$ , the claim of the lemma is true. For positive  $r$ , we use induction:

$$\begin{aligned} \mathbb{P}_i(V_i > r + 1) &= \mathbb{P}_i(T_i^{(r)} < \infty \text{ and } S_i^{(r+1)} < \infty) \\ &= \mathbb{P}_i(S_i^{(r+1)} < \infty | T_i^{(r)} < \infty) \mathbb{P}_i(T_i^{(r)} < \infty) \\ &= \mathbb{P}_i(T_i < \infty) \mathbb{P}_i(T_i^{(r)} < \infty) = \left[ \mathbb{P}_i(T_i < \infty) \right]^{r+1}. \end{aligned}$$

□

*Proof of Theorem 4.3.1.* By Lemma 4.3.3, if  $\mathbb{P}\{T_i < \infty\} = 1$ , then

$$\mathbb{P}_i(V_i = \infty) = \lim_{r \rightarrow \infty} \mathbb{P}_i(V_i > r) = 1,$$

so  $i$  is recurrent and  $\sum P_{ii}^n = \mathbb{E}_i(V_i) = \infty$ .

On the other hand, if  $\mathbb{P}\{T_i < \infty\} < 1$ , then using Lemma 4.3.3 again,

$$\begin{aligned} \sum P_{ii}^n &= \mathbb{E}_i(V_i) = \sum_{r=0}^{\infty} \mathbb{P}_i(V_i > r) \\ &= \frac{1}{1 - \mathbb{P}\{T_i < \infty\}} < \infty, \end{aligned}$$

which implies that  $\mathbb{P}_i(V_i = \infty) = 0$  and, therefore,  $i$  is transient. □

**Corollary 4.3.4.** *The states of a communicating class are either all transient or all recurrent.*

This follows because for  $i$  and  $j$  in the same communicating class it is easy to show that  $\sum P_{ii}^n$  and  $\sum P_{jj}^n$  are both convergent or both divergent.

Hence we can talk about recurrent and transient classes. It is clear from the definition of the recurrent class that we cannot leave it forever. In other words, the following statement holds.

**Corollary 4.3.5.** *Every recurrent class is closed. In particular, if  $x$  is a recurrent state, and  $x$  leads to  $y$ , then  $y$  is in the same class as  $x$  and is recurrent.*

For finite classes we also have a converse statement.

**Corollary 4.3.6.** *Every finite closed class is recurrent.*

Indeed, if a class is closed and finite, then  $\mathbb{P}(\text{"}i \text{ is visited infinitely many times"} > 0$  for at least one  $i$ . This implies that the class is recurrent.

It can happen, however that an infinite closed class is transient.

We also have the following simple result.

**Theorem 4.3.7.** *Suppose that chain is irreducible and recurrent. Then, for all states  $j$ , we have*

$$\mathbb{P}[T_j < \infty] = 1.$$

The meaning of the theorem is that if we start with arbitrary state  $i$  we will reach any other state  $j$  in finite time with probability 1. Intuitively, the idea of the proof is that if it were that  $\mathbb{P}_i[T_j < \infty] < 1$  for some pair of states  $i$  and  $j$ , then there would be a possibility that the chain reaches  $i$  and then  $j$  will not be visited infinitely often. This would contradict the recurrency.

*Proof.* Since

$$\mathbb{P}(T_j < \infty) = \sum_i \mathbb{P}_i(T_j < \infty) \mathbb{P}(X_0 = i),$$

it is enough to prove that  $\mathbb{P}_i[T_j < \infty] = 1$  for all  $i$ .

Since the chain is recurrent, therefore

$$\mathbb{P}_j(X_n = j \text{ for infinitely many } n) = 1.$$

By irreducibility, we can choose  $m$  such that  $P_{ji}^m > 0$ , and then we write

$$\begin{aligned} \mathbb{P}_j(X_n = j \text{ for infinitely many } n) &= \mathbb{P}_j(X_n = j \text{ for some } n > m) \\ &= \sum_k \mathbb{P}_j(X_n = j \text{ for some } n > m | X_m = k) \mathbb{P}_j(X_m = k) \\ &= \sum_k \mathbb{P}_k(T_j < \infty) P_{jk}^m. \end{aligned}$$

Since  $\sum_k P_{jk}^m = 1$ , it must be that  $\mathbb{P}_k(T_j < \infty) = 1$  for all  $k$  such that  $P_{jk}^m > 0$ , in particular, for  $k = i$ .  $\square$

### Examples

*Example 4.3.8* (Simple symmetric random walk on  $\mathbb{Z}$ ). Simple symmetric walk on  $\mathbb{Z}$  has the transition probabilities  $P_{i,i-1} = P_{i,i+1}$ . Then, it is easy to see that

$$P_{ii}^{2n} = \binom{2n}{n} 2^{-2n}.$$

By using Stirling's formula for the factorial one can show that  $P_{ii}^{2n} = c/\sqrt{n}$ , and therefore the chain is recurrent by Theorem 4.3.1.

*Example 4.3.9* (Simple symmetric random walk on  $\mathbb{Z}^2$  and  $\mathbb{Z}^3$ ). By significantly more complicated combinatorics, one can show that  $P_{ii}^{2n} = c/n$  and  $P_{ii}^{2n} = c/n^{3/2}$  in the cases of symmetric random walk on  $\mathbb{Z}^2$  and  $\mathbb{Z}^3$ , respectively. Hence, by Theorem 4.3.1, the random walk is still recurrent in the case of  $\mathbb{Z}^2$ , but it is transient in the case of  $\mathbb{Z}^3$ .



*Example 4.3.10* (Birth and death chain). Consider a variant of the birth and death chain. Recall that the probability of transitions are  $p_{i,i+1} = p_i > 0$  for  $i \geq 0$  and  $p_{i,i-1} = q_i$  for  $i > 0$ . In addition, we assume that  $p_{00} = 1 - p_0$ . Here the state 0 is not absorbing and the chain is irreducible. When is it recurrent?

From the analysis in Example 4.2.3, we know that

$$h_i = \mathbb{P}\{\text{"hit 0 starting with state } i\} = 1$$

if and only if  $\sum \gamma_i = \infty$ , where

$$\gamma_i = \frac{q_i q_{i-1} \cdots q_1}{p_i p_{i-1} \cdots p_1}$$

We can write

$$\mathbb{P}_0(T_0 < \infty) = P_{00} + P_{01}h_1,$$

If  $\sum \gamma_i = \infty$ , then  $h_1 = 1$  and therefore  $\mathbb{P}_0(T_0 < \infty) = 1$ . So the state 0 is recurrent. Since the chain is irreducible, it is recurrent.

Conversely, if the chain is recurrent then,  $\mathbb{P}(T_0 < \infty) = 1$ , and therefore  $h_1 = 1$  and  $\sum \gamma_i = \infty$ .

Hence, the birth and death chain is recurrent if and only if  $\sum \gamma_i = \infty$ .

### Exercises

*Ex. 4.3.11.* Consider a Markov chain having state space  $\{0, 1, \dots, 6\}$  and transition matrix

$$\begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

1. Determine which states are transient and which states are recurrent.
2. Find the hitting probabilities  $h_0^i$  for  $i = 0, \dots, 6$ .

*Ex. 4.3.12.* Consider a Markov chain on non-negative integers such that, starting from  $x$ , the chain goes to state  $x + 1$  with probability  $p > 0$ , and goes to state 0 with probability  $q = 1 - p > 0$ .

1. Show that this chain is irreducible.
2. Find  $\mathbb{P}_0(T_0 = n)$ ,  $n \geq 1$ .
3. Show that this chain is recurrent.

*Ex.* 4.3.13. Let  $X_n$  be a Markov chain on  $\{0, 1, \dots\}$  with transition probabilities given by  $p_{01} = 1$ ,  $p_{i,i+1} + p_{i,i-1}$  for  $i \geq 1$ , and

$$p_{i,i+1} = \left(\frac{i+1}{i}\right)^2 p_{i,i-1},$$

for  $i \geq 1$ . Show that if  $X_0 = 0$ , then the probability that  $X_n \geq 1$  for all  $n \geq 1$  is  $6/\pi^2$ .

*Ex.* 4.3.14. For the Markov chain in the previous example, show that

$$\mathbb{P}(X_n \rightarrow \infty \text{ as } n \rightarrow \infty) = 1.$$

Suppose, instead, the transition probabilities satisfy

$$p_{i,i+1} = \left(\frac{i+1}{i}\right)^\alpha p_{i,i-1}.$$

For each  $\alpha \in (0, \infty)$  find the value of  $\mathbb{P}(X_n \rightarrow \infty \text{ as } n \rightarrow \infty)$ .

*Ex.* 4.3.15. A random sequence of non-negative integers  $F_n$  is obtained by setting  $F_0 = 0$  and  $F_1 = 1$  and, once  $F_0, \dots, F_n$  are known, taking  $F_{n+1}$  to be either the sum or the difference of  $F_{n-1}$  and  $F_n$ , each with probability  $1/2$ . Is  $F_n$  a Markov chain?

By considering the Markov chain  $X_n = (F_{n-1}, F_n)$ , find the probability that  $F_n$  reaches 3 before first returning to 0.

Draw enough of the flow diagram for  $X_n$  to establish a general pattern. Then, using the strong Markov property, show that the hitting probability for  $(1, 1)$  starting from  $(1, 2)$  is  $(3 - \sqrt{5})/2$ .

Deduce that  $X_n$  is transient. Show that, moreover, with probability 1,  $F_n \rightarrow \infty$  as  $n \rightarrow \infty$ .

## 4.4 More about invariant distributions

### 4.4.1 The existence of an invariant distribution



hen does an infinite Markov chain has an invariant distribution? First of all we can ask the question about the existence of an invariant measure, when we do not require that the measure of the whole space equals 1.

An essential feature of Example 4.1.9 is that probability mass escapes to infinity. If we exclude this possibility by requiring that the chain is recurrent, we find that the invariant measure always exists. (However, note that the invariant distribution can also exist for transient chains, so the requirement that the chain is recurrent is only a sufficient condition.)

In fact we can identify the invariant measure (up to a scale) as a expected time spend between visits to a particular state. The choice of this reference state is irrelevant.

Let  $k$  be a state and define  $\gamma_i^k$  as the expected time spent in  $i$  between visits to  $k$ ,

$$\gamma_i^k = \mathbb{E}_k \sum_{n=0}^{T_k-1} \mathbb{1}_{\{X_n=i\}}.$$

We are going to show that for an arbitrary choice of the initial state  $k$ , the vector  $\pi := \gamma^k$  is a positive invariant measure.

**Theorem 4.4.1.** *An irreducible and recurrent Markov chain always has an invariant measure  $\pi$  such that  $\pi_i > 0$  for all  $i$ .*

*Proof.* Since the chain is recurrent and  $T_k < \infty$  with probability 1, and since  $X_{T_k} = X_0 = k$ , we can also change limits in the sum and write

$$\gamma_i^k = \mathbb{E}_k \sum_{n=1}^{T_k} \mathbb{1}_{\{X_n=i\}}.$$

This expression can be written as

$$\begin{aligned} \gamma_i^k &= \sum_{n=1}^{\infty} \mathbb{E}_k \mathbb{1}_{\{X_n=i \text{ and } n \leq T_k\}} \\ &= \sum_{n=1}^{\infty} \mathbb{P}_k \{X_n = i \text{ and } n \leq T_k\} \\ &= \sum_j \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = j, X_n = i \text{ and } n \leq T_k). \end{aligned}$$

Then, we can write

$$\begin{aligned} \mathbb{P}_k(X_{n-1} = j, X_n = i \text{ and } n \leq T_k) &= \mathbb{P}(X_n = i | X_{n-1} = j \text{ and } n \leq T_k) \\ &\quad \times \mathbb{P}(X_{n-1} = j \text{ and } n \leq T_k) \end{aligned}$$

The event  $\{n \leq T_k\}$  means that  $X_1 \neq k, \dots, X_{n-1} \neq k$ , and therefore we can use the Markov property and write

$$\mathbb{P}_k(X_{n-1} = j, X_n = i \text{ and } n \leq T_k) = P_{ji} \mathbb{P}(X_{n-1} = j \text{ and } n \leq T_k)$$

Hence

$$\begin{aligned} \gamma_i^k &= \sum_j P_{ji} \sum_{n=1}^{\infty} \mathbb{P}_k(X_{n-1} = j, \text{ and } n-1 \leq T_k-1) \\ &= \sum_j P_{ji} \sum_{m=0}^{\infty} \mathbb{E}_k \mathbb{1}_{\{X_m = j, \text{ and } m \leq T_k-1\}} \\ &= \sum_j \gamma_j^k P_{ji} \end{aligned}$$

This shows that  $\gamma^k P = \gamma^k$ . Since  $\gamma_k^k = 1$ , the measure is non-zero. It remains to show that it is bounded for every  $i$  and that it is positive. By irreducibility for some  $n$  and  $m$ ,  $P_{ik}^n > 0$  and  $P_{ki}^m > 0$ . Then  $\gamma_i^k \geq \gamma_k^k P_{ki}^m > 0$ , and  $\gamma_i^k P_{ik}^n \leq \gamma_k^k = 1$ . This completes the proof.  $\square$

[Exercise: Why doesn't this proof work for the situation in Example 4.1.9?]

In fact, for irreducible and recurrent chains, the invariant measure is unique up to a multiplication.

**Theorem 4.4.2.** *If  $P$  is an irreducible and recurrent Markov chain and  $\pi$  is an invariant measure such that  $\pi_k = 1$ . Then  $\pi = \gamma^k$ .*

We omit the proof. Let us repeat that besides the existence and uniqueness, the two previous theorem show that the invariant measure of a state  $i$  is proportional to the expected time spend in the state  $i$  between two consecutive visits to some state  $k$ .

When can we say that the invariant measure of an irreducible recurrent chain is actually a probability distribution?

Recall that a state  $i$  is recurrent if  $\mathbb{P}_i(T_i < \infty) = 1$ , where  $T_i$  is the first return time. If in addition the expected return time is finite,  $m_i = \mathbb{E}_i(T_i) < \infty$ , then the state is called *positive recurrent*. A recurrent state which fails to have this stronger property is called *null recurrent*.

**Theorem 4.4.3.** *Let  $P$  be irreducible. Then the following are equivalent:*

1. every state is positive recurrent;
2. some state is positive recurrent;
3.  $P$  has an invariant probability distribution.

Moreover, if one of these conditions holds and  $\pi$  is the invariant probability distribution, then the expected first return time  $\mathbb{E}_i(T_i) = 1/\pi_i$ .

This theorem gives another interpretation for the invariant distribution for recurrent chains. The measure  $\pi_i$  is the inverse of the expected return time for the state  $i$ .

*Proof.* (i)  $\rightsquigarrow$  (ii), obviously.

(ii)  $\rightsquigarrow$  (iii). The sum  $\sum_j \gamma_j^i$  is the expected time before the first return to  $i$ ,  $m_i$ . (It is the expected time spend in all other states before the return to  $i$ .) By the definition of positive recurrence  $m_i$  is finite. Hence,  $\pi_j = \gamma_j^i/m_i$  defines an invariant distribution.

(iii)  $\rightsquigarrow$  (i). Take a state  $k$ . By irreducibility,  $\pi_k = \sum_i \pi_i P_{ik}^n > 0$  for some  $n$ . Set  $\lambda_i = \pi_i/\pi_k$ . Then,  $\lambda \geq \gamma^k$  by Lemma 4.4.4 below. Hence

$$m_k = \sum_i \gamma_i^k \leq \frac{\pi_i}{\pi_k} = \frac{1}{\pi_k} < \infty,$$

and  $k$  is positive recurrent.  $\square$

**Lemma 4.4.4.** *Let  $P$  be irreducible and let  $\lambda$  be an invariant measure for  $P$  with  $\lambda_k = 1$ . Then  $\lambda_j \geq \gamma_j^k$ .*

*Proof.* By invariance and the assumption that  $\lambda_k = 1$ , we can write

$$\begin{aligned}\lambda_j &= \sum_{a \neq k} \lambda_a P_{aj} + P_{kj} \\ &= \sum_{a, b \neq k} \lambda_b P_{ba} P_{aj} + \sum_{a \neq k} P_{ka} P_{aj} + P_{kj} \\ &= \dots\end{aligned}$$

We can continue this expansion and note that the first term is always positive. Than by reading the remaining terms from left to right we find that

$$\lambda_j \geq \mathbb{P}_k(X_1 = j \text{ and } T_k \geq 1) + \mathbb{P}_k(X_2 = j \text{ and } T_k > 2) + \dots$$

However, this is exactly the expected time spend by the chain  $X_n$  at  $j$  before the return to  $k$ . Hence,  $\lambda_j \geq \gamma_j^k$ .  $\square$

*Example 4.4.5* (Simple symmetric walk on  $\mathbb{Z}$ ). It is easy to check that for the symmetric walk on  $\mathbb{Z}$ , the only invariant measure with  $\pi_0 = 1$  is the constant measure  $\pi_i = 1$ . This measure is not a probability distribution and therefore we conclude that all states of this chain are null-recurrent,  $\mathbb{E}T_i = \infty$ .

*Example 4.4.6* (Birth and death chain). Consider the birth-and-death chain from Example 4.3.10. The invariant distribution  $\pi$  has to satisfy equations

$$\pi_{i-1} P_{i-1,i} + \pi_{i+1} P_{i+1,i} = \pi_i,$$

which we can write as

$$\pi_{i-1} p_{i-1} + \pi_{i+1} q_{i+1} = \pi_i,$$

or

$$\begin{aligned}q_{i+1} \pi_{i+1} - p_i \pi_i &= q_i \pi_i - p_{i-1} \pi_{i-1} \\ &= q_{i-1} \pi_{i-1} - p_{i-2} \pi_{i-2} \\ &= \dots \\ &= q_1 \pi_1 - p_0 \pi_0.\end{aligned}$$

However, another invariance equation gives us

$$\pi_1 P_{10} + \pi_0 P_{00} = \pi_0,$$

which simplifies to  $\pi_1 q_1 = \pi_0 p_0$ .

It follows that

$$\begin{aligned}q_{i+1} \pi_{i+1} - p_i \pi_i &= 0, \\ \pi_{i+1} &= \frac{p_i}{q_{i+1}} \pi_i,\end{aligned}$$

for all  $i \geq 0$ , and by induction,

$$\pi_i = \frac{p_0}{q_1} \dots \frac{p_{i-1}}{q_i} \pi_0.$$

This gives us an invariant measure. This measure is a distribution, if and only if

$$\sum_{i=1}^{\infty} \frac{p_0}{q_1} \dots \frac{p_{i-1}}{q_i} < \infty.$$

Note that in Example 4.3.10, we showed that the chain is recurrent if and only if

$$\sum_{i=1}^{\infty} \frac{q_1}{p_1} \dots \frac{q_i}{p_i} = \infty.$$

Consequently, the chain is null-recurrent if and only if this holds simultaneously with the condition

$$\sum_{i=1}^{\infty} \frac{p_0}{q_1} \dots \frac{p_{i-1}}{q_i} = \infty.$$

#### 4.4.2 Convergence

In many cases, we are interested not in calculating the invariant distribution but in sampling from it. This becomes especially relevant if the state space is very large and we cannot even enumerate it efficiently. In these cases the idea is to run the chain for a sufficiently long time so that it becomes in an almost stationary distribution.

This idea is based on the phenomenon of the convergence of the chain probability distribution to the stationary distribution. Namely, as chain  $X_n$  evolves, the transition probabilities  $P_{ij}^n$  become independent from the initial state  $i$  and converge to the invariant distribution  $\pi_j$ . In order to formulate the precise result, we need another definition.

We say that a state  $i$  is *aperiodic* if  $P_{ii}^n > 0$  for all sufficiently large  $n$ .

**Lemma 4.4.7.** *Suppose that Markov chain  $P$  is irreducible and has an aperiodic state  $i$ . Then for all states  $j$  and  $k$ ,  $P_{jk}^n > 0$  for all sufficiently large  $n$ . In particular, all states are aperiodic.*

**Theorem 4.4.8** (Convergence to equilibrium). *Let  $P$  be irreducible and aperiodic, and suppose that  $P$  has an invariant distribution  $\pi$ . Then  $P_{ij}^n \rightarrow \pi_j$  as  $n \rightarrow \infty$  for all  $i$  and  $j$ .*

Note: This is equivalent to the following statement. Let  $\lambda$  be any distribution, and suppose that  $X_0$  has distribution  $\lambda$ . Then,  $\mathbb{P}(X_n = j) \rightarrow \pi_j$  as  $n \rightarrow \infty$  for all  $j$ .

*Proof.* Let  $Y_n$  be a Markov with the same transition matrix as  $X_n$  and with  $Y_0$  distributed according to the invariant distribution  $\pi$ . Assume that  $X_n$  and  $Y_n$  are independent.

Let  $T = \inf\{n \geq 1 : X_n = Y_n = b\}$ , where  $b$  is a particular state.

The process  $W_n = (X_n, Y_n)$  has the state space that consists of pairs  $(i, k)$ . It has transition probabilities  $P_{i,k \rightarrow j,l} = P_{ij}P_{kl}$  and initial distribution  $\lambda_i \pi_k$ .

By aperiodicity of the chain  $\mathbb{P}\{X_n = i, Y_n = k\} = \mathbb{P}\{X_n = i\}\mathbb{P}\{Y_n = k\} > 0$  for all sufficiently large  $n$ , which implies that  $(X_n, Y_n)$  is irreducible. In addition it has an invariant distribution  $\pi_i \pi_k$ .

By Theorem 4.4.3, the chain is positive recurrent, which implies that  $\mathbb{E}(T) < \infty$  and in particular  $\mathbb{P}(T < \infty) = 1$ .

We claim that for any  $n \geq 1$ ,

$$\mathbb{P}(X_n = y, T \leq n) = \mathbb{P}(Y_n = y, T \leq n). \quad (4.5)$$

Indeed, for every  $m \leq n$ , we have

$$\mathbb{P}(X_n = y | T = m) = \mathbb{P}(Y_n = y | T = m),$$

since both probabilities equal  $P_{b,y}^{n-m}$  for  $n > m$  and  $\delta_{b,y}$  for  $n = m$ .

This implies that

$$\mathbb{P}(X_n = y | T \leq n) = \mathbb{P}(Y_n = y | T \leq n),$$

and therefore that equality (4.5) holds.

It follows that


$$\begin{aligned} |\mathbb{P}(X_n = j) - \pi| &= |\mathbb{P}(X_n = j) - \mathbb{P}(Y_n = j)| \\ &= |\mathbb{P}(X_n = j \text{ and } n < T) - \mathbb{P}(Y_n = j \text{ and } n < T)| \\ &\leq \mathbb{P}(n < T). \end{aligned}$$

And  $\mathbb{P}(n < T) \rightarrow 0$ , as  $n \rightarrow \infty$ , because

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(T > n) < \infty.$$

This completes the proof.  $\square$

### 4.4.3 Ergodic theorem

 It is possible to interpret the invariant distribution as the long-run proportion of time spent by the Markov chain in each state. This result is called the ergodic theorem. Define  $V_i(n)$  as the number of visits to  $i$  before time  $n$ :

$$V_i(n) = \sum_{k=0}^{n-1} \mathbb{1}_{\{X_k=i\}}.$$

Then,  $V_i(n)/n$  is the proportion of time before  $n$  spent in state  $i$ .

**Theorem 4.4.9** (Ergodic theorem). *Let  $X_n$  be an irreducible Markov chain. Then*

$$\mathbb{P}\left(\frac{V_i(n)}{n} \rightarrow \frac{1}{m_i} \text{ as } n \rightarrow \infty\right) = 1,$$

where  $m_i = \mathbb{E}(T_i)$  is the expected return time to state  $i$ . Moreover, in the positive recurrent case, for any bounded function  $f$ , we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \rightarrow \bar{f} \text{ as } n \rightarrow \infty\right) = 1,$$

where

$$\bar{f} = \sum_i \pi_i f_i,$$

and  $\pi_i$  is the unique invariant distribution.

*Proof.* For a transient case, the result is clear because then  $V_i(n)$  is finite and  $m_i$  is infinite.

For the recurrent case, we note that

$$S_i^{(1)} + \dots + S_i^{(V_i(n)-1)} \leq n \leq S_i^{(1)} + \dots + S_i^{(V_i(n))},$$

where  $S_i^{(r)}$  is the time between visits to  $i$  number  $(r-1)$  and number  $r$ .

By strong Markov property the random variables  $S_i^{(r)}$  are independent and identically distributed, with expectation  $\mathbb{E}_i S_i^{(r)} = m_i$ . By the strong law of large numbers, with probability 1,

$$\lim_{t \rightarrow \infty} \frac{S_i^{(1)} + \dots + S_i^{(t)}}{t} = m_i.$$

and since the chain is recurrent,  $V_i(n) \rightarrow \infty$  with probability 1. It follows that with probability 1,

$$\frac{n}{V_i(n)} \rightarrow m_i, \text{ as } n \rightarrow \infty,$$

or

$$\frac{V_i(n)}{n} \rightarrow \frac{1}{m_i}, \text{ as } n \rightarrow \infty.$$

For the second statement we assume without loss of generality that  $|f| \leq 1$  and note that

$$\left| \frac{1}{n} \sum_{k=0}^{n-1} f(X_k) - \bar{f} \right| = \left| \sum_i \left( \frac{V_i(n)}{n} - \pi_i \right) f_i \right| \leq \sum_i \left| \frac{V_i(n)}{n} - \pi_i \right|$$



The last sum can be estimated as follows, for an arbitrary choice of a set of states  $J$ ,

$$\begin{aligned}
\sum_i \left| \frac{V_i(n)}{n} - \pi_i \right| &\leq \sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \sum_{i \notin J} \left( \frac{V_i(n)}{n} + \pi_i \right) \\
&= \sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + 1 - \sum_{i \in J} \frac{V_i(n)}{n} + \sum_{i \notin J} \pi_i \\
&= \sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + \sum_{i \in J} \left( \pi_i - \frac{V_i(n)}{n} \right) + 2 \sum_{i \notin J} \pi_i \\
&\leq 2 \sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| + 2 \sum_{i \notin J} \pi_i
\end{aligned}$$

Then, for an arbitrary choice of  $\varepsilon$ , one can find a *finite* set  $J$ , so that  $2 \sum_{i \notin J} \pi_i < \varepsilon/2$ . We know that  $\left| \frac{V_i(n)}{n} - \pi_i \right| \rightarrow 0$  for every fixed  $i$  with probability 1. Since  $J$  is finite it follows that  $\sum_{i \in J} \left| \frac{V_i(n)}{n} - \pi_i \right| \rightarrow 0$  with probability 1. Together, this ensures that

$$\frac{1}{n} \sum_{k=0}^{n-1} n - 1f(X_k) - \bar{f} \rightarrow 0,$$

with probability 1. □

### Exercises

*Ex.* 4.4.10. Five balls are distributed between two urns, labelled  $A$  and  $B$ . Each period, an urn is selected at random, and if it is not empty, a ball from that urn is removed and placed into the other urn. In the long run, what fraction of time is urn  $A$  empty?

*Ex.* 4.4.11. Consider the following transition matrix.

$$P = \begin{bmatrix} 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{bmatrix}.$$

Find all invariant distributions of the corresponding Markov chain.

*Ex.* 4.4.12. A particle moves on eight vertices of a cube in the following way: at each step the particle is equally likely to move to each of the free adjacent vertices, independently of its past motion. Let  $i$  be the initial vertex occupied by the particle,  $o$  be the vertex opposite  $i$ . Calculate each of the following quantities.

1. the expected number of steps until the particle returns to  $i$ ;
2. the expected number of visits to  $o$ , until the first return to  $i$ ;
3. the expected number of steps until the first visit to  $o$ .

#### 4.4.4 Another example of the Markov Chain Monte Carlo (MCMC) algorithm

Here we present another example of the MCMC algorithm. This example is a baby-version of some problems from Bayesian statistics. In contrast to everything in previous sections, in this example the state space is uncountable, so we allow the Markov chain to take continuous values.

Consider that observations  $Y_i$  are independent for  $i = 1, 2, \dots, n$  and have the distribution

$$y_i \sim \mathcal{N}(\mu, \tau^{-1}).$$

We model  $\mu, \tau^{-1}$  as random variables with prior distribution

$$\begin{aligned} \mu &\sim \mathcal{N}(\theta_0, \varphi_0^{-1}), \\ \tau &\sim \Gamma(\alpha_0, \beta_0), \end{aligned}$$

and  $\mu$  and  $\tau$  are independent under prior distribution.

It is assumed here that the parameters of the prior distribution are known. (The choice of the prior is one of the fundamental issues in Bayesian statistics.)

We would like to sample from the posterior joint distribution of the  $\mu$  and  $\tau$  with the goal to learn about some statistics of this distribution.

By Bayes' formula the posterior distribution is

$$\begin{aligned} \pi(\mu, \tau|y) &= \frac{f(y|\mu, \tau)\pi(\mu, \tau)}{f(y)} \\ &= \frac{f(y|\mu, \tau)\pi(\mu, \tau)}{\iint f(y|\mu, \tau)\pi(\mu, \tau) d\mu d\tau} \end{aligned} \quad (4.6)$$

It is clear that

$$\begin{aligned} f(y|\mu, \tau)\pi(\mu, \tau) &= (2\pi)^{-n/2}\tau^{n/2} \exp\left(-\frac{\tau}{2} \sum_i (y_i - \mu)^2\right) \\ &\quad \times (2\pi)^{-1/2}\varphi_0^{1/2} \exp\left(-\frac{\varphi_0}{2}(\mu - \theta_0)^2\right) \\ &\quad \times \beta_0^{-\alpha_0}\Gamma(\alpha_0)^{-1}\tau^{\alpha_0-1}e^{-\beta_0\tau}. \end{aligned}$$

In this example, the denominator in (4.6) can be obtained by direct numeric integration of this formula and the statistics of the posterior distribution  $\pi(\mu, \tau|y)$  can be calculated by the direct numeric integration against this distribution.

However, in more complex examples, the number of parameters is typically larger than two, and the direct integrations are difficult.

Instead, we want to sample points  $(\mu_i, \tau_i)$  from the posterior distribution  $\pi(\mu, \tau|y)$  without explicitly calculating this distribution.

In order to do this, we build a Markov chain on the state space of the pairs  $(\mu, \tau)$  so that the invariant distribution of this chain coincides with the posterior distribution  $\pi(\mu, \tau|y)$ .

We will use the Gibbs sampler to build this chain. Given a realization of  $(\mu_j, \tau_j)$ , we will update  $\mu_j$  and  $\tau_j$  in an alternating fashion, keeping the other variable fixed. The probabilities of the updates will be conditional probabilities given the value of the other variable fixed.

These conditional probability distributions are easy to compute:

$$\pi(\mu|y, \tau) = \mathcal{N}(\theta_n, \varphi_n),$$

where

$$\theta_n = \frac{\varphi_0 \theta_0 + \tau \sum_i y_i}{\varphi_0 + n\tau},$$

$$\varphi_n = \varphi_0 + n\tau,$$

and

$$\pi(\tau|\mu, y) = \Gamma(\alpha_n, \beta_n),$$

where

$$\alpha_n = \alpha_0 + \frac{n}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} \sum_i (y_i - \mu)^2.$$

In a more complicated case we have  $m$  groups of observations, with  $n$  observations in which of them. So, our data is  $Y_{ij}$ , where  $i = 1, \dots, n$  and  $j = 1, \dots, m$ . The model is

$$Y_{ij} = \mathcal{N}(\mu_j, \tau^{-1}).$$

The prior distributions on the parameters are as before,

$$\mu_j \sim \mathcal{N}(\theta_0, \varphi_0^{-1}),$$

$$\tau \sim \Gamma(\alpha_0, \beta_0),$$

It is easy to calculate the joint distribution and to conclude that the posterior conditional distributions on  $\mu_j$  and  $\tau$ , are the normal and the gamma distributions, respectively, with the parameters

$$\theta_{j,n} = \frac{\varphi_0 \theta_0 + \tau \sum_i y_{ij}}{\varphi_0 + n\tau},$$

$$\varphi_n = \varphi_0 + n\tau,$$

and

$$\alpha_n = \alpha_0 + \frac{mn}{2},$$

$$\beta_n = \beta_0 + \frac{1}{2} \sum_{i,j} (y_{ij} - \mu_j)^2.$$

Then, the samples from the posterior distribution  $\pi(\mu, \tau|y)$  can be calculated by a Gibbs sampler method.

# Chapter 5

## Martingales

### 5.1 Filtrations and Stopping Times



Filtrations, stopping times, and martingales are all related to the idea of “the information available at the present time.” This is represented by an increasing family of  $\sigma$ -fields indexed by time and random variables measurable with respect to these  $\sigma$ -algebras.

**Definition 5.1.1.** A *filtration*  $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ , is an increasing sequence of  $\sigma$ -algebras.

We use  $\mathcal{F}_\infty$  to denote  $\sigma(\bigcup_{n=0}^\infty \mathcal{F}_n)$ .

The  $\sigma$ -algebra  $\mathcal{F}_n$  is often interpreted as a  $\sigma$ -algebra generated by “events that are determined by time  $n$ .”

For example, if  $X_1, X_2, \dots$  is a sequence of random variables on probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , then it defines a filtration

$$\mathcal{F}_0 = \{\emptyset, \Omega\} \subset \mathcal{F}_1 = \sigma(X_1) \subset \mathcal{F}_2 = \sigma(X_1, X_2) \subset \dots$$

Conversely we say that a sequence of random variables  $\{X_n\}$  is *adapted* to a filtration  $\{\mathcal{F}_n\}$  if  $X_n$  is measurable with respect to  $\mathcal{F}_n$  for all  $n \geq 1$ . (Sometimes we will abuse notation and write  $X_n \in \mathcal{F}_n$  to indicate that  $X_n$  is measurable with respect to  $\mathcal{F}_n$ .)

A sequence of random variables  $\{X_n\}$  is *predictable* if  $X_n$  is measurable with respect to  $\mathcal{F}_{n-1}$ , for all  $n \geq 1$ .

**Definition 5.1.2.** A *stopping time*  $T$  is a random variable  $T : \Omega \rightarrow \mathbb{Z}^+ \cup \{\infty\}$  such that the event  $\{T = n\}$  is  $\mathcal{F}_n$ -measurable: for every  $n < \infty$ ,  $\{T = n\} \in \mathcal{F}_n$ .

Intuitively, if  $X_i$  is the size of your win/loss at time  $i$  in a casino, then your decision to stop the game at the end of period  $n$  should not depend on the value of  $X_i$  for  $i > n$ .

**Example:** Consider the filtration generated by random variables  $X_i$ ,  $i = 1, \dots, N$ , and the random variable

$$T = \text{first index } i \leq N \text{ s.t. } X_i = \max_{1 \leq j \leq N} X_j.$$

Then,

$$(T = n) = (X_1 < X_n, \dots, X_{n-1} < X_n, X_{n+1} \leq X_n, \dots, X_N \leq X_n).$$

Clearly  $(T = n) \in \mathcal{F}_N$ , but  $(T = n) \notin \mathcal{F}_n$  in general. So,  $T$  is not a stopping time. Another example is  $T = \sup\{n : X_n \in A\}$ , the time of the last visit to a set  $A$ .

Another example of a strategy which is not a stopping time is the following rule for cooking toast: “cook toast until 10 seconds before it starts to smoke.”

Here are examples of valid stopping times.

- Constant stopping time:  $T(\omega) = k$ .
- The time of the first visit to a set  $A$ , that is,  $T = \inf\{n : X_n \in A\}$ .
- If  $T$  is a stopping time and  $f(t) : \mathbb{Z}^+ \rightarrow \mathbb{Z}^+$  is a non-decreasing function that has the property that  $f(t) \geq t$  for all  $t \in \mathbb{Z}^+$ , then  $T' = f(T)$  is again a stopping time.
- If  $T_1$  and  $T_2$  are stopping times, so are  $T_1 \wedge T_2$  and  $T_1 \vee T_2$ .

The last property implies that any stopping time  $T$  is an increasing limit of bounded stopping times  $T_n = T \wedge n$ .

Just as we have  $\sigma$ -algebras  $\mathcal{F}_n$  associated with constant times, we do have a  $\sigma$ -algebra  $\mathcal{F}_T$  associated to any stopping time. This is the information we have when we observe the chain up to time  $T$ . Formally,

$$\mathcal{F}_T = \{A : A \in \mathcal{F}_\infty \text{ and } A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for each } n\}.$$

One can check from the definition that  $T$  is  $\mathcal{F}_T$  measurable and so is  $X_T$  on the set  $T < \infty$ .

## 5.2 Definition of Martingales

**I**n a  $\{\mathcal{F}_n\}$ -adapted sequence of random variables  $\{X_n\}$  is a *martingale* with respect to  $\mathcal{F}_n$  if

1.  $\mathbb{E}|X_n| < \infty$ , and
2.  $\mathbb{E}(X_{n+1} | \mathcal{F}_n) = X_n$  for every  $n \geq 0$ .

We define  $X_0 = \mathbb{E}X_1$  for convenience.

It is useful to think about a martingale  $X_n$  as total winnings in a fair game.

From the definition of a martingale and the tower property of the conditional expectations, it is immediate that  $\mathbb{E}(X_p|\mathcal{F}_n) = X_n$  for  $p > n$ .

An adapted sequence with finite means is called a *submartingale* if  $X_n \leq \mathbb{E}(X_{n+1}|\mathcal{F}_n)$ , and is called a *supermartingale* if  $X_n \geq \mathbb{E}(X_{n+1}|\mathcal{F}_n)$ .

Super-martingale is not an especially “super” thing. It represents winnings in a losing game where your expected wealth in the next hour is less than it is now.

Note that if  $\{X_i\}$  is a submartingale then  $\{-X_i\}$  is a supermartingale and vice versa. For that reason we will usually formulate results either only for submartingales or only for supermartingales. For the other case, the results can be obtained by obvious transformations.

Note also that a martingale is both a submartingale and a supermartingale.

For a submartingale  $X_n$ , we have that  $\mathbb{E}(X_n \leq \mathbb{E}(X_{n+1}))$ . So, a submartingale increases on average and it can be seen as a stochastic analogue of a non-decreasing sequence. In particular, we will see later that if a submartingale is bounded from above then it converges.

We can define  $Y_{i+1} = X_{i+1} - X_i$  and see that

$$\mathbb{E}(Y_{i+1}|\mathcal{F}_i) = 0 \text{ for every } i \geq 0.$$

Such sequences are called *martingale differences*. If  $Y_i$  is a martingale difference, then we can recover the corresponding martingale as

$$X_n = X_0 + Y_1 + Y_2 + \dots + Y_n.$$

### Examples of Martingales

1. (Sums of independent random variables.) If  $Y_i$  is a sequence of independent random variables such that  $\mathbb{E}|Y_i| < \infty$  for all  $i \geq 1$  and  $X_n = \sum_{i=1}^n Y_i$  then the sequence  $\{X_n\}$  is a martingale.

2. (Learning Martingale.) If  $\{\mathcal{F}_i\}$  is an increasing sequence of  $\sigma$ -algebras and  $X$  is an integrable random variable on  $(\Omega, \mathcal{F}, \mathbb{P})$ , we define  $X_i = \mathbb{E}(X|\mathcal{F}_i)$ . Then  $X_i$  is a martingale with respect to  $\{\mathcal{F}_i\}$ .

3. (Martingale Transforms.)

**Definition 5.2.1.** Suppose  $X_n$  is a martingale with respect to  $\{\mathcal{F}_n\}$ ,  $Y_n$  are the differences  $X_n - X_{n-1}$ , and  $a_n$  is a predictable sequence (that is,  $a_n \in \mathcal{F}_{n-1}$ ). Then, a *martingale transform*  $X'_n = (a \cdot X)_n$  is given by the formula

$$X'_n = X'_{n-1} + a_n Y_n. \quad (5.1)$$

**Lemma 5.2.2.** If  $X_n$  is a  $\mathcal{F}_n$ -martingale, and  $a_n$  is a predictable sequence such that  $a_n X_n$  integrable for each  $n$ , then  $(a \cdot X)$  is an  $\mathcal{F}_n$ -martingale.

*Proof.*  $Y_n = (a \cdot X)_n$  is an  $\mathcal{F}_n$ -martingale if

$$\mathbb{E}(Y_n - Y_{n-1} | \mathcal{F}_{n-1}) = \mathbb{E}(a_n(X_n - X_{n-1}) | \mathcal{F}_{n-1}) = 0.$$

Using  $a_n \in \mathcal{F}_{n-1}$ ,  $X_{n-1} \in \mathcal{F}_{n-1}$  and  $X_n$  a martingale, we have

$$\mathbb{E}(Y_n - Y_{n-1} | \mathcal{F}_{n-1}) = a_n(\mathbb{E}(X_n | \mathcal{F}_{n-1}) - X_{n-1}) = 0.$$

□

Now, how we can generate submartingales?

**Lemma 5.2.3.** Suppose  $\{(X_i, \mathcal{F}_i)\}$  is a martingale and  $\varphi$  is a convex function of one variable such that  $\varphi(X_i)$  is integrable for every  $i$ . Then  $\{(\varphi(X_i), \mathcal{F}_i)\}$  is a submartingale.

*Proof.* This is a consequence of Jensen's inequality for conditional expectations. □

In particular, for any  $p \geq 1$ , if  $\{X_i\}$  is a martingale and  $\mathbb{E}|X_i|^p < \infty$  for all  $i$ , then  $\{|X_i|^p\}$  is a submartingale. For example, if  $p = 2$ , then we see that the variance of  $\{X_i\}$  is increasing with  $i$ .

The following inequality generalizes Markov's inequality to the setting of martingale sequences.

**Theorem 5.2.4** (Doob's inequality I). Suppose  $\{X_i\}$  is a martingale sequence of length  $n$ . Then

$$\mathbb{P}\left\{\omega: \sup_{1 \leq i \leq n} |X_i| \geq l\right\} \leq \frac{1}{l} \int |X_n| d\mathbb{P}$$

*Proof.* Let us define  $S(\omega) = \sup_{1 \leq j \leq n} |X_j(\omega)|$  and  $E = \{\omega : S(\omega) \geq l\}$ .

We can represent  $E$  as a union of disjoint events  $E_j$ , where  $E_j$  is the event that  $S(\omega)$  achieved  $l$  for the first time at time  $j$ .

$$E_j = \{\omega : |X_1(\omega)| < l, \dots, |X_{j-1}(\omega)| < l, |X_j(\omega)| \geq l\}.$$

We have

$$\mathbb{P}(E_j) \leq \frac{1}{l} \int_{E_j} |X_j| d\mathbb{P} \leq \frac{1}{l} \int_{E_j} |X_n| d\mathbb{P}$$

The second inequality follows from the fact that  $|X_j|$  is a submartingale. In particular  $\mathbb{E}(|X_n| | \mathcal{F}_j) \geq |X_j|$  a.e. and  $E_j \in \mathcal{F}_j$ . Summing up the inequality over  $j = 1, \dots, n$  we obtain

$$\mathbb{P}\left\{\omega : \sup_{1 \leq i \leq n} |X_i| \geq l\right\} \leq \frac{1}{l} \int_E |X_n| d\mathbb{P} \leq \frac{1}{l} \int |X_n| d\mathbb{P}.$$

□

In the proof, we could have started with

$$\mathbb{P}(E_j) \leq \frac{1}{l^p} \int_{E_j} |X_j|^p d\mathbb{P}$$

and then we would obtain for  $p \geq 1$ ,

$$\mathbb{P}(E) \leq \frac{1}{l^p} \int_E |X_n|^p d\mathbb{P}$$

In particular, for  $p = 2$ , this gives a generalization of Kolmogorov's inequality for sums of independent random variables.

**Theorem 5.2.5.** *If  $X_n$  is a martingale and  $T$  is a stopping time, then  $X_{n \wedge T}$  is also a martingale.*

*Proof.* This follows from Lemma 5.2.2. Indeed,  $a_n = \mathbb{1}_{(T > n-1)} \in \mathcal{F}_{n-1}$  is the predictable bounded sequence and  $X_{n \wedge T} = (a_n \cdot X_n)$ . Hence it is a martingale. □

What about the random variable  $X_T$ ? These are your winnings at the stopping time. Can we claim that  $\mathbb{E}X_T = X_0$ ?

Now, if  $X_n$  is a martingale and  $T$  is a stopping time bounded by  $b$ , then

$$\mathbb{E}(X_T) = \mathbb{E}(X_{T \wedge b}) = X_0,$$

because  $X_{T \wedge b}$  is a martingale by Theorem 5.2.5.

In the case of an unbounded stopping time  $T$ , we have that  $X_{T \wedge n} \rightarrow X_T$  a.s. as  $n \rightarrow \infty$ .

Hence, if we could exchange expectations and limits, then we would write:

$$\mathbb{E}(X_T) = \mathbb{E}\left(\lim_{n \rightarrow \infty} X_{T \wedge n}\right) = \lim_{n \rightarrow \infty} \mathbb{E}(X_{T \wedge n}) = X_0. \quad (5.2)$$

However, this exchange is not always valid.




*Example 5.2.6.* Consider a random symmetric walk  $S_n$  starting at  $S_0 = 1$  and let  $T = \inf\{n | S_n = 0\}$ .

Then we have  $\mathbb{E}S_T = 0$  because  $\mathbb{P}(T < \infty) = 1$ , so we eventually hit zero with probability 1. However, this is different from  $S_0 = 1$ . Despite the fact that we play a fair game, we will eventually lose.

In fact the basic theorems of real analysis imply that a sufficient condition for validity of the exchange is that the sequence of random variables  $X_n$  is uniformly bounded. Hence if  $X_n$  is a martingale,  $T$  is a stopping time, and  $X_{T \wedge n}$  is uniformly bounded then  $\mathbb{E}X_T = X_0$ .

### 5.3 Stopping times and martingales: Examples

*Example 5.3.1 (Gambler ruin).*

 Suppose that we have a coin, with probability  $p$  of heads,  $q = 1 - p$  of tails. Let us define i.i.d. random variables  $X_i$  by  $X_i = 1$  when the  $i^{\text{th}}$  coin toss is a head, and  $-1$  when the  $i^{\text{th}}$  coin toss is a tail.

Define  $S_0 = a$  where  $a$  is a positive integer, and let  $S_n = S_0 + X_1 + \dots + X_n$ . Let  $T = \inf\{n: S_n = 0 \text{ or } b\}$  for  $b > a$  an integer. We argue that  $\mathbb{P}(T < \infty) = 1$  and we want to find  $\mathbb{P}(S_T = b)$  and  $\mathbb{P}(S_T = 0)$ .

First, consider the case of a fair coin  $p = q = 1/2$ . We know that  $S_n$  is a martingale. What we really need is that  $\mathbb{E}(S_T) = \mathbb{E}S_0 = a$ , which holds here because  $S_{T \wedge n}$  is uniformly bounded. If  $T_x = \inf\{n | S_n = x\}$ , then

$$\mathbb{E}(S_T) = 0 \times \mathbb{P}(T_0 < T_b) + b \times (1 - \mathbb{P}(T_0 < T_b)) = a.$$

and

$$\mathbb{P}(S_T = 0) = \mathbb{P}(T_0 < T_b) = 1 - \frac{a}{b}.$$

Now consider the unfair case  $p \neq q$ . Here  $S_n$  is *not* a martingale and so the idea is to find a suitable  $h(x)$  such that  $h(S_n)$  is a martingale.

If  $h(S_n)$  were a martingale, we would have that  $h(x) = ph(x+1) + qh(x-1)$ . So let us try  $h(x) = z^x$ , where  $z$  will be determined. Then, we should have  $z^x = pz^{x+1} + qz^{x-1}$ , and in particular,  $z = pz^2 + q$ . The roots of this quadratic equation are 1 and  $\frac{q}{p}$ . So we conclude that  $(\frac{q}{p})^{S_n}$  is a martingale.

For  $T = \inf\{n: S_n = 0 \text{ or } b\}$ , we observe that  $(q/p)^{S_T}$  is bounded between  $(q/p)^0$  and  $(q/p)^b$ , so  $(q/p)^{S_{n \wedge T}}$  is a bounded martingale. This implies that  $\mathbb{E}[(q/p)^{S_T}] = (q/p)^a$ , where  $S_0 = a$ .

Therefore,

$$\mathbb{P}(S_T = b)(q/p)^b + \mathbb{P}(S_T = 0)(q/p)^0 = (q/p)^a.$$

We also have

$$\mathbb{P}(S_T = b) + \mathbb{P}(S_T = 0) = 1.$$

Solving these two equations, we obtain

$$\mathbb{P}(S_T = b) = \frac{(q/p)^a - 1}{(q/p)^b - 1}$$

and

$$\mathbb{P}(S_T = 0) = \frac{(q/p)^b - (q/p)^a}{(q/p)^b - 1}.$$

*Example 5.3.2* (Wald's First Identity). Let  $X_1, X_2, X_3, \dots$  be i.i.d. random variables with  $\mathbb{E}|X_i| < \infty$ ,  $S_n = X_1 + X_2 + \dots + X_n$ , and  $\mathcal{F}_n = \sigma(X_1, X_2, \dots, X_n)$ . If  $T$  is a stopping time for filtration  $\{\mathcal{F}_n\}$  and  $\mathbb{E}(T) < \infty$ , then  $\mathbb{E}S_T = \mathbb{E}X_1\mathbb{E}T$ .

*Proof.*

$$\mathbb{E}S_T = \mathbb{E}\left(\sum_{n=1}^T X_n\right) = \mathbb{E}\left(\sum_{n=1}^{\infty} X_n \mathbb{1}_{(T \geq n)}\right) = \sum_{n=1}^{\infty} \mathbb{E}(X_n \mathbb{1}_{(T \geq n)})$$

Note that event  $(T \geq n)$  is measurable with respect to  $\mathcal{F}_{n-1}$ . (At the end of period  $n - 1$  we know that we have not yet stopped and therefore  $T \geq n$ , although we do not know if we stop at period  $n$ .) Since  $X_n$  is independent of  $\mathcal{F}_{n-1}$ , therefore,

$$\mathbb{E}S_T = \mathbb{E}X_1 \mathbb{E}\left(\sum_{n=1}^{\infty} \mathbb{1}_{(T \geq n)}\right) = \mathbb{E}X_1\mathbb{E}T.$$

□

*Ex. 5.3.3.* If  $X_n$  is a martingale such that the differences  $Y_n = X_n - X_{n-1}$  are all square integrable, show that for  $n \neq m$ ,  $\mathbb{E}(Y_n Y_m) = 0$ . Therefore

$$\mathbb{E}(X_n^2) = \mathbb{E}(X_0^2) + \sum_{j=1}^n \mathbb{E}Y_j^2.$$

If, in addition,  $\sup_n \mathbb{E}[X_n^2] < \infty$ , then show that there is a random variable  $X$  such that

$$\lim_{n \rightarrow \infty} \mathbb{E}(|X_n - X|^2) = 0.$$

*Example 5.3.4* (Wald's second identity). Let  $X_1, X_2, \dots$  be i.i.d. with  $\mathbb{E}X_n = 0$  and  $\mathbb{E}X_n^2 = \sigma < \infty$ . If  $T$  is a stopping time with  $\mathbb{E}T < \infty$  then show that  $\mathbb{E}S_T^2 = \sigma^2\mathbb{E}T$ .

*Proof.* Recall that  $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ . Let  $M_n := S_n^2 - n\sigma^2$ , where  $\sigma^2 = \mathbb{E}[X^2]$ . Check that  $M_n$  is a martingale:

$$\begin{aligned} \mathbb{E}[M_{n+1}|\mathcal{F}_n] &= \mathbb{E}[S_{n+1}^2 - (n+1)\sigma^2|\mathcal{F}_n] \\ &= \mathbb{E}[S_n^2 + 2X_{n+1}S_n + X_{n+1}^2 - (n+1)\sigma^2|\mathcal{F}_n] \\ &= S_n^2 - n\sigma^2 \\ &= M_n. \end{aligned}$$

First, if  $\mathbb{P}(T \leq N) = 1$  for some non-random  $N < \infty$ . Then, we know  $0 = \mathbb{E}[M_0] = \mathbb{E}[M_T] = \mathbb{E}[S_T^2 - T\sigma^2]$ . Hence, in this case  $\mathbb{E}[S_T^2] = \mathbb{E}[T]\sigma^2$ .

Now consider the general case where only  $\mathbb{E}T < \infty$  is assumed. We have  $\mathbb{E}[S_{T \wedge n}^2] = \mathbb{E}[T \wedge n]\sigma^2$  for every  $n = 1, 2, 3, \dots$ . Since  $T \wedge n$  is an increasing sequence of positive random variable bounded by  $T$  and converging a.s. to  $T$  as  $n$  increases, hence  $\mathbb{E}[T \wedge n] \uparrow \mathbb{E}[T] < \infty$ .

There is a possibility that  $S_{T_n}$  converges to  $S_T$  a.s. but not in  $L^2$ . This could prevent us from taking the limit on the left-hand side of  $\mathbb{E}[S_{T \wedge n}^2] = \mathbb{E}[T \wedge n]\sigma^2$ . In order to rule out this possibility, note that  $S_{T \wedge n}$  is a martingale with square integrable increments and  $\mathbb{E}[S_{T \wedge n}^2] = \mathbb{E}[T \wedge n]\sigma^2 < \sigma^2 \mathbb{E}[T] < \infty$ . So, by applying the result in Exercise 5.3.3 to martingale  $S_{T \wedge n}$ , we find that  $S_{T \wedge n}$  converges to a limit in  $L^2$ . Since we know that  $S_{T \wedge n} \rightarrow S_T$  a.s., the limit in  $L^2$  is also  $S_T$ . Therefore  $\mathbb{E}[S_T^2] = \lim_{n \rightarrow \infty} \mathbb{E}[S_{T \wedge n}^2] = \mathbb{E}[T]\sigma^2$ .  $\square$

## 5.4 Martingale convergence theorem

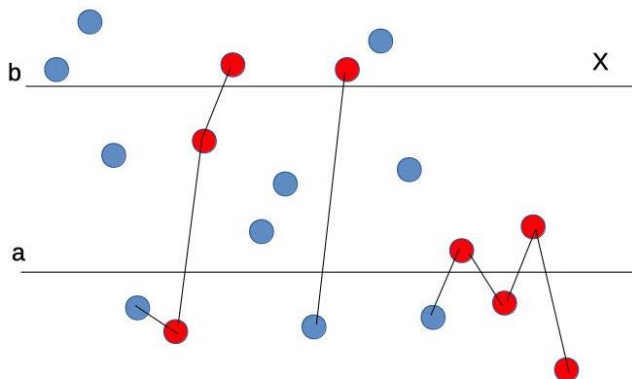


Figure 5.1

**S**uppose that  $X_n$ ,  $n \geq 0$  is a martingale modeling the fortune of a gambler (who is playing a fair game) at time  $n$ , and let  $a$  and  $b$  be two real numbers such that  $X_0 \leq a < b$ . *Upcrossing* is defined as a situation when a sequence of random variables goes from below the lower threshold  $a$  to above the upper threshold  $b$  (see Figure 5.1).

Let  $U_{a,b}(n)$  be the number of upcrossings of  $[a, b]$  by the sequence  $X_0, X_1, \dots, X_n$  and  $U_{a,b}$  be the number of upcrossings of  $[a, b]$  by the entire sequence (so  $U_{a,b}$  can be infinite).

If  $X_n$  is a stock price, and you buy a stock whenever  $X_n \leq a$  and sell it when  $X_n \geq b$ , then every time that an upcrossing completes, you make a profit of at least  $\$(b - a)$ . Our goal is to study the number of upcrossings.

Recall that a super-martingale  $X_n$  is a stochastically declining sequence when  $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ .

**Theorem 5.4.1** (Doob's Upcrossing Inequality). *Let  $\{X_n\}$  be a supermartingale, and  $\mathbb{E}X_0 < a < b$ . Then for every  $n$ ,*

$$\mathbb{E}U_{a,b}(n) \leq \frac{\mathbb{E}(a - X_n)^+}{b - a} \leq \frac{|a| + \mathbb{E}|X_n|}{b - a}.$$

This theorem allows us to conclude that for  $L^1$ -bounded martingales (i.e., when the sequence  $\mathbb{E}|X_n|$  is bounded), the total number of upcrossings  $U_{a,b}$  is bounded with probability one. Indeed,  $\mathbb{E}U_{a,b}(n)$  is increasing and bounded. Hence  $\mathbb{E}U_{a,b} = \lim_{n \rightarrow \infty} \mathbb{E}U_{a,b}(n) < \infty$  and this implies that  $U_{a,b}$  is bounded with probability 1.

*Proof.* Let  $C_1 = \mathbb{1}_{\{X_0 < a\}}$ , and

$$C_n = \mathbb{1}_{\{C_{n-1}=1\}} \mathbb{1}_{\{X_{n-1} \leq b\}} + \mathbb{1}_{\{C_{n-1}=0\}} \mathbb{1}_{\{X_{n-1} < a\}} \text{ for } n \geq 2.$$

In words,  $C_n$  equals 1 when either  $C_{n-1} = 1$  and the stock price at time  $n-1$  was less than upper threshold  $b$  or if  $C_{n-1} = 0$  but the prices at time  $n-1$  dropped below lower threshold  $a$ . That is  $C_n$  is 1 if the trader uses the upcrossing strategy by trading one share of stock.

Define  $Y_n = (C \cdot X)_n$ . This is the wealth of the trader at time  $n$ . The sequence  $C_n$  is predictable and bounded, and therefore  $Y_n$  is a supermartingale. Hence  $\mathbb{E}Y_n < 0$ . On the other hand, it is obvious from the picture that

$$Y_n(\omega) \geq (b - a)U_{a,b}(n)(\omega) - (a - X_n(\omega))^+.$$

Every upcrossing of  $[a, b]$  increases the  $Y$ -value by at least  $(b - a)$ , while the  $(a - X_n(\omega))^+$  overestimates the loss during the last interval of play.

By taking expectations on both sides, we find that

$$(b - a)\mathbb{E}U_{a,b}(n) \leq \mathbb{E}(a - X_n)^+,$$

as claimed. □

If  $\{X_n\}$  is a supermartingale and  $\sup_n \mathbb{E}|X_n| < \infty$ , then  $\{X_n\}$  is called  $L^1$ -bounded supermartingale.

**Theorem 5.4.2** (Doob's Martingale Convergence Theorem). *If  $\{X_n\}$  is an  $L^1$ -bounded supermartingale, then as  $n \rightarrow \infty$ ,  $X_n$  converges almost surely to a limit  $X$  with  $\mathbb{E}|X| < \infty$ .*

*Proof.* From the upcrossing inequality we can infer that for every interval  $[a, b]$  the number of upcrossings is finite with probability 1.

If  $X_n$  does not converge, then either there is an interval that is crossed infinitely many times, which is impossible by the above argument, or  $X_n$  converges to  $+\infty$  or  $-\infty$ .

However the latter is impossible by Fatou's lemma:

$$\begin{aligned}\mathbb{E}(|X_\infty|) &= \mathbb{E}(\liminf |X_n|) \leq \liminf \mathbb{E}(|X_n|) \\ &\leq \sup \mathbb{E}(|X_n|) < \infty,\end{aligned}$$

so that


$$\mathbb{P}(X_\infty \text{ is finite}) = 1.$$

□

**Corollary 5.4.3.** *If  $\{X_n\}$  is a non-negative supermartingale, then  $X_\infty = \lim X_n$  exists almost surely.*

*Proof.* The martingale  $\{X_n\}$  is bounded in  $L^1$ , since  $\mathbb{E}|X_n| = E(X_n) \leq E(X_0)$ . □

## 5.5 Uniformly integrable martingales

 Sometimes we are interested to know when the limit of a martingale sequence has the expectation equal to the limit of expectations.

In general, convergence in probability can be upgraded to convergence in  $L^1$  if the converging sequence of functions  $X_n$  is uniformly integrable (“UI”). (See Appendix.)

A class  $\mathcal{C}$  of random variables is called *uniformly integrable* (UI) if given  $\varepsilon > 0$ , there exists  $K \in [0, \infty)$  such that

$$\mathbb{E}(|X|I_{|X| \geq K}) \leq \varepsilon$$

for all  $X \in \mathcal{C}$ .

Note that if the sequence  $X_n$  is uniformly integrable then it is automatically  $L^1$  bounded.

If, in addition,  $X_n$  is a martingale, then by Doob's theorem we can conclude that  $X_n$  converges a.s. to an integrable r.v.  $X_\infty$ , and, by uniform integrability,  $\mathbb{E}X_n \rightarrow \mathbb{E}X$ .

There are two useful sufficient conditions for  $X_n$  to be a uniformly integrable (UI) martingale. First, if  $X_n$  is  $L^p$ -bounded for a  $p > 1$ , then  $X_n$  is uniformly integrable. (If  $X_n$  is only  $L^1$ -bounded then it is not necessarily UI.)

Second, if  $X_n = \mathbb{E}(X|\mathcal{F}_n)$  for an integrable r.v.  $X$  and a filtration  $\{\mathcal{F}_n\}$ , then  $X_n$  is uniformly integrable.

This is a consequence of the following result.

**Theorem 5.5.1.** *Let  $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$  and let  $\mathcal{G}$  vary over all sub- $\sigma$ -fields of  $\mathcal{F}$ . The family  $\{\mathbb{E}(X|\mathcal{G}) : \mathcal{G} \subset \mathcal{F}\}$  is uniformly integrable.*

For the proof we need a lemma. Let us write  $\mathbb{E}(X; A)$  to denote  $\mathbb{E}(X\mathbb{1}_A)$ .

**Lemma 5.5.2.** *Suppose  $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ . Then, given  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that for every  $A \in \mathcal{F}$ ,  $\mathbb{P}(A) < \delta$  implies that  $\mathbb{E}(|X|; A) < \varepsilon$ .*

*Proof of Lemma.* If the conclusion is false, then, for some  $\epsilon_0 > 0$ , we can find a sequence  $A_n$  of elements of  $\mathcal{F}$ , such that  $\mathbb{P}(A_n) < 2^{-n}$  and  $\mathbb{E}(|X|; A_n) \geq \epsilon_0$ . Let

$$H = \limsup A_n = \bigcap_{m} \bigcup_{n \geq m} A_n = \{\omega : \omega \in A_n \text{ for infinitely many } n\}.$$

Then by Borel-Cantelli Lemma,  $P(H) = 0$ , however,  $|X|\mathbb{1}_{A_n}$  is bounded above by a integrable  $|X|$  hence, the ‘Reverse’ Fatou Lemma shows that the expectation when we pass in the limit to  $|X|\mathbb{1}_{A_n}$  can only jump upward, hence  $\mathbb{E}(|X|; H) \geq \epsilon_0$ , and we have arrived at the required contradiction.  $\square$

*Proof of Theorem: 5.5.1.* We denote  $\mathbb{E}(X|\mathcal{G})$  by  $X_{\mathcal{G}}$ . Let  $\epsilon > 0$  be given. Choose  $\delta > 0$  such that, for all  $A \in \mathcal{F}$ ,

$$\mathbb{P}(A) < \delta \text{ implies that } \mathbb{E}(|X|; A) < \epsilon.$$

Choose  $K$  so that  $K^{-1}\mathbb{E}(|X|) < \delta$

By Jensen’s inequality for conditional expectations,  $|X_{\mathcal{G}}| \leq |X|_{\mathcal{G}}$ .

Hence  $\mathbb{E}|X_{\mathcal{G}}| \leq \mathbb{E}|X|$ , and

$$K\mathbb{P}(|X_{\mathcal{G}}| > K) \leq \mathbb{E}|X_{\mathcal{G}}| \leq \mathbb{E}|X|,$$

so that

$$\mathbb{P}(|X_{\mathcal{G}}| > K) < \delta.$$

Since  $\{|X_{\mathcal{G}}| > K\} \in \mathcal{G}$ , hence

$$\begin{aligned} \mathbb{E}(|X_{\mathcal{G}}|; \{|X_{\mathcal{G}}| > K\}) &\leq \mathbb{E}(|X|_{\mathcal{G}}; \{|X_{\mathcal{G}}| > K\}) \\ &= \mathbb{E}(|X|; \{|X_{\mathcal{G}}| > K\}) \leq \epsilon, \end{aligned}$$

where the equality holds by the definition of the conditional expectation.  $\square$

This situation is in fact the most general as the following theorem shows.

**Theorem 5.5.3.** *Let  $X$  be a UI martingale. Then  $X_{\infty} := \lim X_n$  exists a.s. and in  $L^1$ . Moreover for every  $n$ ,*

$$X_n = \mathbb{E}(X_{\infty}|\mathcal{F}_n).$$

*Proof.* We already know that  $X$  converges a.s. and  $L^1$ . It remains to show the last property. Let  $A \in \mathcal{F}_n$  and let  $r \geq n$ . Then

$$\begin{aligned} |\mathbb{E}(X_r; A) - \mathbb{E}(X_{\infty}; A)| &\leq \mathbb{E}(|X_r - X_{\infty}|; A) \\ &\leq \mathbb{E}(|X_r - X_{\infty}|) \rightarrow 0, \end{aligned}$$

as  $r \rightarrow \infty$ , because  $X_r$  converges to  $X_{\infty}$  in  $L^1$ . On the other hand we know by martingale property that

$$\mathbb{E}(X_r; A) = \mathbb{E}(X_n; A).$$

By taking the limit  $r \rightarrow \infty$  we find that

$$\mathbb{E}(X_{\infty}; A) = \mathbb{E}(X_n; A),$$

which means by definition that  $X_n = \mathbb{E}(X_{\infty}|\mathcal{F}_n)$ .  $\square$

## 5.6 Regular stopping times

In many cases we don't have a uniformly integrable or even  $L^1$ -bounded martingale, so we cannot use the criterion of the previous section directly. In these cases, we hope that we can stop martingales so that the "stopped" martingale is uniformly integrable.

**Definition 5.6.1.** A stopping time  $T$  is called *regular* for the martingale  $\{X_n\}$  if the martingale  $\{X_{n \wedge T}\}$  is uniformly integrable.

For example, for a random walk  $S_n$  that starts from  $a > 0$  and  $b > a$ , the stopping time  $T = \inf\{n : S_n \geq b \text{ or } S_n \leq 0\}$  is regular, because the stopped martingale is bounded.

**Theorem 5.6.2.** Let  $S_n = Y_1 + \dots + Y_n$ , where  $Y_k$  are i.i.d. integrable random variables and  $\mathbb{E}Y_k = 0$ .

Suppose  $T$  is a stopping time which satisfies  $\mathbb{E}(T) < \infty$ . Then,

(i)  $T$  is regular for  $S_n$  assuming  $\mathbb{E}(|Y_1|) < \infty$ .

(ii)  $T$  is regular for  $S_n^2 - n\text{Var}(Y_1)$  assuming  $\mathbb{E}(Y_1^2) < \infty$ .

*Proof.* Since  $\mathbb{E}(T) < \infty$ , hence  $\mathbb{P}(T < \infty) = 1$  and  $S_{T \wedge n} \xrightarrow{a.s.} S_T$ . We will show that the convergence holds also in  $L^1$ , which implies that  $S_{T \wedge n}$  is uniformly integrable.

Note that  $|S_{T \wedge n} - S_T| = 0$  if  $T \leq n$  and

$$|S_{T \wedge n} - S_T| = \left| \sum_{j=n+1}^T Y_j \right|, \text{ if } T > n.$$

Hence,

$$\begin{aligned} |S_{T \wedge n} - S_T| &= \left| \sum_{j=n+1}^{\infty} Y_j \mathbb{1}_{\{j \leq T\}} \right| \\ &\leq \sum_{j=n+1}^{\infty} |Y_j| \mathbb{1}_{\{j \leq T\}} =: \xi_{n+1}. \end{aligned}$$

We claim that  $\mathbb{E}\xi_{n+1} \rightarrow 0$  as  $n \rightarrow \infty$ . Note that if this is true, then the previous inequality immediately gives the desired convergence  $S_{T \wedge n} \rightarrow S_T$  in  $L^1$ .

Clearly  $\xi_{n+1} \rightarrow 0$  almost surely, since the series is zero if  $n > T$ . In addition,  $\xi_{n+1} \leq \xi_1$  for every  $n \geq 0$ . So, it remains to show that  $\xi_1$  is integrable, and the conclusion will follow by the bounded convergence theorem. We have

$$\mathbb{E}\xi_1 = \sum_{j=1}^{\infty} \mathbb{E}(|Y_j| \mathbb{1}_{\{j \leq T\}}).$$

Note that event  $\{j \leq T\}$  is measurable at time  $j - 1$ , since it is known at time  $j - 1$  if  $T < j$  (that is, if the martingale has been stopped).

Therefore, by independence we find that

$$\begin{aligned}\mathbb{E}\xi_1 &= \sum_{j=1}^{\infty} \mathbb{E}|Y_j| \times \mathbb{E}(\mathbb{1}_{\{j \leq T\}}) \\ &= \mathbb{E}|Y_1| \sum_{j=1}^{\infty} \mathbb{E}(\mathbb{1}_{\{j \leq T\}}) = \mathbb{E}|Y_1| \times \mathbb{E}T < \infty.\end{aligned}$$

In order to prove the second claim, we will first show that  $S_{T \wedge n} \rightarrow S_T$  in  $L^2$ .

Indeed, by using independence, we write

$$\begin{aligned}\mathbb{E}(S_{T \wedge n} - S_T)^2 &= \mathbb{E}\left(\sum_{j=n+1}^{\infty} Y_j \mathbb{1}_{\{j \leq T\}}\right)^2 = \sum_{j=n+1}^{\infty} \mathbb{E}(Y_j \mathbb{1}_{\{j \leq T\}})^2 \\ &= \mathbb{E}(Y_1)^2 E(T).\end{aligned}$$

because for  $i < j$ ,

$$\mathbb{E}(Y_i \mathbb{1}_{\{i \leq T\}} Y_j \mathbb{1}_{\{j \leq T\}}) \mathbb{E}(Y_j Y_i \mathbb{1}_{\{i \leq T\}}) = 0,$$

since  $\{i \leq T\}$  is measurable at time  $i - 1$ . □



## 5.7 Applications of Martingales

Applications:

1. Kakutani Theorem and consistency of likelihood ratio test.
2. the Choquet-Deny theorem on bounded harmonic functions for random walks on groups.
3. The double logarithm law for supremum of a random walk.
4. Azuma-Hoeffding inequality
5. Application to math finance
6. Application to optimality in stochastic control
7. Application to filtering

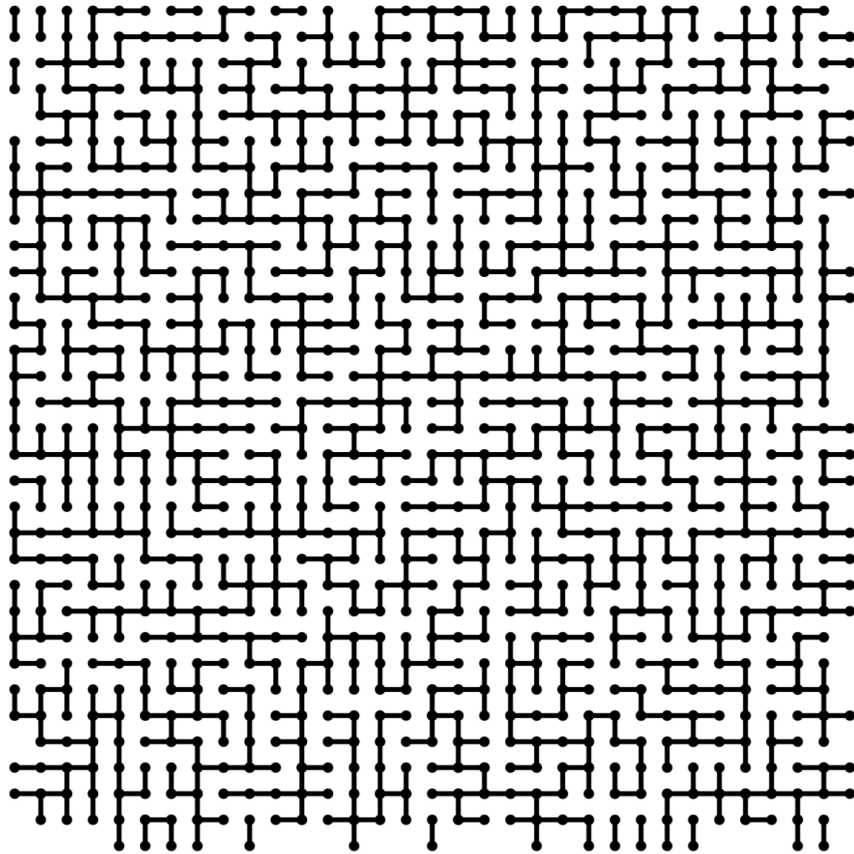



Figure 5.2: A fragment of a random spanning tree for  $\mathbb{Z}^2$

## Chapter 6

# Uniform Spanning Trees

### 6.1 General Results

#### 6.1.1 What is a Uniform Spanning Tree (UST)?

 First of all, recall that a (directed) *graph*  $G = (V, E)$  is a set of vertices  $V$  and a set of edges  $E$ , which are simply a pair of vertices  $E \subset V \times V$ . We will assume that there are no loops, that is, there is no edges  $e = (v, v)$ . A graph  $G_1 = (V_1, E_1)$  is a subgraph of  $G$  if  $V_1 \subset V$  and  $E_1 \subset E$ .

An *undirected graph* is a graph in which edge  $(v_1, v_2)$  is identified with edge  $(v_2, v_1)$ . One can think about this as that these two edges always come together. This is also called a *simple graph* since it does not have loops and multiple edges, that is, there cannot be two distinct edges with the same endpoints. If multiple edges are allowed we have to distinguish between several different edges  $(u, v)$ . Such a graph is called a *multigraph*. Some of our results should also work for multigraphs but as a standing hypothesis it is assumed that all graphs are simple.

A *path* is a sequence of edges  $e_1 = (v_0, v_1), \dots, e_n = (v_{n-1}, v_n)$ . A *circuit* is a path such that all  $v_i$  are distinct except that  $v_n = v_0$ . A graph is *connected* if for every pair of vertices  $u$  and  $v$ , there is a path from  $u$  to  $v$ .

A *tree* is a connected graph such that  $|E| = |V| - 1$ . It can also be defined as a connected graph without circuits. And a *spanning tree*  $T$  of a graph  $G$  is a subgraph of  $G$  that has the same set of vertices as  $G$  and also is a tree.

**Definition 6.1.1.** A *uniform spanning tree* (“UST”)  $T$  of a graph  $G$  is a tree which is chosen uniformly at random from all spanning trees of  $G$ .

#### 6.1.2 Wilson’s algorithm for UST generation

 Wilson’s algorithm is a very efficient method to generate a spanning tree.

Wilson's method applies to a more general situation when one needs to generate a weighted spanning tree on a directed graph associated with a finite Markov chain.

Recall that a *directed graph*  $\vec{G}$  is a pair  $(V, \vec{E})$ , where  $V$  is a finite set of vertices and  $\vec{E}$  is a set of directed edges, that is a subset of  $V \times V$  where the elements  $(v_1, v_2)$  and  $(v_2, v_1)$  are considered different. For an edge  $e = (v_1, v_2)$  the vertex  $v_1$  is called *tail* and denoted  $e^-$  and vertex  $v_2$  is called *head* and denoted  $e^+$ .

For each directed graph  $\vec{G}$  we can associate an undirected graph  $G$  by forgetting the orientation of the edges of  $\vec{G}$ . In this case, however, two edges of  $\vec{G}$  can map to one edge of  $G$ . We can talk about paths and cycles of  $\vec{G}$  as the sequences of edges that map to paths and cycles of  $G$ .

In addition, we have the concept of *directed path*  $e_1, \dots, e_k$ , in which the head of the edge  $e_i$  is the tail of the next edge  $e_{i+1}$ . If  $v_0$  is the tail of the first edge  $e_1$  and  $v_k$  is the head of the last edge  $e_k$  then we say that  $v_0$  and  $v_k$  are connected by a directed path.

Next we define a spanning tree  $T$  of a directed graph  $G$  as a subgraph of  $G$  that includes all vertices in  $V$ , that has a marked vertex, called *root*, and that satisfies a property that every vertex in  $V$  except root is connected to the root by a unique directed path in  $T$ , and this path is directed from the vertex to the root.

In general, not all directed graphs  $G$  have a spanning tree. However, the graphs associated with irreducible Markov chains do have spanning trees.

For a Markov chain on state-space  $V$ , we associate a directed graph  $G$  by assuming that it has a directed edge  $(v_1, v_2)$ , if and only if the probability of one-step transition from  $v_1$  to  $v_2$  is positive,  $P(v_1, v_2) > 0$ .

A usual spanning tree for a graph  $G$  is the directed spanning tree for the simple random walk on graph  $G$  for which we forget the orientation of the edges. Conversely if  $T$  is the spanning tree on  $G$ , then  $\vec{G}$  has  $|V|$  directed spanning trees which are obtained from  $T$  by selecting a vertex  $v$  as a root and choosing those edges in the tree that are directed to the root.

In particular, if we have a method to generate a directed spanning tree uniformly at random, then this gives a method to generate the usual spanning tree uniformly at random.

Wilson's method is a way to generate a directed spanning tree with a specific probability distribution. Namely, let us associate a weight  $w(T)$  to the spanning tree  $T$ , where

$$w(T) = \prod_{e \in T} p(e),$$

and where  $p(e) = p(v_1, v_2)$  is the probability of transition from vertex  $v_1$  to  $v_2$  for every directed edge  $e = (v_1, v_2)$ .

Wilson's method generates trees with probability distribution proportional to these weights. It works by constructing an increasing sequence of trees  $T_i$ ,  $i = 0, \dots, s$ . Choose a random vertex  $r$ . This will be the root of the trees. We

start with  $T_0$  with the vertex set  $V_0 = \{r\}$ . Given a tree  $T_i = (V_i, E_i)$ , which is not spanning, select  $x \in V \setminus V_i$ . Start the Markov chain at  $x$  and stop it when it hits  $T_i$ . Perform a *loop erasure* on the resulting directed path of the edges and add the resulting path to  $T_i$ . The result is declared to be  $T_{i+1}$ .

This algorithm requires explaining what do we mean by “loop erasure” of a directed path. Let  $P = (x_0, x_1, \dots, x_l)$  is a directed path. If  $x_l = x_0$ , then we set the resulting looped erased path is empty. If  $x_l \neq x_0$  then we set  $u_0 = x_0$  and  $u_1 = x_{i+1}$  where  $x_i$  is the last time when the path visited  $x_0$ . Next we check if  $x_l = u_1$ . If yes, we stop and the loop erased path is  $(u_0, u_1)$ . If not then we set  $u_2 = x_{j+1}$ , where  $x_j$  is the last time when the path visited  $u_1$ . We continue this process until  $x_l = u_t$  for some  $t$ . The resulting path  $(u_0, \dots, u_t)$  is directed and has no cycles. We call it the loop erasure of  $P$  and denote  $LE(P)$ .

**Theorem 6.1.2.** *Wilson’s method produces a directed spanning tree  $T$  of the graph associated with Markov chain with probability proportional to  $w(T)$ .*

The proof of this theorem is by a study of a model which is slightly more general than the Markov chain. We call it the *random stacks model*. It is given as a collection of numbers  $S_i^x, i = 1, \dots, \infty$ . These numbers are all independent and  $\mathbb{P}(S_i^x = y) = \mathbb{P}(x, y)$ , where  $\mathbb{P}(x, y)$  is the transition probability of the Markov chain.

Intuitively, every collection  $\{S_i^x\}$  can be thought of as a realization of the Markov chain where the random variable  $S_i^x$  shows which is the next state after the  $i$ -th visit to the state  $x$ .

It is convenient to think about these numbers as organized in stacks lying under the states  $x$ . In this case the Markov chain can be imagined by moving around the state space from  $x$  to whatever is written on the top of the stack  $S^x$  and removing this prescription from the stack.

The top elements of stacks describe a directed graph with edges  $(x, S_1^x)$ . This is a graph in which each vertex has out-degree 1. This graph is called “visible graph”.

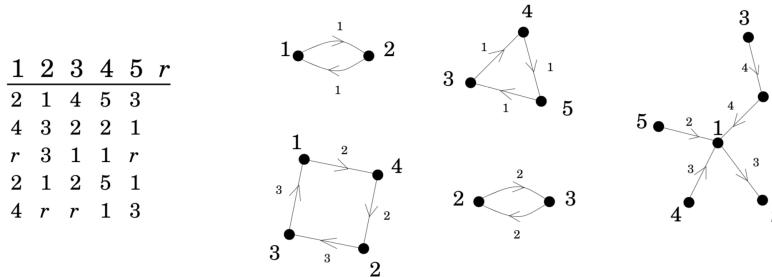
Next we can consider certain operations on the collection of stacks  $\{S^x\}$  which relate two different collections. Every of these operations is a composition of several elementary operations, which are called “popping a cycle”.

*Popping a cycle* is an operation that removes the top elements in stacks  $\{S^x\}$  that form a cycle. Popping a cycle changes the visible graph and it is convenient to distinguish the edges in this graph even if they have the same endpoints. For this reason we say that the edge  $(x, y)$  has color  $i$  if corresponds to the element  $y = S_i^x$  in the original stack configuration.

Note that a sequence of cycle-popping can terminate if there are no cycles to pop. In this case, the visible graph is a directed spanning tree. An example is shown in Figure 6.1.

It is important that if the process of cycle-popping is eventually terminated then the order of cycle-popping does not matter and the resulting spanning tree is always the same. This is the claim of the following lemma.

**Lemma 6.1.3.** *Given any stacks under the states, the order in which cycles are popped is irrelevant in the sense that every order pops an infinite number of*



**Figure 6.1:** An example of cycle popping. On the left is the stacks under the 6 vertices. The stack under the root  $r$  is empty. In the middle is a sequence of popped cycles (shown in the clock-wise order). The numbers on the top of edges show the depth of the edges in stacks. On the right is the resulting spanning tree.

*cycles or every order pops the same (finite set of) colored cycles, thus leaving the same colored spanning tree on top in the latter case.*

*Proof.* Suppose that a colored cycle  $C$  can be popped at some stage of a cycle-popping sequence, that is, that there is a sequence  $C_1, C_2, \dots, C_k = C$  of colored cycles that can be popped, and suppose that  $C' \neq C_1$  is another colored cycle that can be popped at the first stage. Then we claim that either (1)  $C' = C$  or (2) after  $C'$  is popped, there is another cycle-popping sequence such that  $C$  can be popped.

This is enough to prove the lemma. Indeed, if at least one cycle-popping sequence is infinite then for every other sequence there is always an infinite number of cycles that can be popped, so every sequence is infinite. On the other hand, suppose a colored cycle  $C$  is an element of a finite cycle-popping sequence. Then if we consider a particular stage of another sequence, then the claim above shows that either  $C$  is already popped in this sequence, or it can be continued so that  $C$  is popped. Since all sequences are finite in this case, when we conclude that  $C$  will be eventually popped.

So it remains to prove the claim above. Now if all the vertices in  $C'$  are different from vertices in  $C_i$  then obviously,  $C$  can still be popped. Otherwise, let  $C_j$  be the first cycle that has a vertex in common with  $C'$ . Let  $x$  be this vertex and let it be followed by  $y$  in  $C'$ . Since  $C'$  can be popped at the first stage and since it has no vertices in common with  $C_1, \dots, C_{j-1}$  hence  $x$  is still followed by  $y$  in  $C_j$ . By repeating this argument we find that  $C' = C_j$  and either  $C' = C$  or it is possible to pop  $C$  in the order  $C', C_1, \dots, C_{k-1}, C_{k+1}, \dots, C$ .  $\square$

*Proof of Theorem 6.1.2.* We think about Wilson's algorithm as a particular example of cycle popping procedure applied to a realization of random stacks model. By the previous lemma the order of cycle-poppings does not matter. If a particular sequence of cycle-poppings terminates, then all of these sequences

terminate and all them pop exactly the same sequence of cycles. We are interested to know what is the probability that a particular spanning tree will be uncovered as a result of any of these sequences. Clearly, the lemma implies that this probability depends only on the realization of the random-stack model.

Hence, the probability that Wilson's algorithm ends with a spanning tree  $T$  is


$$\sum_O w(O)w(T),$$

where  $w(O) = \prod_{e \in O} p(e)$ , where  $O$  is a collection of cycles in a realization of the random stack model such that popping these cycles will uncover the tree  $T$ . Note that this collection does not depend on  $T$  in the sense that any collection  $O$  is compatible with any spanning tree  $T$ . Hence, this probability is

$$\left( \sum_O w(O) \right) w(T),$$

so it is proportional to  $w(T)$ . □

### 6.1.3 USTs, hitting probabilities, and potentials

 Now let us return to the case of undirected graphs. Consider a random walk on an undirected graph  $G$ . This is a Markov chain with the transition probability  $p(v_1, v_2) = 1/d(v_1)$  where  $(v_1, v_2)$  is an edge of the graph  $G$ , and the vertex  $v_1$  has degree  $d(v_1)$ . Hence, for a given root  $r$ , Wilson's method generates spanning trees with probability proportional to weight

$$w(T) = \prod_{e \in T} \frac{1}{d(e^-)} = \prod_{v \in V \setminus \{r\}} \frac{1}{d(v)}$$

This probability is the same for all trees with the root at  $r$ . Every undirected spanning tree on graph  $G$  corresponds to exactly one directed spanning tree on graph  $\vec{G}$  with a specified root  $r$ . Hence, if after a directed tree is generated by Wilson procedure with a specific  $r$ , we remove the direction of edges, every resulting undirected tree will receive equal probability. This shows that Wilson's method can be used to generate undirected spanning trees of graph  $G$  uniformly at random.

The uniform spanning trees are connected to other models such as domino tilings, perfect matchings in graphs and others and it is important to be able to calculate probabilities that a selected edge or edge belong to the tree. We start with just one edge.

**Theorem 6.1.4** (Uniform spanning tree and hitting probabilities). *Let  $T$  be an unrooted weighted uniform spanning tree of a graph  $G$  and  $e = (e^-, e^+)$  be an edge of  $G$ . Then*

$$\mathbb{P}[e \in T] = P_{e^-}[\text{1st hit } e^+ \text{ via traveling along } e].$$

*Proof.* The first statement follows from Wilson's method applied to graph  $G$  if we set root equal  $e^+$  and start the procedure with  $e^-$ . Indeed, if the first hit of  $e^+$  is via traveling along  $e$ , then the procedure stops at the moment of hit and  $e$  will not be removed during loop erasure. Therefore, it will be a part of the generated tree. Conversely,  $e$  can be part of the tree only if the first hit of  $e^+$  occurred via the edge  $e$ .  $\square$

How do we calculate the probability of the first hit along a specific edge? For this we introduce Green functions.

Recall that we introduced the *hitting time* of a subset  $A$  is the random variable

$$H^A = \inf\{n \geq 0 : X_n \in A\}.$$

Define the *Green function*  $G_A(a, x)$  for a Markov chain absorbed at  $A$  as the expected number of visits to  $x$  strictly before hitting  $A$  by a random walk started at  $a$ . If

$$N_A(x) = \sum_{n=0}^{\infty} \mathbb{1}[X_n = x, n < H^A]$$

denotes the number of visits to  $x$  by the Markov chain  $X_i$  before hitting  $A$ , then

$$G_A(a, x) = \mathbb{E}[N_A(x) | X_0 = a].$$

**Theorem 6.1.5.** *Let  $G$  be a finite connected graph. Start a random walk at  $a$  and absorb it when it first visits a set  $A$ . For an edge  $e = (x, y)$ , let  $S_{xy}$  be the number of transitions from  $x$  to  $y$ . Let  $x \notin A$ . Then,*

$$\mathbb{E}[S_{xy}] = G_A(a, x)/d_x.$$

*Proof.*  $S_{xy}$  is the number of visits to  $x$  followed by a transition to  $y$ . Therefore,  $\mathbb{E}[S_{xy}] = G_A(a, x)/d_x$  by the definition of  $G_A(a, x)$  and the fact that the transition probability from  $x$  to  $y$  equals  $d_x$ .  $\square$

**Theorem 6.1.6 (Kirchhoff's Formula).** *Let  $T$  be an unrooted weighted uniform spanning tree of a graph  $G$  and  $e = (e^-, e^+)$  be an edge of  $G$ . Then,*

$$\mathbb{P}[e \in T] = G_{e^+}(e^-, e^-)/d_{e^-},$$

*Proof.* By Theorem 6.1.4,

$$\mathbb{P}[e \in T] = P_{e^-}[\text{1st hit } e^+ \text{ via traveling along } e]. \quad (6.1)$$

Consider now the random walk on  $G$ , started at  $e^-$  and stopped at  $e^+$ . Then  $S_{e^-, e^+} = 1$  if the first hit of  $e^+$  occurred via  $e$  and zero otherwise. And  $S_{e^+, e^-} = 0$  always. Hence, the probability on the right-hand side of (6.1) equals  $G_{e^+}(e^-, x)/d_x$  with  $x = e^-$  by Theorem 6.1.5.  $\square$



Note that this formula gives an answer how to calculate the probability that an edge belongs to a tree provided that we can calculate the Green functions for a given source and boundary.

Obviously  $G_A(a, x) = 0$  if  $x \in A$ . What can be said about the Green function outside of  $a$  and  $A$ ? First of all, if the chain is reversible, then the Green function is in a certain sense symmetric.

**Lemma 6.1.7.** *If the Markov chain  $X_n$  is reversible with the stationary distribution  $\pi(x)$ , then*

$$\pi(a)G_A(a, x) = \pi(x)G_A(x, a).$$

*Proof.* By the definition of reversibility  $P(x, y) = \pi(x)P_{xy} = \pi(y)P(yx) = P(y, x)$ . For every path  $(x_0, x_1, \dots, x_n)$ , this property can be extended to the property

$$P(x_0, x_1, \dots, x_n) = P(x_n, x_{n-1}, \dots, x_0),$$

where  $P(x_0, x_1, \dots, x_n) := \pi(x_0)P_{x_0x_1} \dots P_{x_{n-1}x_n}$ , is the joint probability mass function for a chain  $P$  started in invariant distribution  $\pi$ . Hence, for every  $n$ ,

$$\begin{aligned} \pi(a)\mathbb{P}[X_n = x, n < H^A | X_0 = a] &= \sum_{x_1, \dots, x_{n-1} \notin A} \pi(a)\mathbb{P}[X_1 = x_1, \dots, X_n = x | X_0 = a] \\ &= \sum_{x_1, \dots, x_{n-1} \notin A} P(a, x_1, \dots, x_{n-1}, x) \\ &= \sum_{x_1, \dots, x_{n-1} \notin A} P(x, x_{n-1}, \dots, x_1, a) \\ &= \pi(x)\mathbb{P}[X_n = a, n < H^A | X_0 = x]. \end{aligned}$$

Since  $G_A(a, x) = \sum_{n=0}^{\infty} \mathbb{P}[X_n = x, n < H^A | X_0 = a]$ , after summing the identity in the previous display over all  $n$ , we obtain the conclusion of the lemma.  $\square$

**Theorem 6.1.8.** *If the Markov chain  $X_n$  is irreducible and reversible with the stationary distribution  $\pi(x) > 0$  for all  $x$ , then the function  $v(x) = G_A(a, x)/\pi(x)$  is harmonic everywhere outside  $a \cup A$ . That is,*

$$v(x) = \sum_y P_{xy}v(y),$$

for  $x \notin a \cup A$ . In addition,

$$v(a) - \sum_y P(a, y)v(y) = \frac{1}{\pi(a)}.$$

*Proof.* By using lemma, we only need to show that the function  $G_A(x, a)$  is harmonic over  $x$  if  $x \notin a \cup A$ . However, this is clear from the formula

$$\begin{aligned} G_A(x, a) &:= \mathbb{E}[N_A(a)|X_0 = x] = \sum_y P(x, y) \mathbb{E}[\tilde{N}_A(a)|X_1 = y] \\ &= \sum_y P(x, y) \mathbb{E}[N_A(a)|X_0 = y] \\ &= \sum_y P(x, y) G_A(y, a), \end{aligned}$$

where  $\tilde{N}_A(a) := \sum_{n=1}^{\infty} \mathbb{1}[X_n = a, n < H^A]$ , valid for all  $x \notin a \cup A$ .  
For  $x = a$ , we have

$$\mathbb{E}[N_A(a)|X_0 = a] = 1 + \sum_y P(a, y) \mathbb{E}[\tilde{N}_A(a)|X_1 = y],$$

which means that

$$\frac{1}{\pi(a)} G_A(a, a) = \frac{1}{\pi(a)} + \sum_y P(a, y) \frac{1}{\pi(a)} G_A(y, a),$$

or

$$v(a) - \sum_y P(a, y) v(y) = \frac{1}{\pi(a)}.$$

□

In our case, we are concerned with random walks on graphs where the transition probabilities  $P_{ij} = 1/d_i$ , where  $d_i$  is the degree of vertex  $i$ . Hence we will call a function  $f$  on vertices of graph  $G$  *harmonic* at vertex  $x$ , if

$$f(x) = \frac{1}{d_x} \sum_{y \sim x} f(y),$$

where the sum is over all  $y$  adjacent to  $x$ .

Theorems 6.1.6 and 6.1.8 gives us tools to calculate the probabilities that a given edge belongs to a uniform spanning tree.

In order to generalize this to many edges, we connect these results with the theory of electric networks. However, for simplicity we will consider only the case when all edge conductances are equal 1. This will be enough for our purposes.

**Definition 6.1.9.** Let  $a$  and  $A$  be a vertex and a subset of vertices of a graph  $G$ , then a *voltage function* is a function on the vertices of the graph that is harmonic at all  $x \notin a \cup A$ .

Another name for the voltage function is *potential*, and one can talk about potential theory for Markov chains. The voltage should be thought of as a harmonic function on vertices of the graph that has exceptions at some vertices where it is not harmonic. We will usually think about  $a$  as a point with non-zero voltage through which the electric current flows in (so it is a “source”) and  $A$  as the set with zero voltage (“ground” or a “sink”). However, this will not play a role in calculations and purely for convenience.

**Definition 6.1.10.** Given a voltage function  $v$ , we define the associated *current function*  $i$  on the pairs of adjacent vertices by

$$i(x, y) := v(x) - v(y).$$

Notice that current function is antisymmetric  $i(x, y) = -i(y, x)$  and whenever  $v$  is harmonic at a vertex  $x$ , we have

$$\sum_{y \sim x} i(x, y) = 0,$$

where  $y \sim x$  means that vertex  $y$  is adjacent to vertex  $x$ .

The definition of voltage function implies that a current function is a flow on  $V \setminus \{a \cup A\}$ , where we use the following important definition.

**Definition 6.1.11.** A function  $f$  on ordered pairs of adjacent vertices in graph  $G$  is called a *flow* between  $a$  and  $A$  if  $f(x, y) = -f(y, x)$  for all neighbors  $x, y$  and  $\sum_{y \sim x} f(x, y) = 0$  for all  $x$  not in  $a \cup A$ .

A flow between  $a$  and  $A$  is called a *unit flow* (out of  $a$ ) if  $\sum_{y \sim a} f(a, y) = 1$ .

**Theorem 6.1.12** (Voltage as expected number of visits). *Let  $G$  be a finite connected graph,  $d_x$  denote the degree of vertex  $x$ , and  $G_A(a, x)$  be the Green function for a walk started at  $a$  and stopped at  $A$ . Then, the formula  $v(x) = G_A(a, x)/d_x$  defines a valid voltage function on  $G$ , which is harmonic everywhere outside  $a \cup A$  with the voltage equal to 0 on  $A$ . It corresponds to a unit current flow between  $a$  and  $A$ .*

*Proof.* Since the stationary distribution  $\pi(x)$  of a simple random walk is proportional to  $d_x$ ,  $\pi(x) = cd_x$ , the claims that  $v(x)$  is harmonic outside of  $a \cup A$  and that  $v(x) = 0$  for  $x \in A$  have been already proved in Theorem 6.1.8. In addition, the same theorem implies that

$$v(a) = \frac{1}{d_a} + \frac{1}{d_a} \sum_{y \sim a} v(y),$$

which we can rewrite as

$$1 = \sum_{y \sim a} (v(a) - v(y)).$$

This shows that  $v$  corresponds to the unit current flow. □

The Kirchhoff formula in Theorem 6.1.8 can be reformulated as a statement about current flows. Namely, for an edge  $e = (e^-, e^+)$ ,

$$\mathbb{P}(e \in T) = i_e(e^-, e^+),$$

where  $i_e$  is the unit current flow from  $e^-$  to  $e^+$ .

More generally, we are interested in a probability  $\mathbb{P}[e_1, \dots, e_k \in T]$  that a specific set of edges  $e_1, \dots, e_k$  belong to a uniform spanning tree.

In order to formulate the result, let us choose an arbitrary orientation on the graph  $G$ , so that we can talk about the start point  $e^-$  and the end point  $e^+$  of an edge. Then  $i_e$  denote the unit current flow from  $e^-$  to  $e^+$ . Specifically, for two adjacent edges  $x$  and  $y$ , define

$$i_e(x, y) = \frac{1}{d_{e^-}} [G_{e^+}(e^-, x) - G_{e^+}(e^-, y)],$$

where  $G$  is the corresponding Green function. Then for two arbitrary edges we define the following function:

$$\mathcal{Y}(e, f) := i_e(f) = i_e(f^-, f^+).$$

**Theorem 6.1.13** (Burton-Pemantle). *Let  $T$  be a uniform spanning tree of graph  $G$ . For any distinct edges  $e_1, \dots, e_k \in G$ ,*

$$\mathbb{P}[e_1, \dots, e_k \in T] = \det \mathcal{Y}(e_i, e_j) \Big|_{i=1, \dots, k; j=1, \dots, k}$$

The case  $k = 1$  of this theorem has been just established. Then we plan to proceed by induction by using the identity:

$$\mathbb{P}[e_1, \dots, e_k \in T] = \mathbb{P}[e_k \in T | e_1, \dots, e_{k-1} \in T] \mathbb{P}[e_1, \dots, e_{k-1} \in T].$$

The proof of the Burton-Pemantle is based on an interpretation of the conditional probability

$$\mathbb{P}(e \in T | e_1, \dots, e_k \in T)$$

in terms of the UST of the graph  $G$  contracted along edges  $e_1, \dots, e_k$ .

A contraction of a graph  $G$  along a set of edges is a different graph  $G'$  in which all endpoints of the edges are replaced with a single vertex and all the resulting loops are removed. In this form, however, the resulting graph can become a multigraph (without loops), in which two vertices can be connected by several different edges.

**Lemma 6.1.14.** *Suppose  $T$  is a spanning tree in  $G$  and  $e_1, \dots, e_k \in T$ . Then  $T' = T / \{e_1, \dots, e_k\}$  is a spanning tree in multigraph  $G' = G / \{e_1, \dots, e_k\}$ .*

*Proof.* A contraction of an edge in  $T$  cannot create a cycle in  $T'$ , so  $T'$  is still a tree and it is also obvious that it is a spanning tree.  $\square$

**Lemma 6.1.15.** *Let  $e_1, \dots, e_k$  is a set of edges in  $G$  and  $T$  is a uniform random spanning tree in  $G$ . If  $e_1, \dots, e_k \in T$ , then  $T' = T/\{e_1, \dots, e_k\}$  is the uniform spanning tree in multigraph  $G' = G/\{e_1, \dots, e_k\}$ .*

*Proof.* This holds because there is a bijection between spanning trees in  $G$  that contain edges  $e_1, \dots, e_k$  and all uniform spanning trees in  $G$ . If  $T$  is a uniform (random) spanning tree and  $t_1, t_2$  are two spanning trees that both contain  $e_1, \dots, e_k$ , then  $\mathbb{P}(T = t_1) = \mathbb{P}(T = t_2)$  by definition of uniform distribution. The bijectivity of the map  $T \rightarrow T' = T/\{e_1, \dots, e_k\}$  implies that  $\mathbb{P}(T' = t'_1) = \mathbb{P}(T' = t'_2)$ .  $\square$

It follows that if  $T$  and  $T'$  denote the corresponding uniform spanning trees in  $G$  and  $G' = G/\{e_1, \dots, e_k\}$ , respectively, then

$$\mathbb{P}(e \in T | e_1, \dots, e_k \in T) = \mathbb{P}(e \in T'), \quad (6.2)$$

We assume here without proof our previous theory can be applied to multigraph  $G'$ . (In particular, we assume that a suitable generalization of Wilson's method generate a uniform spanning tree in multigraph.)


In particular, we assume that

$$\mathbb{P}(e \in T | e_1, \dots, e_k \in T) = \mathbb{P}(e \in T') = i_e^{G'}(e), \quad (6.3)$$

where  $i_e^{G'}$  is the unit current flow from  $e^-$  to  $e^+$  in multigraph  $G'$ .

In order to use this idea we need to understand how the contraction of edges influences the current flows. For this purpose, we will study the space of current flows more attentively in the next section. The main idea will be to represent  $i_e^{G'}$  as an orthogonal projection of  $i_e$  on the space of current flows orthogonal to currents  $i_{e_1}, i_{e_2}, \dots, i_{e_k}$ .

#### 6.1.4 Voltages, Currents, and Projections

 Our first goal will be establishing that the matrix  $\mathcal{Y}(e_i, e_j) = i_{e_i}(e_j)$  in the Burton-Pemantle theorem can be represented as the matrix of a quadratic form  $(P_X e_i^*, e_j^*)$  for certain functions  $e_i^*$  and  $e_j^*$  and a certain orthogonal projection operator  $P_X$ .

First, we need to understand what are distinct features of current flows.

(What is going on in this section is that we are building a discrete version of certain concepts from differential geometry. Current flows correspond to differentials of functions and general flows to differential forms. We are going to build a decomposition of the space of forms as a direct sum of the space of differentials and its orthogonal complement.)

In the following theorem we use directed edges and define the current function  $i$  on a directed edge  $e = (x, y)$  as  $i(e) = v(x) - v(y)$ , where  $v$  is a voltage function.

**Theorem 6.1.16** (Kirchhoff's cycle law). *Let  $e_1 = (v_0, v_1)$ ,  $e_2 = (v_1, v_2)$ , ...,  $e_n = (v_{n-1}, v_n = v_0)$  is a directed cycle in  $V$  and  $i$  is a current flow. Then*

$$\sum_{k=1}^n i(e_k) = 0.$$

This follows from the definition of the current function as the difference of the voltage on endpoints. In fact this property holds for any function on directed edges defined as difference of a fixed function on edge endpoints. Note, however, that this property does not always hold for arbitrary flows. (For example, it does not hold for a flow over the cycle graph that gives value 1 to every counter-clockwise directed edge.) In fact, this property is in a certain sense characterizes the current flows.

**Lemma 6.1.17.** *Suppose that an antisymmetric function  $j$  (meaning that  $j(x, y) = -j(y, x)$ ) on the directed edges of a finite connected graph satisfies Kirchhoff's cycle law. Then there exists a function  $F(x)$  on vertices of the graph such that  $j(x, y) = F(x) - F(y)$ . This function is unique up to an additive constant. Moreover, the function  $F$  is harmonic at every point  $x$ , where the following star condition is satisfied:*

$$\sum_{e:e^-=x} j(e) = 0.$$

*Proof.* Fix a value  $F(x_0)$  on some particular vertex. Then we can define the value  $F(y)$  on any vertex by finding a path  $e_1, \dots, e_n$  from  $x_0$  to  $y$  and defining  $F(y) = F(x_0) - \sum_{k=1}^n j(e_k)$ . The fact that this definition does not depend on the choice of path follows from the cycle law. The uniqueness can be shown by induction. The final statement follows because

$$\frac{1}{d_x} \sum_{y \sim x} F(y) = F(x) - \frac{1}{d_x} \sum_{e:e^-=x} j(e).$$

□

In order to discuss current flows it is convenient to introduce a suitable linear space with a scalar product.

First of all, for a graph  $G = (V, E)$ , let us define  $\vec{E}$ , the set of directed edges of  $G$ . Namely, if  $e = (v_1, v_2) \in E$ , then there are two corresponding elements ("directed edges") in  $\vec{E}$ .

It is convenient to identify vertices of  $G$  with integers  $1, \dots, |V|$ , the edges in  $E$  with ordered pairs  $(v_1, v_2)$ ,  $v_1 < v_2$  and the edges in  $\vec{E}$  with all possible pairs  $(v_1, v_2)$ ,  $v_1 \neq v_2$ . For an edge  $e = (v_1, v_2) \in \vec{E}$ , we define  $-e = (v_2, v_1)$ . For the edge  $e$ , we also define  $e^- = v_1$  and  $e^+ = v_2$ .

Next we introduce an  $|E|$ -dimensional linear space of all antisymmetric functions on  $\vec{E}$ , that is, all functions  $s : \vec{E} \rightarrow \mathbb{R}$ , such that  $s(-e) = -s(e)$ . These functions are meant to represent flows over the graph edges, although at this

stage we do not impose a restriction that a total flow at a non-source vertex is zero. Intuitively, if the directed edge is  $e = (v_1, v_2)$ , then  $s(e)$  represents the amount of flow from  $v_1$  to  $v_2$ .

We introduce a scalar product on this space of flows as

$$(s, t) = \frac{1}{2} \sum_{e \in \vec{E}} s(e)t(e) = \sum_{e \in E} s(e)t(e).$$

We denote this space  $l^2(E)$ .

For an element  $e \in \vec{E}$ , we will use  $e^*$  to denote the function  $e^*(\cdot) = \mathbb{1}_e(\cdot) - \mathbb{1}_{-e}(\cdot) \in l^2(E)$ . Here,  $\mathbb{1}_e$  is a function which maps  $e$  to 1 and all other elements of  $\vec{E}$  to 0. (The function  $\mathbb{1}_e$  is not antisymmetric and  $\notin l^2(E)$ .) In particular,  $e^*(e) = 1$  and  $e^*(-e) = -1$ .

A convenient orthonormal basis in the space  $l^2(E)$  is given by functions  $e_i^*$ , where  $e_i$  are all edges in  $E$ .

Next, we define the linear subspace  $X \subset l^2(E)$  spanned by “star” functions

$$f_v = \sum_{e \in \vec{E}: e^- = v} e^* \in l^2(E).$$

Note that if a function  $g \in l^2(E)$  is orthogonal to  $X$ , then for every vertex  $v$  in the graph, we have

$$\sum_{e \in \vec{E}: e^- = v} g(e) = 0.$$

One can interpret this by saying that the flow represented by  $g$  has the property that the net amount that passes vertex  $v$  is zero.

Let  $P_X$  be the orthogonal projection in  $l^2(E)$  on the subspace  $X$  and define the symmetric bilinear form,

$$\mathcal{Y}(f, g) = (P_X f, g).$$

In particular, if  $e_i$  and  $e_j$  are two edges in  $E$ , then we define  $\mathcal{Y}(e_i, e_j) = \mathcal{Y}(e_i^*, e_j^*)$ . At this moment, it is not clear if the form defined in this way coincides with the form  $\mathcal{Y}$  we defined above in terms of unit current flows. The fact that these two definitions are in agreement will be justified in Lemma 6.1.22 below.

Let us also define the subspace  $Y \subset l^2(E)$  spanned by “cycle” functions,

$$g_C = \sum_{e \in C} e^* \in l^2(E),$$

where  $C$  is a directed cycle in  $\vec{G}$ . Note that every  $h \in Y^\perp$  satisfies the Kirchoff’s cycle law and by Lemma 6.1.17 can be written as  $h(e) = F(e^-) - F(e^+)$  for a function  $F$  determined up to a constant.

**Theorem 6.1.18.** *For a finite connected graph  $G$ , we have the orthogonal decomposition  $l^2(E) = X + Y$ .*

*Proof.* Every cycle function is orthogonal to every star function. This holds by an observation that the number of times a directed cycle  $C$  enters a given point  $v$  equals to the number of time that it exits this points. The first number gives the number of  $-1$ s in the calculation of the the scalar product and the second number is the number of  $+1$ s.

Hence  $X$  and  $Y$  are orthogonal. Now suppose that function  $j \in l^2(E)$  is orthogonal to both  $X$  and  $Y$ . The assumption  $j \in Y^\perp$  implies that  $j(e) = F(e^-) - F(e^+)$  for some function  $F \in l^2(V)$ . Indeed, fix a value of  $F(x)$  on some vertex of the graph and successively define the values of  $F$  on other vertices of the graph by using  $F(y) = F(x) - j(x, y)$ . This will never lead to a contradiction by the assumption that  $j$  is in  $Y^\perp$ . (If a contradiction is obtained, then this would produce a cycle function which is not orthogonal to  $j$ .)

By the assumption that  $j \in X^\perp$ , it follows that for this function  $F(x)$  we have:

$$\sum_{y \sim x} F(y) = d_x F(x) - \sum_{e: e^- = x} j(e) = d_x F(x),$$

where  $d_x$  is the degree of  $x$ . Hence,  $F(x)$  is a function which is harmonic everywhere on a finite connected graph  $G$ . It is known that all such functions are constants, which implies that  $j = 0$ .  $\square$

Let a vertex  $x$  be called a *source* of a function  $j \in l^2(E)$  if  $\sum_{e: e^- = x} j(e) \neq 0$ , with the *outflow* of the source equal to the sum. Also, let a function  $j \in l^2(E)$  be called *sourceless* if it has no sources. By Theorem 6.1.18 we see that  $Y = X^\perp$  is the subspace of sourceless functions.

*Ex.* 6.1.19. What are sources of a star function  $f_v$ ? What are sources of a cycle function  $f_C$ ?

Intuitively,  $X = Y^\perp$  is the space of current flows and  $Y = X^\perp$  is the space of flows with no sources.

**Lemma 6.1.20.** *Let  $P_X$  denote the orthogonal projection on subspace  $X$  and let  $j \in l^2(E)$ . Then  $P_X(j)$  has the same set of sources as  $j$  and the outflow at each source of  $j$  equal to the outflow at the corresponding source of  $P_X(j)$ . In addition,  $P_X(j)$  is a current flow for the voltage function which is harmonic outside of the set of sources of  $j$ .*

*Proof.* The statement about sources is clear because the difference  $j - P_X(j)$  is in  $Y$  and therefore it is sourceless. The second statement follows because of Lemma 6.1.17.  $\square$

**Lemma 6.1.21.** *Let  $e = (e^-, e^+)$  be a directed edge in  $\vec{E}$ , and  $e^*$  is the associated asymmetric indicator function, that is  $e^*(e) = 1$ ,  $e^*(-e) = -1$ , and  $e^*(f) = 0$ , for all other edges  $f$ . Then,*

$$P_X e^* = i_e,$$

where  $i_e$  is the unit current flow for the voltage function  $v(x)$  which is harmonic on  $V \setminus \{a \cup A\}$  with  $a = e^-$ ,  $A = \{e^+\}$ , and  $v(e^+) = 0$ .



*Proof.* The function  $j_e := P_X e^* \in X = Y^\perp$ . Hence, it satisfies the Kirchhoff's cycle law and by Lemma 6.1.17 can be written as  $j_e(f) = F(f^-) - F(f^+)$  for some function  $F$ . By Lemma 6.1.20, it has the same set of sources as  $e^*$ , that is, only  $e^-$  and  $e^+$ , with the outflows equal to 1 and  $-1$ , respectively. By Lemma 6.1.17, the function  $F$  is harmonic everywhere outside of these two vertices. These conditions determine the unit current flow  $i_e$ . Hence,  $P_X e^* = i_e$ .  $\square$

Now let us reformulate the Kirchhoff's formula in Theorem 6.1.6 for the probability that an edge is in the UST. We will use the operator  $P_X$ , which is the orthogonal projection in  $l^2(E)$  on the subspace  $X$ , spanned by star functions, and the symmetric bilinear form on functions in  $l^2(E)$ ,

$$\mathcal{Y}(f, g) = (P_X f, g).$$

The following lemma shows that our two definitions of  $\mathcal{Y}$  are in agreement.

**Lemma 6.1.22.** *For arbitrary two edges  $e_i, e_j$  in a connected graph  $G$ ,*

$$Y(e_i^*, e_j^*) = i_{e_i}(e_j),$$

where  $i_e$  denotes the unit current flow from  $a = e^-$  to  $A = \{e^+\}$ .

The claim of the lemma is a direct consequence of the definition of  $Y(e_i, e_j)$  as  $(P_X e_i^*, e_j^*)$  and Lemma 6.1.21.

In particular, this lemma and Theorem 6.1.6 imply that the basic case of Theorem 6.1.13 is valid.

**Corollary 6.1.23.** *Let  $T$  be an unrooted weighted uniform spanning tree of a graph  $G$  and  $e = (e^-, e^+)$  be an edge of  $G$ . Then,  $\mathbb{P}[e \in T] = \mathcal{Y}(e, e)$ .*

## 6.1.5 Proof of the Burton-Pemantle theorem



We need to understand how the current flows in the contracted graph are related to current flows in the original graph.

Recall that for any directed edge  $e = (e^-, e^+)$ , the function  $i_e \in l^2(E)$  is defined as the unit flow from  $e^-$  to  $e^+$  with the voltage function which is harmonic outside of  $\{e^-, e^+\}$ , and  $v(e^-) = 1$ ,  $v(e^+) = 0$ . These current functions satisfy the Kirchhoff's cycle law (as differences of potential functions) and therefore are in  $X = Y^\perp$ .

**Lemma 6.1.24.** *Let  $F = \{f_1, \dots, f_k\}$  be the set of edges in graph  $G$ , and  $L$  be the subspace spanned by functions  $i_{f_1}, \dots, i_{f_k}$  in  $X$ . Let  $P_{L^\perp}$  denote the orthogonal projection in  $l^2(E)$  on the ortho-complement of  $L$  in  $X$ . Suppose that edge  $e \notin F$ , and define*

$$i_{e,F} := P_{L^\perp}(i_e).$$

Then  $i_{e,F}$  is a current flow that has no sources outside of the set  $\{e^-, e^+\} \cup Z$ , where  $Z$  is the set of endpoints of  $f_j$ . In addition, it satisfy the property  $i_{e,F}(f_j) = 0$  for every  $f_j \in F$ .

*Proof.* The orthogonal projection along  $L$ ,  $P_{L^\perp}$ , subtracts a linear combinations of functions  $i_{f_j}$ . Since a linear combination of these current flows has no sources in  $V \setminus (\{e^-, e^+\} \cup Z)$ , hence  $P_{L^\perp}(i_e)$  is a current flow with no sources outside of  $\{e^-, e^+\} \cup Z$ .

Let us check the property  $i_{e,F}(f) = 0$  for every  $f \in F$ . Since  $i_{e,F} \in X$ , we have

$$\begin{aligned} i_{e,F}(f) &= (i_{e,F}, f^*) \\ &= (P_X i_{e,F}, f^*) = (i_{e,F}, P_X f^*) \\ &= (i_{e,F}, i_f) = (P_{L^\perp}(i_e), i_f) = (i_e, P_{L^\perp}(i_f)) \\ &= 0. \end{aligned}$$

The first line is by definition of the scalar product, the second and third lines use the self-adjointness of the orthogonal projections  $P_X$  and  $P_{L^\perp}$ , and Lemma 6.1.21.  $\square$

Now consider the restriction of  $i_{e,F}$  to the edges the graph  $G' = G/F$ , which is the graph  $G$  contracted along all the edges in  $F$ . Let us denote this restriction as  $\bar{i}_{e,F}$ .

**Lemma 6.1.25.** *Let the assumptions of the previous lemma hold. In addition, suppose that for the edge  $e \notin F$ , no circuit can be formed by edges in  $e \cup F$ . Then, the flow  $\bar{i}_{e,F}$  is the unit current flow from  $e_-$  to  $e_+$  in  $G' = G/F$ .*

*Proof.* If there is no circuit formed by edges in  $e \cup F$ , then the edge  $e$  is still present in the graph  $G' = G/F$ . The previous lemma implies that  $i_{e,F}$  is a current flow in  $G'$ . (We use the same potential on the vertices of  $G'$ ). It remains to show that this is a *unit* current flow out of  $e^-$  in the contracted graph.

Indeed,  $i_{e,F}$  is a linear combination of  $i_e$  and  $i_{f_i}$ . The flows  $i_{f_i}$  have the same sources as  $f_i^*$ . If the edges  $f_i$  are not incident to the vertex  $e^-$  then it is clear that the flow of  $i_{e,F}$  out of  $e^-$  is the same as the flow of  $i_e$  out of  $e^-$  and therefore equals 1.

Otherwise, if  $e^-$  is incident to an edge in  $\{f_i\}$ , then let  $Z$  be the endpoints of the edges  $\{f_i\}$  and  $e$ , and let  $C$  be a connected component of the subgraph induced by  $Z$  that includes vertex  $e^-$ . After contraction this will be represented by a single vertex  $(e^-)'$  in the graph  $G'$  and therefore we need to show that the total outflow of  $i_{e,F}$  out of  $C$  equals 1.

Indeed, the total outflow of  $i_{f_i}$  out of  $C$  is equal to the total outflow of  $f_i^*$  out of  $C$  and hence equals 0. This implies that total outflow of  $i_{e,F}$  out of  $C$  equals the total outflow of  $i_e$  out of  $C$ . This equals total outflow of  $e^*$  out of  $C$ , and this equals 1 because  $C$  cannot include both  $e^-$  and  $e^+$ . (This is by assumption that the edges in  $e \cup \{f_i\}$  cannot form a cycle.)  $\square$

It follows from formula (6.3) and Lemma 6.1.25, that

$$\mathbb{P}(e \in T | e_1, \dots, e_k \in T) = i_{e,F}(e) = P_{L^\perp}[i_e](e),$$

where  $L$  is the subspace spanned by functions  $i_{e_1}, \dots, i_{e_k}$ . This is the basis for the proof of Theorem 6.1.13.

*Proof of Theorem 6.1.13.* First of all, note that it is enough to consider the case then there is no cycle in the subgraph formed by edges  $e_1, \dots, e_k$ . Indeed if we can find such a cycle then the probability is zero. On the other hand, then we have  $\sum_i e_i^* = 0$  for the edges in the cycle, which implies that  $\sum_i (P_X e_s^*, e_i^*) = 0$  for all  $s = 1, \dots, k$ , which means that sum of several columns in the matrix  $Y(e_i, e_j)$  is zero. This means that the determinant is zero.

Now, we will prove the theorem by induction. The base case  $k = 1$  has been already established. So suppose the theorem was proved for  $k$  and let us prove it for  $k + 1$ .

Let  $Y_k$  denotes the matrix  $(Y(e_i, e_j))$  for  $i, j = 1, \dots, k$ . Then we need to show that

$$\frac{\det Y_{k+1}}{\det Y_k} = \mathbb{P}(e_{k+1} \in T | e_1 \in T, \dots, e_k \in T) = P_{L^\perp}[i_{e_{k+1}}](e_{k+1}),$$

where  $L$  is the subspace spanned by functions  $i_{e_1}, \dots, i_{e_k}$ .

The projection operator  $P_{L^\perp}$  is the projection along the subspace  $L$  and it acts on functions in  $X$  by the rule:  $f \rightarrow f - \sum_{j=1}^k a_j(f) i_{e_j}$ , where the  $a_j(f)$  are certain real coefficient that may depend on function  $f$ . (It is not difficult to write a formula for these coefficients. However, it is not really needed here.) Let  $\alpha_j = a_j(e_{k+1})$ .

The  $k + 1$  column of matrix  $Y_{k+1}$  has elements  $(i_{e_1}, i_{e_{k+1}}), \dots, (i_{e_k}, i_{e_{k+1}}), (i_{e_{k+1}}, i_{e_{k+1}})$ . We can subtract from this column a linear combination of the first  $k$  columns with coefficients  $\alpha_j$ . Then the determinant will not change and the elements of the last column will become  $(i_{e_1}, P_{L^\perp} i_{e_{k+1}}), \dots, (i_{e_k}, P_{L^\perp} i_{e_{k+1}}), (i_{e_{k+1}}, P_{L^\perp} i_{e_{k+1}})$ .

Since  $P_{L^\perp}$  is the orthogonal projection on the space orthogonal to the linear span of  $i_{e_j}$ , the entries  $(i_{e_j}, P_{L^\perp} i_{e_{k+1}})$  are all zero for  $j = 1, \dots, k$ . And the entry


$$\begin{aligned} (i_{e_{k+1}}, P_{L^\perp} i_{e_{k+1}}) &= (e_{k+1}, P_X P_{L^\perp} i_{e_{k+1}}) \\ &= (e_{k+1}, P_{L^\perp} i_{e_{k+1}}) = P_{L^\perp}[i_{e_{k+1}}](e_{k+1}). \end{aligned}$$

By expanding the determinant of  $k + 1$ -by- $k + 1$  modified matrix by the entries in the last column we find that

$$\det Y_{k+1} = P_{L^\perp}[i_{e_{k+1}}](e_{k+1}) \det Y_k,$$

and this is exactly what we wanted to prove.  $\square$

## 6.2 Square Lattice

The results for the previous graph can be applied to infinite graphs as well, provided a measure on spanning trees in infinite graphs is well-defined. For a square lattice graphs in  $\mathbb{Z}^d$  this can be done by considering a sequence of larger and larger boxes.

Can we say something about local statistics in these infinite graphs? For example, what is the distribution of vertex degrees in a random spanning tree? What is the probability that a specific edge belongs to the spanning tree?

In this section we will consider as an example the case of the graph  $\mathbb{Z}^2$ .

### The voltage function on $\mathbb{Z}^2$

Let  $\mathbb{T}$  denote the torus  $\mathbb{R}/\mathbb{Z}^2$ . For every  $k \in \mathbb{Z}^2$  and  $\alpha \in \mathbb{T}$ , define the character function

$$\chi_k(\alpha) = e^{-2\pi i k \cdot \alpha}.$$

For every  $\alpha = (\alpha_1, \alpha_2) \in \mathbb{T}$ , define also

$$\varphi(\alpha) = 4 - 2(\cos 2\pi\alpha_1 + \cos 2\pi\alpha_2).$$

**Theorem 6.2.1** (Voltage on  $\mathbb{Z}^2$ ). *The voltage at  $u$  when a unit current flows from  $x$  to  $y$  in  $\mathbb{Z}^2$  and when  $v(y) = 0$  is*

$$v(u) = f(u) - f(y),$$

where

$$f(u) = \int_{\mathbb{T}^2} \frac{\chi_x(\alpha) - \chi_y(\alpha)}{\varphi(\alpha)} \chi_{-u}(\alpha) d\alpha,$$

where the integration is with respect to the Lebesgue measure on  $\mathbb{T}$ .

*Proof.* The discrete Laplacian on  $\mathbb{Z}^2$  is defined by the formula

$$\Delta f(x) = 4f(x) - (f(x + e_1) + f(x + e_2) + f(x - e_1) + f(x - e_2)).$$

We are looking for functions which are harmonic with respect to this Laplacian except for a set of sources. For example, if we have two sources:  $+1$  at  $x$  and  $-1$  at  $y$ , then we need to solve the equation:

$$\Delta v = \delta_x - \delta_y$$

The main tool here is the Fourier transform  $\mathcal{F}$ . For every function  $f$  on  $\mathbb{Z}^2$  we can define a function  $\hat{f} = \mathcal{F}(f)$  on  $\mathbb{T}$  as

$$\hat{f}(\alpha) = \sum_{k \in \mathbb{Z}^2} f(k) \chi_k(\alpha).$$

Then the significance of function  $\varphi(\alpha)$  is that the Laplacian operator corresponds to multiplication by the function  $\varphi$ :

$$\widehat{\Delta f}(\alpha) = \varphi(\alpha)\widehat{f}(\alpha).$$

Hence, we can find the Fourier transform of  $v(x)$  as

$$\widehat{v}(\alpha) = \frac{\widehat{\delta}_x(\alpha) - \widehat{\delta}_y(\alpha)}{\varphi(\alpha)} = \frac{\chi_x(\alpha) - \chi_y(\alpha)}{\varphi(\alpha)}$$

and the result of the theorem follows by taking the inverse Fourier transform.  $\square$

Can we calculate the integral explicitly?

For a  $u \in \mathbb{Z}^2$ , define

$$H(u) = 4 \int_{\mathbb{T}^2} \frac{1 - \chi_{-u}(\alpha)}{\varphi(\alpha)} d\alpha.$$

It is clear that in order to calculate  $v(u)$ , it is sufficient to be able to calculate function  $H(u)$  for all  $u \in \mathbb{Z}^2$ .

For us, the most interesting is the function  $Y(e, f)$ . If we set  $x = e^-$  and  $y = e^+$ , calculate voltage  $v$ , and evaluate  $i_e(f) = v(f^+) - v(f^-)$ , then a calculation shows that

$$Y(e, f) = \frac{1}{4} \left[ H(f^- - e^+) - H(f^- - e^-) - H(f^+ - e^+) + H(f^+ - e^-) \right]$$

**Lemma 6.2.2.**

$$H(n, n) := 4 \int_{\mathbb{T}^2} \frac{1 - e^{2\pi n(\alpha_1 + \alpha_2)}}{\varphi(\alpha)} d\alpha = \frac{4}{\pi} \sum_{k=1}^n \frac{1}{2k-1}.$$

The values off the diagonal can be calculated by checking that  $H(x, y) = H(y, x)$ ,  $H(x, -y) = H(x, y)$ ,  $H(-x, y) = H(x, y)$  and that the identity  $\Delta H = -4\delta_0$  holds. Then one can calculate  $H$  recursively at gradually increasing distances from the diagonal.

### Some results about UST on $\mathbb{Z}^2$

Using the voltage function and the Burton-Pemantle one can calculate the distribution of vertex degrees in the UST on  $\mathbb{Z}^2$ . The results are in Figure 6.2. This is given as Exercise 4.10 in Lyons-Peres book.

### Exercises

*Ex. 6.2.3.* Calculate  $H(3, 2)$ .

Degree	Probability
1	$\frac{8}{\pi^2} \left(1 - \frac{2}{\pi}\right) = .294^+$
2	$\frac{4}{\pi} \left(2 - \frac{9}{\pi} + \frac{12}{\pi^2}\right) = .447^-$
3	$2 \left(1 - \frac{2}{\pi}\right) \left(1 - \frac{6}{\pi} + \frac{12}{\pi^2}\right) = .222^+$
4	$\left(\frac{4}{\pi} - 1\right) \left(1 - \frac{2}{\pi}\right)^2 = .036^+$

**Figure 6.2:** Probability distribution of the vertex degrees in the UST on  $\mathbb{Z}^2$ .

*Ex. 6.2.4.* If  $G$  is a graph and  $K$  is a subset of vertices, then the *edge boundary* of  $K$  is the set of edges  $\partial K$  that connect  $K$  to its complement. An infinite graph  $G$  is called *edge-amenable* if there is a sequence of finite subgraphs  $G_n = (V_n, E_n)$  such that  $G$  is the union of these subgraphs and

$$\lim_{n \rightarrow \infty} \frac{|\partial V_n|}{|V_n|} = 0.$$

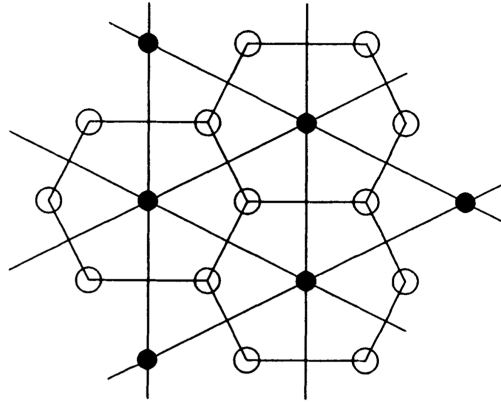
If  $T$  is a spanning tree in an edge-amenable graph  $G$ , then

$$\lim_{n \rightarrow \infty} |V_n|^{-1} \sum_{x \in V_n} d_t(x) = 2.$$

### 6.3 Bijection with Domino Tilings

Suppose now that the graph  $G$  is planar, that is that it can be embedded in  $\mathbb{R}^2$  (or more generally, in some other 2-dimensional surface), such that vertices are represented by points and the edges – by non-intersecting intervals that connect these points.

Then the graph  $G$  splits the plane in connected components, which are called faces, one of which is infinite. If the graph is finite, then the number of faces can be calculated by using the Euler formula:  $|V| - |E| + |F| = 2$ , or more generally,  $= 2 - g$ , where  $g$  is the genus of the imbedding surface.



**Figure 6.3:** A part of infinite triangular lattice graph and its dual.

We can define the dual graph  $G^*$  as the graph that has faces of the embedded diagram as its vertices. Two vertices in the dual graph are connected if the corresponding faces share the same edge in the diagram of the original graph. Note that because of this definition we can essentially identify the edges in the original and the dual.

See an example in Figure 6.3. The edge that connects two vertices of the dual graph is identified with the edge of the original graph that it crosses.

We also define a bipartite graph  $\widehat{G}$ . One class of vertices in this graph is the union of the vertices of the original and the dual graph,  $V \cup V^*$ , and the other class of vertices corresponds to the edges of the original graph  $E$ , – or equivalently to edges of the dual graph. These new vertices can be graphically imagined as the midpoints of the edges. Then a “vertex”  $e \in E$  is connected to a vertex  $v \in V$  if edge  $e$  is incident to the vertex  $v$  in the original graph  $G$ . Similarly, vertex  $e \in E$  is connected to vertex  $v^* \in V^*$ , if edge  $e$  is incident to the face  $v^*$  in the imbedding of the graph  $G$ .

The bipartite graphs that can be represented as  $\widehat{G}$  for some planar graph are called *Temperleyan*; Temperley found that square lattice graph  $\mathbb{Z}^2$  has this

representation and connected domino tilings of  $\mathbb{Z}^2$  to spanning trees of the corresponding graph  $G$ .

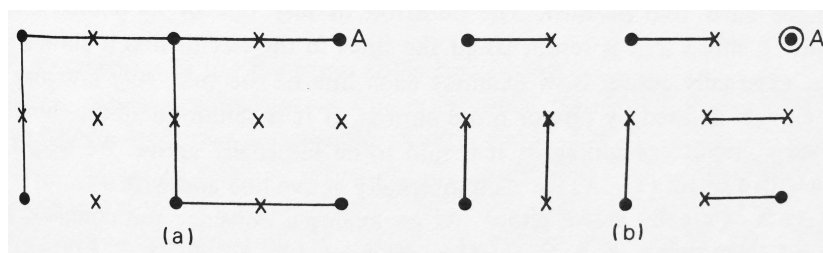
Next, there is a correspondence between subgraphs of  $G$  and  $G^*$ . If  $H$  is a subgraph of  $G$  with the same set of vertices  $V$ , then a subgraph  $H^*$  of  $G^*$  contains an edge  $e$  if and only if the corresponding edge is absent in  $H$ .

For example, the subgraph  $G \subset G$  corresponds to the subgraph with no edges in  $G^*$ .

This correspondence is clearly involutive:  $(H^*)^* = H$ .

**Lemma 6.3.1.** *Let  $G$  be a finite planar graph. Then  $T^*$  is a spanning tree of  $G^*$  if and only if  $T$  is a spanning tree of  $G$ .*

Remarkably, the spanning trees in the graph  $G$  correspond to perfect matchings in the graph  $\widehat{G}$ . Recall that a *matching* is a collection of edges such that no two edges have a common vertex. A *perfect matching* is a matching such that every vertex belongs to an edge in the matching.



**Figure 6.4:** Bijection between spanning trees and matchings

The bijection is easier to explain by an example shown in Figure 6.4. The graph  $G$  is a part of the lattice  $\mathbb{Z}^2$  and its vertices are shown by dots. The additional vertices in the graph  $\widehat{G}$ , – that is, the vertices of the dual graph and half-points of the edges, – represented by crosses.

Assume that we are given a spanning tree on graph  $G$  and select a vertex of this tree as a root. This is vertex  $A$  in the example. Then we can define a perfect matching on the graph  $\widehat{G} - A$ . The recipe is to follow the tree from  $A$  to leaves and include the half-edges that go from the mid-point of the edge to the vertex of  $G$  (from a cross to a dot).

The remaining crosses belong to the spanning tree of the dual graph  $G^*$  and we can do the similar procedure on this tree. (By convention, we can choose the root of this tree equal to the vertex of  $G^*$  that corresponds to the outer face.)

By using this correspondence, the results about the spanning trees can be translated to results about the perfect matchings (or domino tilings) and vice versa.



## 6.4 Connection with Eulerian circuits

An *Euler circuit* (also called an *Eulerian circuit*) is a circuit that uses every edge of a graph exactly once.

**Theorem 6.4.1** (Euler). *A finite, strongly connected, directed graph which is balanced (each vertex has in-degree = out-degree) has an Euler circuit.*

There is a quick algorithm to find an Euler circuit called the Fleury algorithm. For us what is interesting is the following connection with uniform spanning trees.

In a balanced, strongly connected, directed graph, take any spanning tree  $T$ , with directed edges toward an arbitrary root. From the root do an arbitrary walk, at each stage choosing an unused edge at random but saving the spanning-tree-edge until last. One can check that this method always produces an Euler circuit.

**Theorem 6.4.2.** *If  $T$  is a uniform random spanning tree and the random walk steps are chosen uniformly at random we get a random Euler circuit which is distributed uniformly at the set of all Euler circuits.*

*Proof.* ???

□

As an example consider the discrete torus  $\mathbb{Z}_N^d$ . Replace each edge by 2 directed edges. So, in-degree = out-degree =  $2d$ . Any Eulerian circuit consists of  $2d$  “loop” from the origin.

Simulations suggest the following conjecture, which is on David Aldous’ list of favorite open problems.

**Conjecture 6.4.3** (Aldous). For  $d \geq 3$ , out of  $2d$  loops at the origin, some have length  $O(1)$ , all others have length of order  $N^d$  as  $N \rightarrow \infty$ .

The point is that there is no loops with an “in-between” size.




**Figure 6.5:** Trees by Tomioka Soichiro

## Chapter 7

# Galton-Watson Trees

### 7.1 Galton – Watson process

The Galton-Watson process is a particular simple example of a branching process, a class of stochastic processes that finds many applications, from the study of epidemics to the study of neutron proliferation in a nuclear reaction. The original application in the paper by Galton and Watson was to a problem in genetics, namely, to the problem of family name extinction. How fertile should family members be to insure that the family name will not die out in future generations? The branching processes has also been used in the study of queues. In this case the offspring of a customer are those, who arrive while the customer is being served.

Recently, branching processes have been applied to the study of random graphs and other random geometric objects.

The *Galton-Watson process* is a Markov Chain  $X_n$  on the non-negative integers, where  $X_n$  represents the size of  $n$ -th generation. The random variables  $X_{n+1}$  and  $X_n$  are related by certain transition probabilities, and the evolution of the generation size at time  $n$  can be described by the equation:

$$X_{n+1} = \sum_{i=1}^{X_n} L_i^{(n)},$$

where  $L_i^{(n)}$  are independent copies of a random variable  $L$  that has the *offspring distribution*  $\mathbb{P}(L = k) = p_k$ .

**Assumption:** In all considerations and results below we assume that the offspring distribution is not trivial in the sense that  $p_k > 0$  for some  $k > 1$ .

It is often useful to enrich the Galton-Watson process by keeping information not only about the number of the individuals at time  $n$  but also about the details of the family tree  $T$ . In this description, the random variable  $X_n$  is the number of vertices of  $T$  at the depth  $n$ .

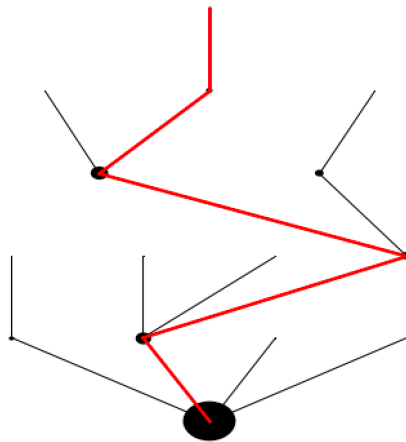
The randomness in the tree  $T$  is described by the following rule: “Each vertex can have  $k$  children with probability  $p_k$ . The numbers of children for different vertices are independent.”

More formally, the probability on the set of rooted locally-finite trees  $T$  is defined by probabilities of the cylinder sets  $\{T_n = t_n\}$ ,  $n = 0, 1, \dots$ , where  $T_n$  is the restriction of the random tree  $T$  to its first  $n$  generations. In particular, the tree  $T_0$  always consists of a single vertex, the root. The tree  $T$  is determined by an increasing sequence of its finite subtrees  $T_0 \subset T_1 \subset \dots \subset T_n \subset \dots$

The conditional probability of  $T_{n+1}$  given  $T_n$  is given by the formula

$$\prod_{v_i : l(v_i)=n} p_{d(v_i)},$$

where the product is over all vertices  $v_i$  that have level  $n$  in  $T_n$  and  $d(v_i)$  denotes the number of children that vertex  $v_i \in T_n$  has in  $T_{n+1}$ .



**Figure 7.1:** A Galton-Watson tree with  $X_0 = 1$ ,  $X_1 = 4$ ,  $X_2 = 4$ ,  $X_3 = 2$ ,  $X_4 = 3$ ,  $X_5 = 1$  and  $X_k = 0$  for  $k > 5$ .

The resulting tree  $T = \lim_{n \rightarrow \infty} T_n$  is called the *Galton - Watson tree*, and the sequence  $X_n$  is the *Galton-Watson process*.

For the study of the Galton-Watson process  $X_n$  it is useful to define the probability generating function of the random variable  $L$ ,

$$f(s) = \sum_{k=0}^{\infty} p_k s^k = p_0 + p_1 s + p_2 s^2 + \dots$$

It is often called the *generating function* of the Galton-Watson process.

The *mean* of the Galton-Watson process is the expectation of the offspring number,

$$\mu := \mathbb{E}L = f'(1).$$

The GW process is called *subcritical*, *critical*, or *supercritical*, depending on whether  $\mu < 1$ ,  $\mu = 1$ , or  $\mu > 1$ , respectively.

**Lemma 7.1.1.** *Let  $X_n$  be the Galton-Watson process. Then  $\frac{X_n}{\mu^n}$  is a martingale with respect to  $\mathcal{F}_n = \sigma(X_0, \dots, X_n)$ .*

*Proof.* Since  $X_n$  is a Markov process, it is enough to condition on  $X_n$  instead of  $\sigma(X_0, \dots, X_n)$ . We have

$$\mathbb{E}\left[\frac{X_{n+1}}{\mu^{n+1}} \mid X_n\right] = \frac{1}{\mu^{n+1}} \sum_{i=1}^{X_n} \mathbb{E}[L_i | X_n] = \frac{X_n}{\mu^n},$$

since  $\mathbb{E}[L_i | X_n] = \mathbb{E}[L_i] = \mu$ . □

The martingale  $X_n/\mu^n$  is non-negative, and therefore by Doob's theorem it has an almost sure limit  $W \geq 0$ .

Now, what is the probability of extinction of the GW branching process? This is the original question posed by Galton and Watson. We denote this probability by  $q$ ,

$$q := \mathbb{P}(X_n = 0 \text{ for some } n > 0).$$

It turns out that  $q = 1$  for all subcritical and critical processes.

**Theorem 7.1.2.** *If  $\mu \leq 1$ , then  $q = 1$ , that is,  $X_n = 0$  for all sufficiently large  $n$ . In particular,  $\frac{X_n}{\mu^n} \xrightarrow{a.s.} W = 0$ .*

(By the way, this is an example of an  $L^1$ -bounded martingale which is convergent almost surely but not in  $L^1$ .)

For the proof we need the following lemma.

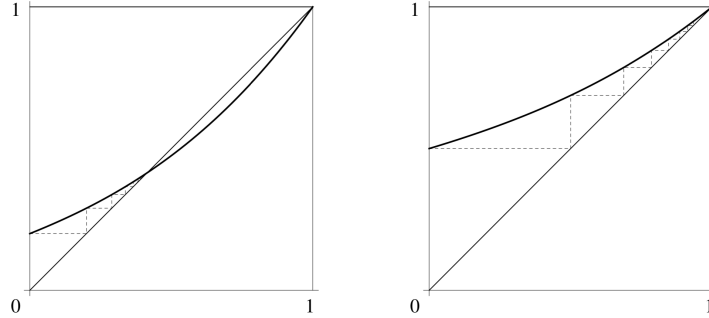
**Lemma 7.1.3.** *The generating function for the Galton Watson process at time  $n$ ,  $X_n$ , is*

$$\mathbb{E}s^{X_n} = f^{(n)}(s) := \underbrace{f \circ \dots \circ f}_{n \text{ times}}(s)$$

*Proof.* We have  $\mathbb{E}s^{X_0} = s$ , and then we proceed by induction:

$$\begin{aligned} \mathbb{E}s^{X_n} &= \mathbb{E}\left[\mathbb{E}\left(s^{\sum_{i=1}^{X_{n-1}} L_i} \mid X_{n-1}\right)\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{X_{n-1}} \mathbb{E}(s^{L_i} | X_{n-1})\right] = \mathbb{E}f(s)^{X_{n-1}} \\ &= f^{(n-1)}(f(s)) = f^{(n)}(s). \end{aligned}$$

□



**Figure 7.2:** Generating function  $f(s)$  for  $\mu > 1$  and  $\mu \leq 1$ , (on the left and right graphs, respectively)

*Proof of Theorem 7.1.2.* By the lemma above, we have  $\mathbb{P}(X_n = 0) = f^{(n)}(0)$ . Therefore, the probability of extinction is  $q = \lim_{n \rightarrow \infty} f^{(n)}(0)$ . It is easy to check that  $f(s)$  is differentiable, increasing, concave and  $f(1) = 1$ . In addition,  $\mu$  is the slope of  $f(s)$  at  $s = 1$ . Then it is clear from the picture in Figure 7.2 that if  $\mu \leq 1$ , then  $f^{(n)}(0)$  converges to 1. This proves the theorem.  $\square$

Theorem 7.1.2 settles the question about the extinction probability and about the limit of the Galton-Watson martingale in the subcritical and critical case. The argument in its proof also shows how to calculate the extinction probability for the super-critical process.

**Theorem 7.1.4.** *The extinction probability  $q := \mathbb{P}(X_n = 0 \text{ for some } n)$  equals the smallest root of the equation  $s = f(s)$ .*

Thus, if  $\mu > 1$  then  $q < 1$  and  $X_n > 0$  for all  $n$  with positive probability. However, even though  $X_n$  stays positive with positive probability, there is still a possibility of the zero limit for the Galton-Watson martingale, that is, it is still possible that  $\frac{X_n}{\mu^n} \xrightarrow{a.s.} 0$ . This means that  $X_n$  grows slower than  $\mu^n$ . The Kesten-Stigum theorem shows that it does not happen under very mild conditions on the distribution of the offspring.

**Theorem 7.1.5 (Kesten-Stigum).** *Suppose that  $X_n$  is a super-critical Galton-Watson process with offspring random variable  $L$  and let  $\mu = \mathbb{E}L > 1$ . Let  $W = \lim_{n \rightarrow \infty} X_n / \mu^n$ . The limit of  $\frac{X_n}{\mu^n}$  is not identically 0 (a.s.) if and only if  $E(L \log^+ L) = \sum_{k=1}^{\infty} (k \log k) p_k < \infty$ .*

This condition is only slightly stronger than the condition on the existence of  $\mu = \mathbb{E}L$ . For example, this condition holds if  $L$  has finite variance.

For the classical proof see Athreya-Ney (1972), Part I.C. We will also give another proof in Section 7.4 below.

Even if this condition does not hold,  $X_n$  grows only slightly slower than  $\mu^n$ , as the following theorem show.

**Theorem 7.1.6** (Seneta-Heyde). *If  $1 < \mu < \infty$ , then there exists constants  $c_n$  such that*

1.  $\lim X_n/c_n$  exists a.s. in  $[0, \infty)$ ;
2.  $\mathbb{P}[\lim X_n/c_n = 0] = q$ ;
3.  $c_{n+1}/c_n \rightarrow \mu$ .

We will prove this theorem by using methods of martingale theory and a zero - one law for Galton-Watson trees.

Call a property of trees *inherited* if two conditions are satisfied:

1. every finite tree has this property, and
2. if a tree has this property, then all the descendant trees of the children of the root also have this property.

*Example 7.1.7.* For a rooted tree  $T$ , let  $X_n$  be the number of its vertices in the level  $n$ . Suppose that  $c_n$  is a sequence of positive constants such that  $\lim_{n \rightarrow \infty} c_{n+1}/c_n = a > 1$ . Define a property  $P$  by requiring that it holds for tree  $T$  if  $\lim_{n \rightarrow \infty} X_n/c_n = 0$ . This property is inherited. Indeed, it is obviously satisfied for every finite tree. Then, if the property holds for a tree  $T$  and  $T^{(i)}$  is a descendant tree with the corresponding sequence  $X_n^{(i)}$ , then we observe that  $X_n^{(i)} \leq X_{n+1}$  and therefore,

$$\frac{X_n^{(i)}}{c_n} \leq \frac{X_{n+1}}{c_{n+1}} \frac{c_{n+1}}{c_n} \rightarrow 0 \times a = 0.$$

*Ex. 7.1.8.* Suppose that  $c_n$  is a sequence of positive constants such that  $\lim_{n \rightarrow \infty} c_{n+1}/c_n = a > 1$ . The tree property  $\lim X_n/c_n < \infty$  is inherited.

**Lemma 7.1.9.** *For a supercritical Galton-Watson tree, each inherited property has probability either  $q$  or  $1$ , where  $q$  is the probability of extinction.*

*Proof.* Let  $A$  be the set of trees with the given property. We are going to show that  $\mathbb{P}(A)$  is either  $q$  or  $1$ .

For a tree  $T$  with  $k$  children of the root, let  $T^{(1)}, \dots, T^{(k)}$  denote the descendant trees of these children. Then

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{E} \left[ \mathbb{P}(T \in A | X_1) \right] \\ &\leq \mathbb{E} \left[ \mathbb{P}(T^{(1)} \in A, \dots, T^{(X_1)} \in A | X_1) \right] \end{aligned}$$

by the definition of the inherited property and the monotonicity of probability. Since  $T^{(i)}$  are independent and have same distribution as  $T$ , we find that

$$\mathbb{E} \left[ \mathbb{P}(T^{(1)} \in A, \dots, T^{(X_1)} \in A | X_1) \right] = \mathbb{E}[P(A)^{X_1}] = f(\mathbb{P}(A)).$$

Hence  $\mathbb{P}(A) \leq f(\mathbb{P}(A))$ . In addition,  $\mathbb{P}(A) \geq q$ , since the property holds for every finite tree. Hence, by Theorem 7.1.4 and properties of the function  $f$ , which are seen in Figure 7.2,  $\mathbb{P}(A)$  is either  $q$  or  $1$ .  $\square$

*Proof of Theorem 7.1.6.* Let  $s_0 \in (q, 1)$  and set  $s_{n+1} = f^{(-1)}(s_n)$  for  $n \geq 0$ . Then  $s_n \rightarrow 1$ . In the proof of Lemma 7.1.3 we established that  $\mathbb{E}[s^{X_{n+1}} | X_1, \dots, X_n] = f(s)^{X_n}$ . This implies that  $s_n^{X_n}$  is a martingale,

$$\mathbb{E}[(s_{n+1})^{X_{n+1}} | X_1, \dots, X_n] = f(s_{n+1})^{X_n} = s_n^{X_n}.$$

Since this martingale is positive and bounded it converges both almost surely and in  $L^1$  to a limit  $Y \in [0, 1]$ , such that  $\mathbb{E}(Y) = \mathbb{E}(s_0^{X_0}) = s_0$ .

Let

$$c_n := -1/\log s_n.$$

It is easy to check that  $\log f(s)/\log s \rightarrow \mu$  as  $s \rightarrow 1$ , hence  $c_{n+1}/c_n \rightarrow \mu$ , which is claim (3).

Then  $s_n^{X_n} = e^{-X_n/c_n}$ , and therefore  $X_n/c_n$  converges a.s. to a limit  $\tilde{W} = -\log Y$  supported in  $[0, \infty]$ .

By Example 7.1.7, the property  $X_n/c_n \rightarrow 0$  is inherited. Probability of this property cannot be 1 since this would imply that the random variable  $Y$  is 1 but we know that  $\mathbb{E}(Y) = s_0 < 1$ . Therefore this probability is  $q$  and claim (2) is established.

Similarly, the property  $\lim X_n/c_n < \infty$  is inherited. It cannot have probability  $q < 1$  because this would imply that with probability  $1 - q$  the random variable  $Y$  is 0. Since  $Y \in [0, 1]$ , this would imply that  $\mathbb{E}Y \leq q$ , while we know that  $\mathbb{E}Y = s_0 \in (q, 1)$ . Hence,  $\lim X_n/c_n < \infty$  with probability 1. This implies claim (1).  $\square$

From the theory of Markov chains, we can extract that if  $\varphi(t)$  denote the moment generating function of the limit random variable  $W$ , that is,  $\varphi(t) = \mathbb{E}(e^{-tW})$ , then the following equation holds,

$$\varphi(\mu t) = f(\varphi(t)).$$

This equation, which is sometimes called Abel's equation, can be used to study the properties of  $W$ . In particular, it can be shown that  $W$  is absolutely continuous.

## 7.2 GW process with immigration

In the Galton – Watson process with immigration there is a new element, - at time  $n$  additional  $Y_i$  particles arrive. So, the equation for the generation sizes is now

$$X_{n+1} = \sum_{i=1}^{X_n} L_i^{(n)} + Y_{n+1},$$

where  $L_i^{(n)}$  are independent copies of a random variable that have the *offspring distribution*  $\mathbb{P}(L = k) = p_k$ , and  $Y_{n+1}$  are non-negative random variables, the immigration process. Assume also that  $X_0 = 0$ .



We know that without immigration,  $X_n/\mu^n$  has a.s. a finite limit  $W$ . What can we say about the processes with immigration?

Consider the super-critical case.

**Theorem 7.2.1** (Seneta). *Let  $X_n$  be the generation sizes of GW process with immigration, and let  $Y_n$  be i.i.d random variables. Suppose that  $\mu = \mathbb{E}(L) > 1$ . If  $\mathbb{E} \log^+ Y < \infty$ , then  $\lim X_n/\mu^n$  exists and the limit is finite a.s., whereas if  $\mathbb{E} \log^+ Y = \infty$ , then  $\limsup X_n/c^n = \infty$  a.s. for every constant  $c > 0$ .*

In the proof we will need several auxiliary results.

**Lemma 7.2.2.** *Suppose  $X, X_1, X_2, \dots$  are non-negative i.i.d. random variables. Then*

$$\limsup_{n \rightarrow \infty} X_n/n = \begin{cases} 0, & \text{if } \mathbb{E}X < \infty, \\ \infty, & \text{if } \mathbb{E}X = \infty. \end{cases}$$

*Proof.* Exercise on the Borel-Cantelli lemma. □

**Lemma 7.2.3.** *Suppose  $X, X_1, X_2, \dots$  are non-negative i.i.d. random variables and  $\mathbb{E}X < \infty$ . Then,*

$$\sum_{n=1}^{\infty} e^{X_n}/\mu^n < \infty$$

for all  $\mu > 1$ .

*Proof.* By the previous lemma, for every  $\lambda > 0$ ,  $X_n < \lambda n$  a.s. for all sufficiently large  $n$ . By choosing a value of  $\lambda < \log \mu$  we can infer that a.s.  $e^{X_n}/\mu^n$  is exponentially declining for all sufficiently large  $n$  and therefore the series  $\sum_{n=1}^{\infty} e^{X_n}/\mu^n$  is convergent. □

*Proof of Theorem 7.2.1.* First, if  $\mathbb{E} \log^+ Y = \infty$ , then by Lemma 7.2.2,  $\limsup(\log^+ Y_n)/n = +\infty$ , which implies that  $\limsup(\log^+(Y_n/c^n))/n = \infty$  for any positive  $c$ . This implies that  $\limsup Y_n/c^n = \infty$ . Since  $X_n \geq Y_n$ , this implies the second claim of the theorem.

Now, assume that  $\mathbb{E}[\log^+ Y] < \infty$ . It is not difficult to see that  $\mu^{-n}X_n$  is a sub-martingale. (Let  $\mathcal{F}$  denote the  $\sigma$ -field generated by  $Y_i, i = 1, 2, \dots$ . By the same calculation as in the proof of Lemma 7.1.1 ,

$$\mathbb{E}[\mu^{-(n+1)}X_{n+1}|X_n, \mathcal{F}] = \mu^{-n}X_n + \mu^{-(n+1)}Y_{n+1} \geq \mu^{-n}X_n,$$

and taking the expectation over  $\mathcal{F}$  we recover the sub-martingale property.)

Now let us calculate the conditional expectation of  $\mu^{-n}X_n$  with respect to the  $\sigma$ -algebra  $\mathcal{F}$ . Let  $X_{n,k}$  be all descendants at time  $n$  of the immigrants that arrived at time  $k$ . Then, we have

$$\begin{aligned} \mathbb{E}(\mu^{-n}X_n|\mathcal{F}) &= \mathbb{E}\left(\mu^{-n} \sum_{k=1}^n X_{n,k}|\mathcal{F}\right) \\ &= \sum_{k=1}^n \mu^{-k} \mathbb{E}(\mu^{-(n-k)}X_{n,k}|\mathcal{F}). \end{aligned}$$

It is clear that  $X_{n,k}$  is the usual Galton-Watson process (without immigration) that started with  $Y_k$ , instead of 1 particle. By Lemma 7.1.1, the sequence  $\mu^{-(n-k)}X_{n,k}$  is a martingale and  $\mathbb{E}(\mu^{-(n-k)}X_{n,k}|\mathcal{F}) = Y_k$ . Therefore,

$$\mathbb{E}(\mu^{-n}X_n|\mathcal{F}) = \sum_{k=1}^n \mu^{-k}Y_k.$$

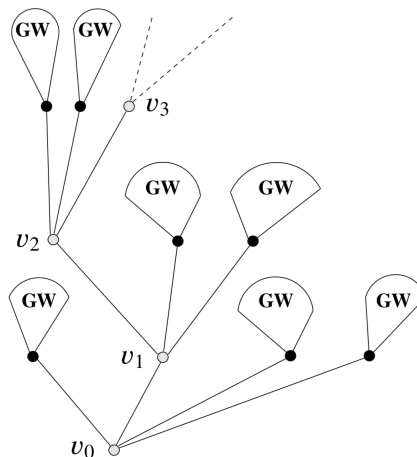
By applying Lemma 7.2.3 to  $X_k = \log^+ Y_k$ , we find that this series is convergent a.s.

It follows that almost surely, for a fixed sequence  $Y_i$ ,  $\mu^{-n}X_n$  is a non-negative sub-martingale bounded in  $L^1$ . By Doob's convergence theorem, this implies that  $\mu^{-n}X_n$  almost surely converges to a random variable  $W$  with finite expectation. This implies the statement of the theorem.  $\square$

### 7.3 Size-biased GW trees

In the next section we are going to prove the Kesten-Stigum theorem following the method by Lyons, Pemantle, and Peres. Recall that this theorem says that in a supercritical process the ratio  $X_n/\mu^n$  converges to a non-zero limit if and only if  $\mathbb{E}L \log^+ L < \infty$ .

The proof is going to be based on a new (that is, different from the Galton-Watson measure  $GW$ ) measure  $\widehat{GW}$  on trees, which we define in this section.



**Figure 7.3:** A scheme for the GW size-biased tree

Suppose a random variable  $X$  takes value in *non-negative* integers, and has the probability distribution  $p_k = \mathbb{P}(X = k)$ ,  $k = 0, 1, \dots$ , with a finite expectation  $\mu$ . Then, a size-biased version of  $X$ , a random variable  $\hat{X}$ , takes values in the set of *positive* integers and has the probability distribution  $\hat{p}_k = kp_k/\mu$ . (This is obviously a probability distribution by the definition of  $\mu$ .)

Let the probability distribution  $p_k$  be interpreted as the offspring distribution for a Galton-Watson tree. We define another random tree, the *size-biased* Galton-Watson tree by using the size-biased distribution.

In contrast with usual GW trees, this new random tree is always infinite and it has a distinguished path from the root to infinity. In particular, we define a *stemmed tree* as a pair  $[t, v]$ , where  $t$  is an infinite rooted tree with root  $v_0$ , and  $v = (v_0, v_1, v_2, \dots)$  is its *stem*, that is, a path on this tree, such that  $d(v_k, v_0) = k$ . We assume that the infinite tree is locally-finite, that is, every vertex has only finite number of children. The *size-biased Galton-Watson tree* is a random stemmed tree with a specific distribution that we describe below.

Schematically, this tree is represented in Figure 7.3 with the distinguished path given by the sequence of vertices  $v_0, v_1, \dots$ .

The tree is random and can be generated in the following way. The vertices will be of two types, usual and special. There will always be exactly one special vertex in each generation. Start with the original vertex  $v_0$ , which is special, and give it a random number of children distributed according to the size-biased law  $\{\hat{p}_k\}$ . In particular, note that this number is always positive. From these children, pick one at random. This will be  $v_1$ , the special vertex in generation 1, and the second vertex in the distinguished path. On all other children, the descendant trees are the ordinary GW trees with the offspring distribution  $\{p_k\}$ . (In particular, these trees can be finite.) However, we treat  $v_1$  in the same way as the initial vertex,  $v_0$ . In particular, the number of its children is distributed according to size-biased law  $\{\hat{p}_k\}$  and one of these children (grand-children of  $v_0$ ) is selected as a new special vertex, and a new vertex in the distinguished path,  $v_2$ . We continue in this way indefinitely.

This procedure defines a probability measure on all stemmed trees. We denote this measure by  $\widehat{GW}$ . How can we calculate the probabilities of cylinder sets using this measure?

Suppose that  $[t, v]_n$  be the set of all stemmed trees, which agrees with the stemmed tree  $[t, v]$  up to the  $n$ -th generation. Also, let  $[t]_n$  denote the set of all rooted trees that agree with tree  $t$  up to the  $n$ -th generation.

**Lemma 7.3.1.** *For all sets  $[t, v]_n$  and all  $n$ ,*

$$\widehat{GW}([t, v]_n) = \mu^{-n} GW([t]_n),$$

where  $GW$  is the Galton-Watson measure on the space of rooted trees.

*Proof.* We prove it by induction. For  $n = 0$ , there is only one tree, that consists of the root  $v_0$  and the statement is true. Now consider the restriction of a stemmed tree  $[t, v]$  up to the generations  $n + 1$ . If the root of  $t$  has  $k$  children, then we denote the corresponding descendant trees as  $t^{(1)}, \dots, t^{(k)}$ . Then, by the definition of the size-biased trees we have

$$\widehat{GW}([t, v]_{n+1}) = \frac{kp_k}{\mu} \frac{1}{k} \widehat{GW}([t^{(i)}, v]_n) \prod_{j \neq i} GW([t^{(j)}]_n),$$

where  $t^{(i)}$  is the descendant tree in which the distinguished vertex  $v_1$  lies. The first factor here is the probability that the root of the tree has  $k$  children and the second is that  $v_1$  is in the sub-tree  $t^{(i)}$ .

By using the induction hypothesis, we can write this as

$$\widehat{GW}([t, v]_{n+1}) = \frac{p_k}{\mu} \frac{1}{\mu^n} GW([t^{(i)}]_n) \prod_{j \neq i} GW([t^{(j)}]_n),$$

At the same time for the corresponding Galton-Watson measure on trees, we have:

$$GW([t]_{n+1}) = p_k \prod_{j=1}^k GW([t^{(j)}]_n)$$

By comparing these expressions we find that

$$\widehat{GW}([t, v]_{n+1}) = \frac{1}{\mu^{n+1}} GW([t]_{n+1}),$$

which completes the induction.  $\square$

Given the measure  $\widehat{GW}$  over all stemmed trees, we can define a new measure over all rooted trees as the marginal of  $\widehat{GW}$ . For cylinder sets we have,

$$\widehat{GW}([t]_n) = \sum_v \widehat{GW}([t, v]_n),$$

where the summation is over all stems  $v$  of the tree  $t$ . Since a portion of the stem from the root to generation  $n$  is completely determined by its vertex  $v_n$  in generation  $n$ , and since the probabilities  $\widehat{GW}([t, v]_n)$  are all equal to each other and equal to  $\mu^{-n} GW([t]_n)$  by Lemma 7.3.1, we have

$$\widehat{GW}([t]_n) = X_n(t) \widehat{GW}([t, v]_n) = \frac{X_n(t)}{\mu^n} GW([t]_n), \quad (7.1)$$

where  $X_n(t)$  is the number of vertices in generation  $n$  of the tree  $t$ . (Informally, measure  $\widehat{GW}$  pays more attention to trees with larger offspring than measure  $GW$ . This is why it is called size-biased measure on trees.)

## 7.4 Supercritical case and a proof of the Kesten-Stigum Theorem (Thm. 7.1.5)

One interpretation of the equation (7.1) is that the random variable  $X_n/\mu^n$  is the ratio of two measures of a cylindrical set of trees that agree on the first  $n$  generations. Therefore, questions about the limit of this random variable are related to questions of absolute continuity of two measures  $\widehat{GW}$  and  $GW$  with respect to each other.

Recall that a measure  $\mu$  is called *absolutely continuous* with respect to a measure  $\nu$ , denoted  $\mu \ll \nu$ , if for each  $A \in \mathcal{F}$ ,  $\nu(A) = 0$  implies that  $\mu(A) = 0$ . In this case the Radon-Nikodym derivative  $f = d\mu/d\nu$  exists and allows us to compute the measure  $\mu$  using formula:

$$\mu(A) = \int_A f d\nu,$$

We call  $f$  density. There can be several such densities but the set where two densities are different from each other, this set have  $\nu$  measure 0.

*Example 7.4.1.* Consider the Lebesgue measure  $\mu$  on  $[0, 1)$ , and  $\nu = \mu + \delta_0$ . Then,  $\mu \ll \nu$  and the density  $f(x) = 1$  at all points  $x \neq 0$ . At zero,  $f(0) = 0$ . The density  $f$  can be changed at any Lebesgue set of measure zero that does not include  $x = 0$ .

What we want to prove is that the limit of  $X_n/\mu^n$  have positive expectation. Typically, this is done by showing the uniform integrability of  $X_n/\mu^n$ , which implies that the expectation of the limiting random variable  $W$  equals to the limit of the expectations of  $X_n/\mu^n$ , which is not zero. Here, we will use an alternative method.

The idea is to consider the martingale sequence  $X_n/\mu^n$  as a change of measure factor. If the expectation of the limit of this sequence is different from 1, then it means that part of the probability disappeared into a set of measure zero. The idea is to rule this possibility out and show that the limiting measure  $\widehat{GW}$  is absolutely continuous with respect to the old measure  $GW$ .

Namely, we will show that if the condition  $\mathbb{E}L \log^+ L < \infty$  holds, then under the *new* measure  $\widehat{GW}$  the random variables  $X_n/\mu^n$  almost surely converge to a *finite* limit  $W$ . Since random variables  $X_n/\mu^n$  have the meaning of Radon-Nikodym derivatives on cylindrical sets, it turns out that their convergence to the finite limit under  $\widehat{GW}$  implies that the new measure  $\widehat{GW}$  is absolutely continuous with respect to the old measure  $GW$ , with the Radon-Nikodym derivative  $W = d\widehat{GW}/dGW$ . It follows by the property of the Radon-Nikodym derivatives that  $\int W dGW = \int d\widehat{GW} = 1$ , and therefore  $W$  is not identically zero.

In contrast, if  $\mathbb{E}L \log^+ L = \infty$  then the limit random variable  $W$  is almost surely infinite under the new measure  $\widehat{GW}$  and almost surely zero under the measure  $GW$ . The intuitive meaning of this fact is that if  $\mathbb{E}L \log^+ L = \infty$  the Galton-Watson tree is almost surely dies out, but if it is conditioned not to die out (and follows the law of  $\widehat{GW}$ ) then it grows faster than  $\mu^n$ .

Formally, we will use the following result, which is essentially a variant of the fundamental theorem about the Radon-Nikodym derivative plus some of its properties.

**Lemma 7.4.2.** *Let  $\widehat{\nu}$  be a finite measure and  $\nu$  be a probability measure on a  $\sigma$ -field  $\mathcal{F}$ . Suppose that  $\mathcal{F}_n$  are increasing sub- $\sigma$ -fields whose union generates  $\mathcal{F}$ . Suppose also that  $(\widehat{\nu}|\mathcal{F}_n)$  is absolutely continuous with respect to  $(\nu|\mathcal{F}_n)$  with*

Radon-Nykodým derivative  $W_n$ . Set  $W := \limsup_{n \rightarrow \infty} W_n$ . Then

$$\left(\widehat{\nu} \ll \nu\right) \Leftrightarrow \left(W < \infty \quad \widehat{\nu}\text{-a.e.}\right) \Leftrightarrow \left(\int d\widehat{\nu} = \int W d\nu\right)$$

and

$$\left(\widehat{\nu} \perp \nu\right) \Leftrightarrow \left(W = \infty \quad \widehat{\nu}\text{-a.e.}\right) \Leftrightarrow \left(\int W d\nu = 0\right)$$

(with the integrals taken over the whole space).

*Example 7.4.3.* In order to understand this result better consider the Lebesgue measure  $\nu$  and the atomic measure  $\widehat{\nu} = \nu + \delta_0$  on  $[0, 1)$ , with the  $\sigma$ -fields  $\mathcal{F}_n$  generated by intervals  $[(k-1)/n, k/n)$ ,  $k = 1, \dots, n$ . Then  $W_n$  converge to the random variable  $W$  which equals 1 everywhere except at  $\omega = 0$ , and at  $\omega = 0$ ,  $W(0) = \infty$ .

In this case  $\widehat{\nu}$  is not absolutely continuous with respect to  $\nu$ , and we see that while  $W < \infty$ ,  $\nu$ -almost everywhere, however  $W = \infty$  on a set of positive  $\widehat{\nu}$  measure. The implication is that one has to check that  $W$  is finite  $\widehat{\nu}$ -everywhere in order to show the absolute continuity of  $\widehat{\nu}$  with respect to  $\nu$  and ensure that  $W$  is the Radon-Nikodym derivative  $d\widehat{\nu}/d\nu$ .

This example is useful to keep in mind when reading the proof of Lemma 7.4.2.

The proof of Lemma 7.4.2 can be found in the Lyons-Peres book. (Lemma 12.2 on p. 414). Here is a wordy paraphrase of this proof.

*Proof of Lemma 7.4.2.* First, one can check that  $W_n$  is a non-negative martingale with respect to  $\nu$  and the filtration generated by  $\mathcal{F}_n$ , and therefore by Doob's theorem the sequence  $W_n$  converges  $\nu$ -almost surely to the random variable  $W$ , which is finite  $\nu$ -almost surely. However, as we have seen in Example 7.4.3, this fact does not guarantee that  $\widehat{\nu}$  is absolutely continuous with respect to  $\nu$ . So we proceed in two steps.

**Part 1:** Decomposition of measure  $\widehat{\nu}$  in the absolutely continuous and singular parts with respect to  $\nu$ .

Let  $\rho = \widehat{\nu} + \nu$ . The measure  $\rho$  is finite and  $\int \rho = C := \int (\widehat{\nu} + \nu)$ . Both  $\widehat{\nu}$  and  $\nu$  are  $\ll \rho$ . Define the Radon-Nikodym derivatives of these measures restricted to the  $\sigma$ -fields  $\mathcal{F}_n$ .

$$f_n = \left. \frac{d\widehat{\nu}}{d\rho} \right|_{\mathcal{F}_n},$$

$$g_n = \left. \frac{d\nu}{d\rho} \right|_{\mathcal{F}_n},$$

and let  $f = \limsup f_n \geq 0$ ,  $g = \limsup g_n \geq 0$ . Note that  $f_n + g_n \leq 1$  and therefore  $f + g \leq 1$ .

One can check that  $f_n + g_n$  is a martingale with respect to filtration generated by  $\mathcal{F}_k$ ,  $k \leq n$ , and measure  $\rho$ , and therefore it converges  $\rho$ -almost surely to

$f + g$ . Moreover, it is an  $L^1(d\rho)$ -bounded martingale, hence the convergence of expectations also holds and  $\mathbb{E}(f + g) := \int (f + g) d\rho = \int d(\widehat{\nu} + \nu) = C$ . Since we know that  $\int d\rho = C$  and that  $f + g \leq 1$ , it follows that  $\rho[f = g = 0] = 0$  and therefore  $\rho$ -almost surely

$$\frac{f}{g} = \frac{\lim f_n}{\lim g_n} = \lim \frac{f_n}{g_n} = \lim \frac{d\widehat{\nu}}{d\nu} \Big|_{\mathcal{F}_n} = \lim W_n = W, \quad (7.2)$$

Then,  $\rho$ -almost surely we have

$$\begin{aligned} \widehat{\nu} &= \mathbb{1}_{\{W \neq \infty\}} \widehat{\nu} + \mathbb{1}_{\{W = \infty\}} \widehat{\nu} \\ &= \mathbb{1}_{\{W \neq \infty\}} (f\rho) + \mathbb{1}_{\{W = \infty\}} \widehat{\nu} \\ &= \mathbb{1}_{\{W \neq \infty\}} (Wg\rho) + \mathbb{1}_{\{W = \infty\}} \widehat{\nu} \\ &= \mathbb{1}_{\{W \neq \infty\}} (W\nu) + \mathbb{1}_{\{W = \infty\}} \widehat{\nu} \end{aligned}$$

where in the second line we use that  $\widehat{\nu} \ll \rho$  with the Radon-Nikodym derivative  $f$ , the third line uses (7.2), and the fourth line uses that  $g$  is the Radon-Nikodym of  $\nu$  with respect to  $\rho$ .

We have seen that  $W$  is finite  $\nu$ -almost surely, hence we can re-write the last expression as the following identity (valid up to the sets that have measure 0 both with respect to  $\nu$  and  $\widehat{\nu}$ ),

$$\widehat{\nu} = W\nu + \mathbb{1}_{\{W = \infty\}} \widehat{\nu}. \quad (7.3)$$

**Part 2:** If  $\widehat{\nu} \ll \nu$  then  $W < \infty$ ,  $\widehat{\nu}$ -surely. (We already know that  $W < \infty$ ,  $\nu$ -surely). If  $W < \infty$ ,  $\widehat{\nu}$ -almost surely, then by (7.3),  $\widehat{\nu} = W\nu$  and therefore  $\int W\nu = \int \widehat{\nu}$ . Finally, if  $\int W\nu = \int \widehat{\nu}$ , then (7.3) implies that  $W < \infty$ ,  $\widehat{\nu}$ -almost surely and  $\widehat{\nu} \ll \nu$  with Radon-Nikodym derivative  $W$ .

If  $\widehat{\nu}$  and  $\nu$  are mutually singular, then (7.3) implies that  $W = \infty$ ,  $\widehat{\nu}$ -almost surely. Then if  $W = \infty$ ,  $\widehat{\nu}$ -almost surely, then, integrating (7.3), we find that  $\int W\nu = 0$ . Finally, if  $\int W\nu = 0$  then (7.3) implies that  $\mu = \mathbb{1}_{\{W = \infty\}} \widehat{\nu}$ , hence  $W = \infty$ ,  $\widehat{\nu}$ -almost surely. Since  $W < \infty$ ,  $\nu$ -almost surely, this implies that  $\widehat{\nu}$  and  $\nu$  are mutually singular. □

In our application we will set  $\widehat{\nu} = \widehat{GW}$  and  $\nu = GW$ . The  $\sigma$ -algebras  $\mathcal{F}_n$  are the algebras of cylindrical sets of trees that depends only on the generations at or below  $n$ . The random variables  $W_n = X_n/\mu$ , as can be seen from equation (7.1).

*Proof of Theorem 7.1.5.* Consider those vertices in the  $n$ -th generation of the size-biased GW tree which are off the distinguished path. Their number follows the Galton-Watson process with immigration. The immigration process is given by  $Y_n = \widehat{L}_n - 1$  since all except one vertices born from the distinguished particle are immigrant particles. Hence we can apply Seneta Theorem 7.2.1.

This theorem says, in particular, that if  $\mathbb{E} \log^+(\widehat{L} - 1) < \infty$  then  $W_n \rightarrow W < \infty$  almost surely, and  $\mathbb{E} \log^+(\widehat{L} - 1) = \infty$ , then  $W_n \rightarrow W = \infty$  a.s.

Note that

$$\mathbb{E} \log^+(\widehat{L} - 1) = \frac{1}{\mu} \sum_{k=2}^{\infty} \log(k-1) k p_k.$$

Hence,  $\mathbb{E} \log^+(\widehat{L} - 1) < \infty$  if and only if  $\mathbb{E}(L \log^+ L) < \infty$ .

This establishes, that the Radon-Nicodym derivative of  $\widehat{GW}$  with respect to  $GW$  is bounded  $\widehat{GW}$ -almost surely if and only if  $\mathbb{E}(L \log^+ L) < \infty$ .

By Lemma 7.4.2, we find that under the measure  $GW$ ,  $\mathbb{E}W = 1$  if  $\mathbb{E}(L \log^+ L) < \infty$  and  $\mathbb{E}W = 0$  if  $\mathbb{E}(L \log^+ L) = \infty$ . This completes the proof of the theorem.  $\square$

## 7.5 Critical Case

The results in the following theorem were obtained by Kolmogorov (1938) and Yaglom (1947) under the assumption  $\mathbb{E}(L^3) < \infty$ . In the more general case, given here, it was established by Kesten, Ney, and Spitzer (1966).

**Theorem 7.5.1.** *Suppose that  $\mu = 1$  and  $\sigma^2 := \text{Var}(L) = \mathbb{E}(L^2) - 1$ . Then we have*

1. **Kolmogorov's Estimate:**  $\lim_{n \rightarrow \infty} n \mathbb{P}[X_n > 0] = 2/\sigma^2$ ;
2. **Yaglom's limit law:** *If  $\sigma < \infty$ , then the conditional distribution of  $X_n/n$  given  $X_n > 0$  converges as  $n \rightarrow \infty$  to an exponential law with mean  $\sigma^2/2$ .*

Interestingly, according to Lyons and Peres book, the case  $\sigma = \infty$  in the second statement appears to be open.

We will prove only the first part of this theorem. The proof of the second part is more difficult and the details given in the Lyons-Peres book are involved and not clear enough to me. So this proof is skipped.

In order to establish this theorem we want to show that conditional on the non-extinction, critical and subcritical Galton-Watson trees have the distribution of a corresponding size-biased Galton-Watson tree.

This is based on the following lemma, in which  $L$  is the offspring distribution and  $H_i^{(n)}$  is the event that the descendants of child  $i$ ,  $1 \leq i \leq L$ , are not extinct at generation  $n$ .

**Lemma 7.5.2.** *Let  $L$  be a random variable taking non-negative integer values with distribution  $\mathbb{P}(L = k) = p_k$ , and  $0 < \mathbb{E}L < \infty$ . Suppose that given  $L$ , events  $H_1^{(n)}, \dots, H_L^{(n)}$  are independent and have probability  $h_n \rightarrow 0$  as  $n \rightarrow \infty$ . Let*

$$Y_n(\omega) = \sum_{i=1}^L \mathbb{1}_{H_i^{(n)}}(\omega),$$



the (random) number of events  $H_i^{(n)}$  that occur. Then conditional on  $Y_n > 0$ , and asymptotically almost surely,

1. only one event  $H_i^{(n)}$  occurs:

$$\lim_{n \rightarrow \infty} \mathbb{P}[Y_n = 1 | Y_n > 0] = 1;$$

2. the law of  $L$  is that of the size-biased r.v.  $\widehat{L}$ :

$$\lim_{n \rightarrow \infty} \mathbb{P}[L = k | Y_n > 0] = kp_k / \mathbb{E}L;$$

3. every of the events  $H_i^{(n)}$  has the same probability to occur:

$$\lim_{n \rightarrow \infty} \mathbb{P}[H_i^{(n)} | Y_n > 0, L = k] = 1/k.$$

for  $1 \leq i \leq k$ .

*Proof.* We have

$$\mathbb{P}[Y_n = 1 | Y_n > 0, L = k] = \frac{kh_n(1-h_n)^{k-1}}{\sum_{s=1}^k \binom{k}{s} h_n^s (1-h_n)^{k-s}}.$$

Using the notation  $t_n = 1 - h_n$ , we re-write this as

$$\frac{k(1-t_n)t_n^{k-1}}{1-t_n^k} = \frac{kt_n^{k-1}}{1+t_n+\dots+t_n^{k-1}} \geq t_n^{k-1}.$$

Hence,

$$\mathbb{P}[Y_n = 1 | Y_n > 0] \geq \sum_{k=1}^{\infty} p_k t_n^{k-1} \rightarrow 1,$$

as  $n \rightarrow \infty$  and therefore  $t_n \rightarrow 1$ . This proves (1).

By using this, we have

$$\lim_{n \rightarrow \infty} \mathbb{P}[L = k | Y_n > 0] = \lim_{n \rightarrow \infty} \mathbb{P}[L = k | Y_n = 1],$$

and by the Bayes rule,

$$\begin{aligned} \mathbb{P}[L = k | Y_n = 1] &= \frac{\mathbb{P}[Y_n = 1 | L = k] \times \mathbb{P}[L = k]}{\mathbb{P}[Y_n = 1]} \\ &= \frac{kp_k h_n (1-h_n)^{k-1}}{\sum_{j=1}^{\infty} jp_j h_n (1-h_n)^{j-1}}. \end{aligned}$$

By using the monotone convergence, we can take the limit  $(1 - h_n) \rightarrow 1$  under the summation sign and get

$$\mathbb{P}[L = k | Y_n = 1] = \frac{kp_k}{\sum_{j=1}^{\infty} jp_j},$$

which proves (2).

Then, again by using (1), we have

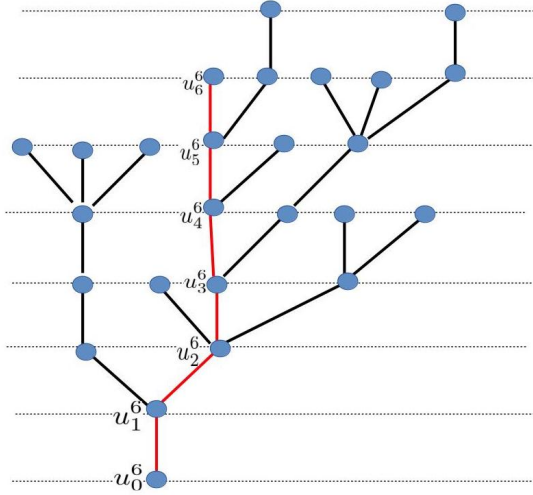
$$\lim_{n \rightarrow \infty} \mathbb{P}[H_i^{(n)} | Y_n > 0, L = k] = \lim_{n \rightarrow \infty} \mathbb{P}[H_i^{(n)} | Y_n = 1, L = k] = 1/k,$$

where the last equality is obvious by the symmetry of the events  $H_i^{(n)}$ .  $\square$

*Proof of Kolmogorov's estimate in Theorem 7.5.1.* Since  $\mathbb{E}(X_n) = \mathbb{E}(X_n | X_n > 0) \mathbb{P}(X_n > 0)$ , we have

$$\mathbb{E}(X_n | X_n > 0) = \frac{\mathbb{E}(X_n)}{\mathbb{P}(X_n > 0)} = \frac{1}{\mathbb{P}(X_n > 0)}.$$

The idea is to show that  $\mathbb{E}(X_n | X_n > 0) \sim \frac{\sigma^2}{2} n$  by decomposing this variable into a sum of descendants from different special vertices.



**Figure 7.4:** A Galton-Watson tree

Let  $u_n^n$  be the left-most individual in generation  $n$  when  $X_n > 0$ , and let its ancestors be  $u_{n-1}^n, \dots, u_0^n$ . (See Figure 7.4 for an example.) Let  $Y_i$  be the number of children of  $u_i^n$  that are to the right of  $u_{i+1}^n$ . In the example we have  $Y_0 = Y_1 = 0, Y_2 = Y_3 = Y_4 = Y_5 = 1$ . Then, let  $Y'_i$  be the number of descendants of  $u_i^n$  in generation  $n$ , which are not descendants of  $u_{i+1}^n$ . In the example,  $Y'_0 = Y'_1 = Y'_2 = 0, Y'_3 = 3, Y'_4 = 0, Y'_5 = 1$ .

Clearly,  $X_n = 1 + \sum_{i=0}^{n-1} Y'_i$ . In addition,  $\mathbb{E}(Y'_i | X_n > 0) = \mathbb{E}(Y_i | X_n > 0)$  since the children of  $u_i^n$  generate an independent Galton-Watson tree. The condition  $X_n > 0$  has no effect on the distribution of  $Y'_i$  or  $Y_i$  because it is satisfied by the existence of the vertex  $u_n^n$ .

Hence,

$$\mathbb{E}(X_n | X_n > 0) = 1 + \sum_{i=0}^{n-1} \mathbb{E}(Y_i | X_n > 0).$$

For the right-hand side we note that as  $n$  grows, by Lemma 7.5.2 the distribution of  $Y_i$  given  $X_n > 0$  (recall,  $Y_i$  is the number of children of  $u_i^n$  that are to the right of  $u_{i+1}$ ) tends to the distribution which is uniform on the set  $[0, \dots, \widehat{L} - 1]$ , conditional on  $\widehat{L}$ . Hence, we can calculate  $\lim_{n \rightarrow \infty} \mathbb{E}(Y_i | X_n > 0)$  as  $\mathbb{E}[(\widehat{L} - 1)/2] = \sigma^2/2$ , where we used

$$\mathbb{E}\widehat{L} = \sum_{k=1}^{\infty} k \frac{kp_k}{\mathbb{E}L} = \sigma^2 + 1,$$

since  $\mathbb{E}L = 1$ . (The passage to the limit of expectations from the limit of distributions needs a justification, which can be found in Lyons-Peres book.)

It follows that

$$\begin{aligned} \frac{1}{\mathbb{P}(X_n > 0)} &= 1 + \sum_{i=0}^{n-1} \mathbb{E}(Y_i | X_n > 0). \\ &= 1 + n(\sigma^2/2 + o(1)), \end{aligned}$$

which implies Kolmogorov's estimate  $\lim_{n \rightarrow \infty} nP(X_n > 0) = 2/\sigma^2$ .  $\square$

It also worthwhile to note that in the subcritical case  $\mu < 1$ , the random variable  $X_n |_{X_n > 0}$  converges almost surely to a random variable (without additional normalization by  $n^{-1}$ ). This is Yaglom's theorem, and it can be proved by an analogous method (see Geiger 1999 Journal of Applied Probability. "Elementary proofs of classical theorems about the Galton Watson trees".) The conditions that ensure that this random variable has finite mean were given by Heathcote.

Here is another interesting (and surprising) result about critical Galton - Watson trees. Let  $G_n$  be the generation of the most recent common ancestor of all particles in generation  $n$ .

**Theorem 7.5.3** (Zubkov). *Suppose  $\mathbb{E}L = 1$  (the process is critical) and  $\mathbb{E}L^2 < \infty$ , then*

$$\lim_{n \rightarrow \infty} \mathbb{P}(n^{-1}G_n \leq u | X_n > 0) = u,$$

for  $u \leq 0 \leq 1$ .

In other words, the generation of the common ancestor is distributed approximately uniformly between 0 and  $n$ .

The proof can be found in Geiger's paper.

## 7.6 Simply generated trees

### 7.6.1 Definition and main properties

Random trees are studied extensively in combinatorics. Here we compare the models from combinatorics with Galton-Watson trees. The material in this section are based on Flajolet - Sedgewick book and lecture notes by S. Janson.

Two typical models are *binary rooted planar trees* and *general rooted planar trees*.

A planar rooted tree is a tree imbedded in a plane with one marked vertex (root). They are equivalent if we can find an oriented homeomorphism of the plane that move one tree to another, with the root moved to the root. A vertex of the binary tree has zero or two children and a vertex of a general tree can have an arbitrary number of children.

Typically, one fixes the number of vertices and considers a tree taken uniformly at random in a given class.

Moon and Meir suggested a class of more general models for trees, which they called *simply generated families* of trees.

Recall that the out-degree of a node in a rooted tree is its number of children, that is, the number of edges that lead away from the root. Let  $c_0 = 1, c_1, c_2, \dots$  be a sequence of non-negative numbers (“weights”), let the weight  $w(v)$  of a vertex  $v$  be equal to  $c_{d(v)}$ , where  $d(v)$  is its out-degree, and let the weight of a tree  $T$  be defined as

$$w(T) = \prod_{v \in T} w(v) = \prod_{d=0}^{\infty} c_d^{N_d(T)},$$

where the last product is over all possible out-degrees  $d$ , and  $N_d(T)$  denote the number of vertices of degree  $d$  in the tree  $T$ .

**Definition 7.6.1.** A *simply-generated random tree* is the tree in  $\mathcal{T}_n$  drawn at random with probability  $w(T)/y_n$ , where

$$y_n = \sum_{T \in \mathcal{T}_n} w(T),$$

and the sum is over all planar rooted trees with  $n$  vertices.

For example, if  $c_0 = c_2 = 1$  and all other  $c_i = 0$ , then  $y_n$  is the number of binary planar trees with  $n$  vertices, and if  $c_i = 1$  for all  $i$  then this is the number of all planar rooted trees with  $n$  vertices. The first model is called the *random binary planar tree* and the second, – *the random general planar tree*.

Moon and Meir studied the *profile* of the simply generated random trees  $\mu_n(k)$ , which is the number of vertices in generation  $k$ , when the total number of vertices  $n$  is large.

They found that the average height of a vertex in a simply generated tree is  $A\sqrt{n}$ , for some positive constant  $A$  and that if  $k = O(\sqrt{n})$  then the profile

behaves according to the Rayleigh distribution,

$$\mathbb{E}\mu_n(k) \sim Ak \exp\left(-\frac{k^2}{2n/A}\right).$$

with the mean proportional to  $k$ . In particular for initial generations of a random binary tree, the growth of population is linear (proportional to  $k$ ).

Moon and Meir use the method of generating functions in their study.

Now what is the relation of these trees with Galton-Watson trees?

If  $\sum_{i=0}^{\infty} c_i = C < \infty$ , then the sequences of weights  $c_i$  can be scaled so that it becomes a probability distribution  $p_i = c_i/C$ . This will not change the relative weight of the trees in  $\mathcal{T}_n$ , since every weight will be simply multiplied by  $C^n$ , where  $n$  is the number of vertices in the tree. In this case  $w(T)$  becomes a probability of a Galton-Watson tree with the offspring distribution  $p_i$ . However, note that we condition this Galton-Watson measure on trees by requiring that the tree has the total progeny equal to  $n$ . Hence in this case the simply generated random tree can be identified with a conditioned Galton-Watson tree.

For example, the random planar binary trees can be described in this way, by setting  $p_0 = p_2 = 1/2$ .

What if  $\sum_{i=0}^{\infty} c_i = \infty$ , as, for example, the situation for the general planar trees?

**Theorem 7.6.2.** *Let the probability on the set of trees  $\mathcal{T}_n$  is given by the weights  $c_k$ ,  $k = 0, 1, \dots$ . Suppose that the series  $\theta(t) = \sum_{k=0}^{\infty} c_k t^k$  has a positive radius of convergence so that  $\sum_{k=0}^{\infty} c_k \rho^k < \infty$  for some  $\rho > 0$ . Then this probability distribution coincides with the probability distribution of a Galton-Watson tree conditioned to have  $n$  vertices.*

*Proof.* We use a modification of the weight sequence which will not change the relative weight of the trees, and so it will not change the probability distribution of the trees. Namely, let  $\bar{c}_i := at^i c_i$ , where  $a$  and  $t$  are parameters. This means that a vertex of out-degree  $d$  now has weight  $at^d c_d$ . After multiplying over all vertices we get an additional factor  $a^n t^{\#\text{edges in } T} = a^n t^{n-1}$ . In other words, for these weights, we have

$$\begin{aligned} \bar{w}(T) &= \prod_{d=0}^{\infty} (\bar{c}_d)^{N_d(T)} = a^n t^{n-1} \prod_{d=0}^{\infty} c_d^{N_d(T)}, \\ \bar{y}_n &= a^n t^{n-1} y_n, \end{aligned}$$

and therefore, the probabilities of trees are the same as before weight modification,  $\bar{w}(T)/\bar{y}_n = w(T)/y_n$ .

Since the new weights are  $\bar{c}_i = ac_i t^i$ , therefore by our assumption on the series  $\sum_{k=0}^{\infty} c_k t^k$ , we can find a modification of weights such that  $\sum \bar{c}_i = 1$ . Then the corresponding family of random trees is the same as that of the Galton-Watson trees with probabilities of offspring  $p_i = \bar{c}_i$  and conditioned to have  $n$  vertices.  $\square$

In fact, since we have two parameters,  $a$  and  $t$ , at our disposal, we can also target a specific mean for the offspring distribution in the Galton-Watson tree. In particular we can target  $\mathbb{E}L = 1$ . If it is possible, the simply generated random tree is equivalent to a critical random Galton-Watson tree conditioned on the total number of progeny being equal to  $n$ .

So, for example, a random general planar tree corresponds to the sequence  $c_i = 1$  for all  $i$ . By the original definition, this is simply a tree chosen uniformly at random from all general planar trees of size  $n$ . If we use the weight transformation with  $a = t = 1/2$ , we find that this random tree corresponds to a random Galton-Watson tree with the geometric offspring distribution with parameter  $p = 1/2$ , conditioned on having  $n$  vertices.

For the critical Galton-Watson trees, we already know that conditional on non-extinction in generation  $h \rightarrow \infty$ , the initial portion of the tree converges to the size-biased random tree. It turns out that the same claim holds if we condition on the total progeny of a critical tree  $n$  and let  $n$  go to  $\infty$ . The initial portion of the tree converges to the size-biased random tree.

*Example 7.6.3.* Consider a random planar tree with large  $n$ . The convergence to the size-biased tree allows us identify (with large probability) in  $k$ -th generation a distinguished vertex, the ancestor of the vertices in the most distant generation. This vertex is likely to have offspring distributed as the size-biased distribution  $\bar{p}_i = i(1/2)^i$  while all other vertices in this generation will have the out-degree with geometric distribution  $p_i = (1/2)^i$ .

Now, supposing that we can re-weight the original weights  $c_i$  to a sequence of probability weights, is it always possible to ensure that the resulting probability distribution has mean 1?

The answer is “no”.

Note that we can use the transformation  $c_i \rightarrow ac_i t^i$  and obtain a probability distribution only if  $t < \rho$  where  $\rho$  is the radius of convergence for the series  $\theta(t) = \sum_{i=0}^{\infty} c_i t^i$ , of if  $t = \rho$  and  $\theta(\rho) < \infty$ .

If we do the transformation and the result is the probability sequence, then we must have  $a = 1/\theta(t)$ . So, let the probabilities after the transformation be  $p_k = c_k t^k / \theta(t)$ . Then the off-spring distribution of the corresponding Galton-Watson tree has the mean

$$\mathbb{E}L = \sum_{k=0}^{\infty} k p_k = \frac{1}{\theta(t)} \sum_{k=0}^{\infty} k c_k t^k = t \frac{\theta'(t)}{\theta(t)}.$$

**Lemma 7.6.4.** *The function  $\psi(t) = t\theta'(t)/\theta(t)$  is increasing on  $[0, \rho)$ .*

*Ex. 7.6.5.* Prove this lemma.

Define

$$\nu = \lim_{t \uparrow \rho} t \frac{\theta'(t)}{\theta(t)}.$$

Then it is clear that we can choose a re-weighting transformation with the critical offspring distribution, only if  $\nu \geq 1$ .

The point is that for certain choices of weights  $\nu < 1$ . The trees in which this situation is realized are called *non-generic random trees*. It can only happen if the radius of convergence  $\rho < \infty$ .

*Example 7.6.6* (Non-generic random tree).

Let  $c_k = 1/k^\beta$  for  $k > 0$  and  $\beta > 2$ . Let  $c_0 = 1 + \delta$ , where we choose  $\delta$  later. Then  $\rho = 1$  and  $\theta(1) = \zeta(\beta) + \delta < \infty$ . (Here  $\zeta(z)$  denote the Riemann zeta function.) In addition

$$\nu = \psi(1) = \frac{\zeta(\beta - 1)}{\zeta(\beta) + \delta}.$$

By a suitable choice of  $\delta$  this can be made smaller than 1. In this situation, the tree cannot be represented as a critical Galton-Watson tree.

From the arguments in the previous section we know that if we condition a critical or a subcritical tree on the existence of surviving vertices in the  $n$ -th generation, and let  $n \rightarrow \infty$ , then the limit is the size-biased Galton-Watson tree. Note, however, that when we study random trees from  $\mathcal{T}_n$ , we are conditioning not on the existence of the vertices in the  $n$ -th generation but rather on the total size of the tree. It turns out that for generic trees, this does not make much difference in the limit, and the limit tree is the same, – the size-biased critical GW tree.

A surprise comes with non-generic trees. It turns out that for them the limit is different.

Suppose that we investigate non-generic trees  $T_n$  with the offspring distribution  $\{p_k\}$ , such that  $\mu < 1$ . Then the limit tree can be described as follows. The vertices are again divided into two classes, normal and special, and in one generation there can be no more than one special vertex. However, now in some generations there are no special vertices.

The root vertex is special. The offspring of every special vertex is distributed according to the following distribution:

$$\bar{L} = \begin{cases} k, & \text{with probability } kp_k \\ \infty, & \text{with probability } 1 - \mu, \end{cases}$$

where  $\mu := \mathbb{E}L = \sum_{k=0}^{\infty} kp_k$ .

If the offspring of a special vertex is finite, then one of the offspring is randomly chosen as the special vertex in the next generation. However, if the offspring is infinite, then none of the children is chosen as special. All vertices in the offspring are normal. The offspring of every normal vertex is distributed according to the usual Galton-Watson law and consists of normal vertices.

With probability 1, this limit random tree has only a finite path of special vertices, and at the end of this tree there is an explosion.

## 7.6.2 The main convergence theorem about simply generated trees

### Definition of convergence of rooted ordered trees

It is convenient to consider all trees as subsets of one very large non-locally finite tree which is called the *Harris-Ulam tree*  $T_{HU}$ . Namely, consider the alphabet of natural numbers  $\mathbb{N} = \{1, 2, \dots\}$ , and the space  $V_\infty$  of all finite strings on this alphabet, including the empty string  $\emptyset$ . For example, a typical element is a sequence  $(100, 1, 10, 2)$ . These are the vertices of the tree  $T_{HU}$ .

The root of this tree is the empty string  $\emptyset$ . The edges connect vertices  $i_1, \dots, i_k$  and  $i_1, \dots, i_k, i_{k+1}$ . This gives  $T_{HU}$  the structure of a connected graph which is in fact a tree. Since the tree has a root, we can talk about parent-child relation between vertices. For example the vertex  $(100, 1, 10)$  is the parent of the vertex  $(100, 1, 10, 2)$ .

The embedding  $\varphi$  of a planar tree  $T$  in  $T_{HU}$  is defined recursively. The root of  $T$  is mapped to  $\emptyset$ . If  $v$  is mapped to  $(i_1, \dots, i_k)$  and has  $s$  children  $v_1, \dots, v_s$  (which are ordered by the definition of a planar tree), then  $v_j$  is mapped to  $(i_1, \dots, i_k, j)$ . For example, if a vertex of  $T$  that corresponds to  $(100, 1, 10)$  has three children, then they are mapped to  $(100, 1, 10, 1)$ ,  $(100, 1, 10, 2)$ , and  $(100, 1, 10, 3)$ , respectively.

Let  $\mathcal{T}_f \subset \mathcal{T}_{lf} \subset \mathcal{T}$  be the sets of all finite (locally finite, arbitrary) rooted planar trees, respectively.

For a tree  $T$  in  $\mathcal{T}$  one can define the out-degree function on vertices of  $T_{HU}$ . Namely,  $d_T(x) = 0$  if  $x$  is not in the image of the embedding  $\varphi(T)$ , and  $d_T(\varphi(v)) = \text{out-degree of vertex } V \text{ in } T$ . Note that this function can take value  $\infty$  if the vertex  $v$  has infinite number of children.

**Definition 7.6.7.** We say that a sequence of planar trees  $T_n \in \mathcal{T}$  converges to  $T \in \mathcal{T}$ , if the functions  $d_{T_n}$  converge to  $d_T$  pointwise, that is if for every  $v \in T_{HU}$ ,

$$d_{T_n}(v) \rightarrow d_T(v).$$

Let  $T^{(m)}$  denote the subtree of  $T$  consisting of the vertices in generations  $0, \dots, m$ , that is, the truncation of tree  $T$  at height  $m$ . Then for locally finite limit trees  $T$ , the situation is simple.

**Lemma 7.6.8.** *If  $T \in \mathcal{T}$  is locally finite, then, for any sequence of trees  $T_n \in \mathcal{T}$*

$$\begin{aligned} T_n \rightarrow T &\iff T_n^{(m)} \rightarrow T^{(m)} \text{ for each } m \\ &\iff T_n^{(m)} = T^{(m)} \text{ for each } m \text{ and } n > n_0(m) \end{aligned}$$

where  $n_0(m)$  is a certain function of  $m$ .

If  $T$  is not locally finite then the second equivalence in this statement does not hold.



*Example 7.6.9.* Let  $S_n$ ,  $1 \leq n \leq \infty$ , be a star where the root have out-degree  $n$  and its children have out-degree 0, then  $S_n \rightarrow S_\infty$  but  $S_n^{(m)} \neq S_\infty^{(m)}$  for all  $n$  and  $m \geq 1$ .

Let  $V^{[m]}$  be the subset of  $V_\infty$  that consists of strings of length at most  $m$  and with string elements at most  $m$ . The  $T^{[m]}$  is the subtree of  $T$  with the vertices in  $V^{[m]}$ . That is,  $T^{[m]}$  is obtained from  $T$  by truncating at height  $m$  and pruning the tree so that all out-degrees are at most  $m$ . Then, the following result holds.

**Lemma 7.6.10.** *For any  $T \in TT$  and for any sequence of trees  $T_n \in TT$*

$$\begin{aligned} T_n \rightarrow T &\iff T_n^{[m]} \rightarrow T^{[m]} \text{ for each } m \\ &\iff T_n^{[m]} = T^{[m]} \text{ for each } m \text{ and } n > n_0(m) \end{aligned}$$

where  $n_0(m)$  is a certain function of  $m$ .

Janson notes that if the trees are random then the analogues of these two lemmas holds for the convergence in probability or in distribution. For example

$$T_n \xrightarrow{d} T \iff T_n^{[m]} \xrightarrow{d} T^{[m]} \text{ for each } m$$

For the proof, Janson refers to the methods in Aldous and Pitman 1998 "Tree-valued Markov chains" (in the case of locally finite limit tree), or more generally to Billingsley "Convergence of probability measures".

### The main convergence theorem

Let us here give a formal statement about the limit of simply generated trees. Recall that we defined  $\theta(t) = \sum_{i=0}^{\infty} c_i t^i$ , where  $c_i$  are the weights of a given family of trees, and  $\rho$  as the radius of convergence of  $\theta(t)$ . We have also defined the function  $\psi(t) = t\theta'(t)/\theta(t)$  and  $\nu = \lim_{t \uparrow \rho} \psi(t)$ . This parameter represents the largest mean of the offspring random variable  $L$  that can be obtained through re-scaling of the weight sequence.

**Theorem 7.6.11.** *Let  $(c_k)$  be any weight sequence with  $c_0 > 0$  and  $c_k > 0$  for some  $k \geq 2$ .*

1. *If  $\nu \geq 1$ , let  $\tau$  be the unique number in  $[0, \rho]$  such that  $\psi(\tau) = 1$ .*
2. *If  $\nu < 1$ , let  $\tau = \rho$ .*

*Let  $\pi_k = c_k \tau^k / \theta(\tau)$  for all  $k \geq 0$ . Then  $(\pi_k)$  is a probability distribution with expectation  $\mu = \psi(\tau) = \min\{\nu, 1\}$ , and (possibly infinite) variance  $\sigma^2 = \tau\psi'(\tau)$ . In the case (i)  $\mu = 1$  and the simply generated random tree  $T_n$  converges to the size-biased Galton-Watson tree that corresponds to distribution  $\pi$ . In the case (ii)  $\mu = \nu < 1$ ,  $T_n$  converges to the modified size-biased Galton-Watson tree that has a finite spine ending with an explosion.*

[Proof ???]

### The convergence for the degree of a random vertex

Recall that we use  $d_T(v)$  to denote the out-degree of vertex  $v$  in a rooted tree  $T$ .

**Theorem 7.6.12.** *Let  $T_n$  be a simply generated tree and let the probability distribution  $\pi$  be as defined in Theorem 7.6.11.*

1. *Let  $v$  be a uniformly random node in a random tree  $T_n$ . Then, for every  $k \geq 0$ , as  $n \rightarrow \infty$ ,*

$$\mathbb{P}[d_{T_n} = k] \rightarrow \pi_k.$$

2. *Let  $N_k(T_n)$  be the number of vertices of out-degree  $k$  in a random tree  $T_n$ . Then for every  $k \geq 0$ ,*

$$\frac{N_k(T_n)}{n} \xrightarrow{\mathbb{P}} \pi_k.$$

Note that the second statement is stronger and more useful than the first. It means that the empirical distribution of vertices in a given random tree converges to the distribution  $\pi$ . It corresponds to the quenched results in terminology of statistical physics, while (i) corresponds to an annealed result.

For generic case  $\nu > 1$  a form of this result was proven by Meir and Moon (see Flajolet - Sedgewick Proposition VII.2 on p. 460). In the presented form it is due to Jansson (Theorem 7.11). Jansson also gives a reference to an early work by Otter (1948).

### 7.6.3 Further Examples

- 1) Non-planar trees, or unordered trees.
- 2) Binary tree II
- 3) Motzkin tree
- 4)  $w_k = k!$

### 7.6.4 How to generate simply generated random trees?

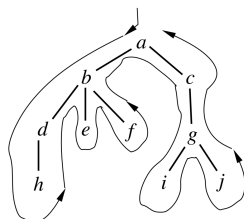
We will describe here Devroy's algorithm for sampling *generic* simply-generated trees, and we describe it below. Devroy assumes in his algorithm that the weight sequence is rescaled to a probability distribution  $\{p_k\}$  with mean 1. This is always possible to do for a finite  $n$  by setting the weights  $c_k$  equal to 0 for  $k > n$ . In particular, this algorithm works both for the generic and non-generic case.

Devroy assumes also that  $0 < \sigma^2 := \text{Var}(L) < \infty$  and shows that the algorithm is linear in  $n$  under this assumption. He explains what happens if the assumption on the variance is relaxed. He warns, however, that for  $\mathbb{E}(L) = \infty$ , the explicit results about the complexity of the algorithm are more difficult to obtain.

### Bijection with Lukasiewicz paths

First, we note that rooted planar trees are in bijection with a special kind of random walk paths, the *Lukasiewicz paths*.

The pictures and discussion below are taken from Flajolet-Segdewick book “Analytical Combinatorics”

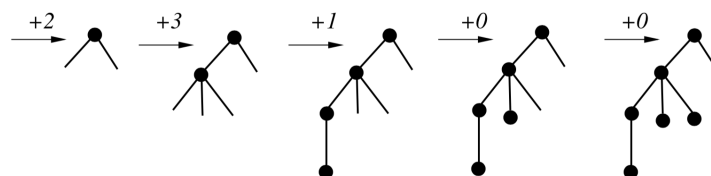


**Figure 7.5:** The depth - first exploration of a tree

Every plane tree can be traversed by starting at the root, proceeding depth-first and left-to-right and backtracking to the root once a sub-tree has been completely explored. For example, in the tree in Figure 7.5 the *first* visits to vertices take place in the following order:  $(a, b, d, h, e, f, c, g, i, j)$ .

If we replace the vertex in this order with the out-degree of this vertex, than we obtain the *Lukasiewicz code* of the tree. For the example in Figure 7.5, the code is

$$\sigma = (2, 3, 1, 0, 0, 0, 1, 2, 0, 0).$$

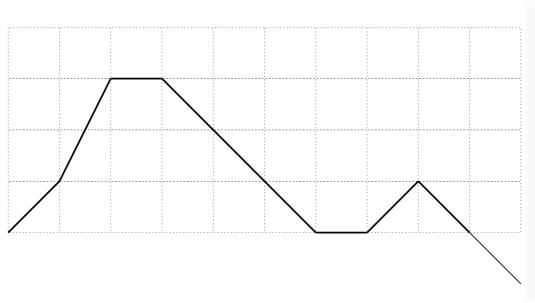


**Figure 7.6:** Reconstruction of a tree from its Lukasiewicz code

One can check that the code determines the tree unambiguously. Given a code, one reconstruct the tree step by step adding vertices one after the other in the left-most available place. For our example, the first steps in this process are illustrated in Figure 7.6.

Lukasiewicz introduced these codes to describe the order of evaluation of logical expressions. These codes are also a basic instrument in development of parsers and compilers in computer science.

The Lukasiewicz codes can be represented as paths on the discrete lattice  $\mathbb{Z} \times \mathbb{Z}$  where the  $j$ -th element of the code  $\sigma_j$  corresponds to the path displacement  $(1, \sigma_j - 1)$ . For our example, we have the path shown in Figure 7.7.



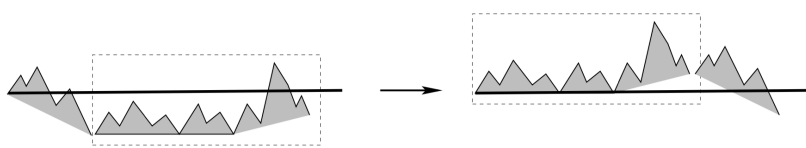
**Figure 7.7:** A Lukasiewicz path

In this way, we obtain a bijection between the planar rooted trees with  $n$  vertices and lattice paths from  $(0, 0)$  to  $(n, -1)$  where the steps have sizes in the set  $\{-1, 0, 1, 2, \dots\}$ , and such that the path does not go below the horizontal axis except at the last step. Such lattices paths are called *Lukasiewicz paths*

So, in order to generate a random planar tree  $T_n$  according to the distribution  $w(T)$ , we need to generate a Lukasiewicz path  $P_n$  with  $n$  steps, according to the distribution  $w(P)$  that assigns weights  $c_i$  to increments with value  $i - 1$ .

### Cycle lemma

For the purposes of path generation, it is somewhat inconvenient that the path is restricted to be always above the horizontal axis except at the last step. So consider the class of *relaxed Lukasiewicz paths* that start at  $(0, 0)$ , finish at  $(n, -1)$  and have step sizes in the set  $\{-1, 0, 1, 2, \dots\}$ . However, they are allowed to take negative values in between times 0 and  $n$ .



**Figure 7.8:** A example of the Vervaat Transform

Every relaxed Lukasiewicz path can be transformed to a regular Lukasiewicz path by a certain transformation. We call it the *Vervaat transform* since it is analogous to a similar transform in the theory of Brownian motion introduced by Wim Vervaat. It works as follows. Let  $x_1, x_2, \dots, x_n$  be the steps of a relaxed Lukasiewicz path, and  $s_j = \sum_{i=1}^j x_i$  is the vertical coordinate of the path at time  $j$ . Let  $\hat{j}$  is the first time when the path reaches its minimum value. Then the Vervaat transform of the original path is given by the step sequence  $(x_{\hat{j}+1}, \dots, x_n, x_1, \dots, x_{\hat{j}})$ . An example is shown in Figure 7.8.

**Lemma 7.6.13** (Dvoretzky-Milman Cycle Lemma). *The Vervaat transform*

is an  $n$ -to-1 map of the set of relaxed Lukasiewicz path to the set of regular Lukasiewicz paths.

It is also clear that the Vervaat transform preserves the number of steps that have a given size. It remains to generate relaxed Lukasiewicz paths according to distribution that gives a path  $P$  the weight  $w(P)$ , where

$$w(P) = c_0^{n_0} c_1^{n_1} \dots c_k^{n_k} \dots,$$

if the path contains  $n_j$  steps of size  $j - 1$ , and we also have the conditions

$$\begin{aligned} n_0 + n_1 + \dots &= n, \\ n_0 \times (-1) + n_1 \times (0) + n_2 \times 1 + \dots + n_k \times (k - 1) + \dots &= -1. \end{aligned}$$

Adding the first condition to the second, we obtain:

$$n_1 + 2n_2 + 3n_3 + \dots = n - 1.$$

By assumption,  $c_i = p_i$  form a probability distribution, so we have a problem of getting a sample  $(n_0, n_1, n_2, \dots)$  from the multinomial distribution with parameters  $n$  and  $(p_0, p_1, p_2, \dots)$  and an additional linear condition  $\sum_{k=1}^{\infty} kn_k = n - 1$ .

### 7.6.5 Sampling from the multinomial distribution

It is easy to sample from the multinomial distribution since its marginal and conditional distributions are binomial and multinomial distributions respectively. First, we sample  $n_0$  as a binomial random variable with parameters  $(n, p_0)$ , then we sample  $n_1$  as a binomial r.v. with parameters  $[n - n_0, p_1/(1 - p_0)]$ , then we sample  $n_2$  as a binomial r.v. with parameters  $[n - (n_0 + n_1), p_2/(1 - (p_0 + p_1))]$ , and so on, until we obtain  $n_K$  such that  $n_1 + \dots + n_K = n$ .

According to Devroye, a binomial random variable can be generated in constant time independent of  $n$  and  $p$ . Therefore, the generation of the sequence  $(n_0, \dots, n_K)$  will on average take time proportional to the expectation of  $K$  above. If the distribution of  $L$  has finite support then this expectation is constant. Otherwise, we need to evaluate the expectation of the maximum of  $n$  independent copies of  $L$  and according to Devroye, this quantity is bounded by  $o(n^{1/\rho})$  if  $\mathbb{E}(L^\rho) < \infty$ . (See the argument in Devroye's paper.)

It follows that the time to generate the sequence  $(n_0, \dots, n_K)$  is  $O(n^{1/2})$  for  $L$  with finite variance.

Note, however, that the generated sequence of  $n_k$  might fail to satisfy an additional condition  $\sum_k kn_k = n - 1$ . In order to satisfy this condition, Devroye suggests repeating the procedure until the condition is satisfied.

Note that

$$\mathbb{E} \sum_k kn_k = \sum_k kp_k n = \mu n.$$

For  $\mu = 1$  this is close to the target condition for  $\sum_k kn_k$ .

Also note that this sum can also be written as  $\sum_{i=1}^n L_i$  where  $L_i$  are i.i.d random variables distributed as  $L$ . Hence one can use theorems about sums of independent random variables. If  $\mu = 1$  and  $\text{Var}(L) < \infty$ , then local limit laws hold and the probability of the sum to hit  $n - 1$  is at least  $cn^{-1/2}$  unless some condition on the parity of the sum interferes (Devroye refers here to Petrov's book.)

Hence, on average one needs  $O(n^{1/2})$  randomly generated sequences to obtain one sequence  $(n_0, \dots, n_K)$  that satisfies the condition  $\sum_k kn_k = n - 1$ . Generation of one sequence takes time  $O(n^{1/2})$ . Overall, this gives  $O(n)$  time to generate a good sequence  $(n_0, \dots, n_K)$ .

### Overall description of the sampling algorithm

So, first we generate the sequence  $(n_0, \dots, n_K)$  that satisfies the condition  $\sum_k kn_k = n - 1$ . Then we form a sequence

$$(-1, -1, \dots, -1, 0, \dots, 0, 1, \dots, 1, 2, \dots, 2, \dots, K - 1, \dots, K - 1).$$

where  $-1$  is repeated  $n_0$  times,  $0$  repeated  $n_1$ , and more generally, the value  $k - 1$  is repeated  $n_k$ .

After that we permute randomly this sequence and apply the Vervaat transform. This can be done in  $O(n)$  time. Finally, we map the resulting Lukasiewicz path to the corresponding tree. This again takes no more than  $O(n)$  time.

Hence, if  $\mu = 1$  the overall running time is linear  $O(n)$ .

### Extensions

What do we do if  $\mathbb{E}L < 1$  and the sequence of weights cannot be re-weighted to the critical case  $\mathbb{E}L = 1$ ? In other words, how do we generate non-generic trees?

In fact for any fixed  $n$ , the probabilities  $p_k$  with  $k \geq n$  are irrelevant after we condition on the progeny size equal to  $n$ . Hence, for every fixed  $n$ , we can use a truncated probability distribution for  $L$ , where  $p_k = 0$ , if  $k \geq n$ . This probability distribution can be conjugated (that is, re-weighted) to a critical case and therefore, we can sample from this family of trees.

The re-weighting scheme is changing from time to time and this is what causes the limit to be different from the case of generic trees.

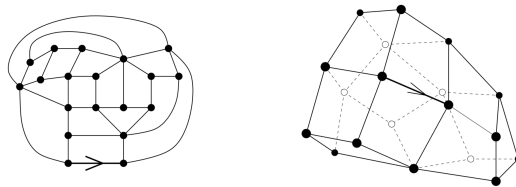
## Chapter 8

# Random planar maps

### 8.1 Planar maps and Quadrangulations

[This section is based on the Chassaing-Schaeffer paper.]

Planar maps are planar graphs with a fixed imbedding in a plane (or sphere). One can also consider graphs embedded in surfaces of higher genus. A priori, loops and multiple edges are allowed in a planar map. We assume here that a planar map is rooted. That is, it has a *root*, which is a distinguished edge with a specified orientation. The starting vertex of the root is called the *root vertex*. The face which stays on the right when we move along the root edge is called the *root face* or *outer face*. Two maps are identical if there is a homeomorphism of the plane that sends one map onto another (roots included).

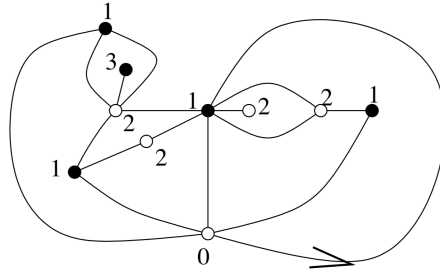


**Figure 8.1:** A planar quadrangulation, in planar and spherical representation; picture from Chassaing and Schaeffer

One particular case of planar maps consists of triangulations, another one, which we consider here consists of quadrangulations. These are planar maps with 4-regular faces. See Figure 8.1.

The 4-faces are allowed to be degenerate, in the sense that they can have as its boundary two edges from  $u$  to  $v$  and another edge in the region bounded by these two edges. See an example of these faces in Figure 8.2.

One can check that *the quadrangulations are necessary bipartite*, that is, its vertices can be colored in *two colors*, so that no two vertices of two colors are connected by an edge. In particular, this implies that there can be no loops in



**Figure 8.2:** A planar quadrangulation, with vertices labelled by minimal distance from the root vertex.

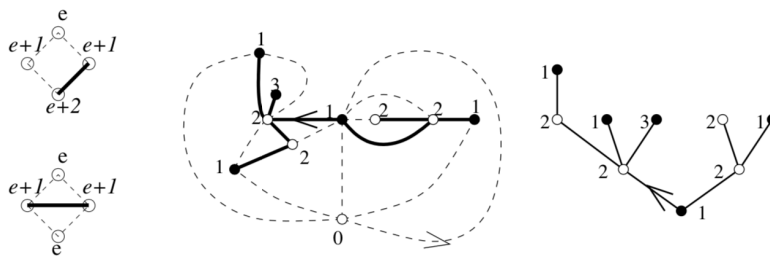
any quadrangulation. However, there still can be multiple edges. See Figure 8.2.

One can check that if a quadrangulation has  $f = n$  faces, then it must have  $e = 2n$  edges. The Euler formula says that there is a relation between the number of faces, edges and vertices in any planar map. Namely,  $f + v = e + 2$ . It follows that the number of vertices in a planar quadrangulation with  $n$  faces is  $v = e + 2 - f = 2n + 2 - n = n + 2$ .

The picture in Figure 8.2 also shows the minimal distances of vertices of quadrangulation from the root vertex.

The distribution of the minimal distance across vertices is called the *profile of the quadrangulation*, that is, the profile is defined as the vector  $(H_k)$ ,  $k = 1, \dots$ , where  $H_k$  is the number of vertices at the distance  $k$  from the root vertex.

Let  $d(v)$  denote the distance from the root. In the following we need some properties of this function.



**Figure 8.3:** The map from quadrangulations to well-labelled trees. On the left: rule for simple (top) and confluent (bottom) faces.

1) **If  $u$  and  $v$  are joined by an edge then  $|d(u) - d(v)| = 1$ .**

This is clear by triangle inequality for the distances. The case  $d(u) = d(v)$  is ruled out because  $u$  and  $v$  has different colors and the function  $d$  has the



different parity for vertices of different color.

The second property is immediate consequence of the first one.

2) **Around a face, four vertices appear: two blacks,  $x_1$  and  $x_2$  and two whites,  $y_1$  and  $y_2$ . These vertices satisfy at least one of the two inequalities  $d(x_1) = d(x_2)$  and  $d(y_1) = d(y_2)$ .**

(Note that it might happen that  $x_1 = x_2$  or  $y_1 = y_2$ .)

In the following, a face will be called *simple* when only one of these equalities is satisfied and *confluent* if both are satisfied.

### 8.1.1 Quadrangulations and well-labelled trees

Here we describe Schaeffer's bijection between rooted planar quadrangulations and well-labelled planar trees. The benefit of this bijection is that we can study random triangulations in the framework of random trees.

A *well-labelled planar tree* is a rooted planar tree, in which every vertex is labelled by a positive integer. The root vertex is labelled by 1 and the labels on adjacent vertices are different by no more than 1.

The label distribution of the well-labelled tree is the vector  $(\lambda_k)$ ,  $k = 1, \dots$ , where  $\lambda_k$  is the number of vertices with label  $k$ .

**Theorem 8.1.1** (Schaeffer). *There is a bijection  $\mathcal{T}$  between planar quadrangulations with  $n$  faces (counting the outer face) and well-labelled trees with  $n$  edges. Under this bijection, the label distribution of the tree  $\mathcal{T}(Q)$  equals the profile of the quadrangulation  $Q$ .*

*Proof.* Here is the map of quadrangulations to well labelled trees.

For every confluent face, take an edge (absent in the original map) that connects two vertices with maximal label  $d(v)$ . For every simple face  $f$  take an edge that have a vertex  $v$  with maximal label as one end-point and leaves  $v$  with face  $f$  on its left. See picture in Figure 8.3

Finally, let the distinguished edge will be the edge that was chosen for outer face of the map.

The first claim is that *the result of this operation is a well-labelled tree with the distinguished edge as its root, and the labels equal to the distances  $d(v)$* . The second claim is that *it is a bijection*, and hence the original planar map can be recovered from the well-labelled tree.

Let us prove the first claim.

**Lemma 8.1.2.** *The mapping  $\mathcal{T}$  sends a quadrangulation  $Q$  with  $n$  faces on a well labeled tree  $\mathcal{T}(Q)$  with  $n$  edges.*

*Proof.* First, why all vertices of  $Q$  belong to  $\mathcal{T}(Q)$ ? Consider a vertex  $x$  which is not the root vertex of  $Q$ , then one of its neighbors, say  $y$ , is located on the path to the root and therefore has a smaller label. For this edge  $(x, y)$  we have the following possibilities:

1. it is incident to a confluent face;

2. it is incident to a simple face in which  $x$  takes the maximal label, and
3. it is incident to two simple faces in which  $x$  takes the intermediate level.

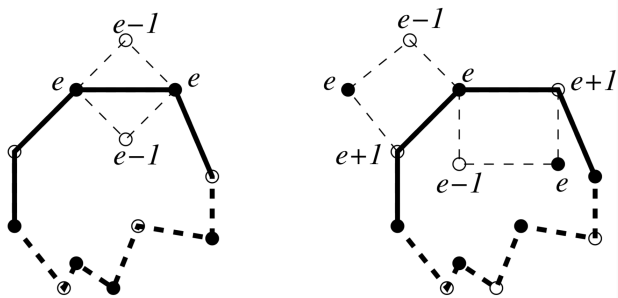
In all of these cases, our rules ensure that  $x$  belongs to at least one edge selected for  $\mathcal{T}(Q)$ . It follows that all vertices except root belong to  $\mathcal{T}(Q)$ . The root is minimal vertex in all its faces and therefore it is not incident to any edge in  $\mathcal{T}(Q)$ .

Since the total number of vertices is  $n + 2$ , it follows that  $n + 1$  vertices are incident to an edge in  $\mathcal{T}(Q)$ . For each face, we have one edge chosen. For two different faces, these edges are different. For confluent faces this is obvious, and for a couple of adjacent simple faces, the edges must be different because of our rule of how to chose an edge from two possible for simple cases. It follows that there are  $n$  edges in  $\mathcal{T}(Q)$ .

It is clear that  $\mathcal{T}(Q)$  is planar by planarity of  $Q$ .

In order to show that  $\mathcal{T}(Q)$  is a tree it remains to prove either that  $\mathcal{T}(Q)$  is connected, or that it does not have cycles.

Here is how cycles can be ruled out.



**Figure 8.4:** Impossibility of cycles

Suppose there exists a cycle in  $\mathcal{T}(Q)$  and let  $e \geq 0$  be the value of the smallest label of a vertex of this cycle. Either all these labels are equal, or there is in the cycle an edge  $(e, e + 1)$  and an edge  $(e + 1, e)$ . In both cases, as we can see from Figure 8.4, the rules of edge selection imply that there is a vertex with label  $e - 1$  in both components of the plane defined by the cycle. Now consider the shortest paths from the root to these vertices. Since the vertices are in different components, one of these paths must intersect the cycle, and the vertices on the path have the labels smaller than  $e - 1$ . This contradicts to the assumption that  $e$  was the smallest label on the path.

It follows that there are no cycles and therefore  $\mathcal{T}(Q)$  is a tree on with  $n$  edges.  $\square$

The second claim is that there is an inverse mapping from well-labeled trees to planar quadrangulations.



**Figure 8.5:** First step of reconstruction algorithm

Here is how the transformation works. The first step is illustrated in Figure 8.5. We view a planar tree as a planar map with a unique face  $F_0$ . We define a *corner* as a sector between two consecutive edges around a vertex. A vertex of degree  $k$  defines  $k$  corners and the total number of corners of a tree with  $n$  edges is  $2n$ . (Adding an edge adds 2 corners to a tree.) The label of a corner is by definition the label of corresponding vertex.

So, first we place a vertex  $v_0$  with label 0 in the face  $F_0$  and add an edge between this vertex and each of the corners with label 1. The new root is the edge from  $v_0$  at the corner before the root of  $T$ . The “before” here is in the sense of counter-clockwise orientation.

If we had  $l$  corners with label 1 then after this procedure we obtain  $l$  faces in the resulting planar map (including the outer face  $F_0$ ).

Then we process every face separately. Let  $k \geq 3$  be the degree of the face  $F$ . Number the corners of  $F$  from 1 to  $k$  starting right after  $v_0$  and going clockwise. Let  $e_i$  be the label of corner  $i$ . (So, in particular  $e_1 = e_{k-1} = 1$  and  $e_k = 0$ .) An example is shown in Figure 8.5 for one of the faces.

Then for each corner  $i \geq 2$  one adds an edge  $(i, s(i))$  inside the face, unless  $s(i) = i + 1$ . Here  $s(i)$  is the successor function:

$$s(i) := \inf\{j \mid j > i, e_j = e_i - 1\}$$

In words, for every corner  $i \geq 2$ , we go along the boundary clock-wise starting with  $i + 1$  and look for a corner that has a label which is smaller by 1 than label of corner  $i$ . If we find such a corner  $j$ , then we connect  $i$  to  $j$ , unless  $j = i + 1$  and it is already connected by an edge.

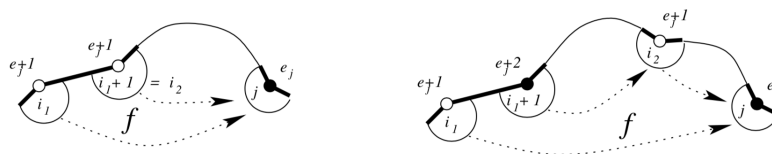
We will prove later that it is possible to add the edges in such a way that they do not intersect.

After these procedure is finished for each face, we obtain a planar map  $\mathcal{Q}'(T)$ . In this planar map we remove the edges with labels of the form  $(e, e)$ . The result is a planar map called  $\mathcal{Q}(T)$ .

The first claim is that  $\mathcal{Q}(T)$  is a quadrangulation with  $n$  faces, and the second claim is that this mapping is inverse for the mapping  $\mathcal{T}$  above.

**Lemma 8.1.3.** *The edges  $(i, s(i))$  do not intersect.*

*Proof.* If we have intersecting edges, then we can choose  $i$  and  $j$  such that  $i < j < s(i) < s(j)$ . Then  $e_s i = e_i - 1$  and the label of  $j$  cannot be  $\leq e_i - 1$ , since otherwise we would encounter the label  $e_i - 1$  earlier before coming to  $s(i)$ , – the labels cannot change by more than 1. Hence  $e_j > e_s(i)$ . However by a similar argument we have  $e_s(i) > e_s(j)$ . It follows that  $e_s(j) \leq e_j - 2$ , which contradicts the definition of  $s(j)$ .  $\square$



**Figure 8.6:** Possible types of faces in  $\mathcal{Q}'(T)$

**Lemma 8.1.4.** *The faces of  $\mathcal{Q}'(T)$  can have only the two types shown in Figure 8.6. They are either triangular with labels  $e, e + 1, e + 1$ , or quadrangular with labels  $e, e + 1, e + 2, e + 1$ . The faces of  $\mathcal{Q}(T)$  are all quadrangular.*

*Proof.* The picture in Figure 8.6 is self-explanatory. However a couple of observations. Let  $f$  be a face in  $\mathcal{Q}'(T)$  and  $j$  be the corner with largest number in the corresponding face  $F$  in  $T_0$ . Let  $i_1 < i_2 < j$  be its two neighbors in  $f$ . (Here we use the numbering inherited from  $F$ ). Then, by the definition for the successor function, we must have  $e_{i_1} = e_{i_2} = e_j + 1$ .

Then two situations are possible. Either  $i_2 = i_1 + 1$ , and the  $f$  is a triangle. Or  $i_2 > i_1$ . In this case  $i_1 + 1$  is not connected with  $j$  and the rules for the successor function imply that  $e_{i_1+1} = e_j + 2$ . and that  $i_1 + 1$  is connected with  $i_1$ . This shows that  $f$  is quadrangular with the properties stated in the lemma.

Finally, when the edges of the type  $(e, e)$  are removed, this joins two triangular faces in a quadrangular face. In particular, it removes all triangular faces and shows that  $\mathcal{Q}(T)$  is a quadrangulation.  $\square$

Now we want to prove our second claim.

**Proposition 8.1.5.** *The mapping  $\mathcal{Q}$  is the inverse of the mapping  $\mathcal{T}$ .*

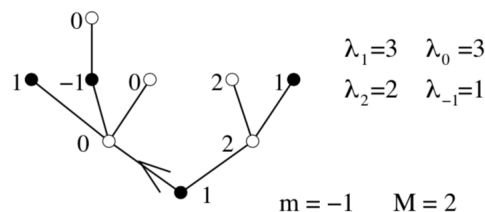
*Proof.* If quadrangulation is an image of the map  $\mathcal{Q}(T)$  then its faces look as quadrangular faces in Figure 8.6 or as pairs of triangular faces in Figure 8.6 joined by the edge  $(e, e)$ . Then we can see that the mapping  $\mathcal{T}$  recovers back the relevant part of the tree by selecting the right edges.

One unclear point is that the set of  $\mathcal{Q}(T)$  gives all possible quadrangulations. This can be resolved either by comparing the cardinalities of the set (which are known from other methods), or directly by showing that  $\mathcal{Q}(\mathcal{T}\mathcal{Q})$  is identity for every quadrangulation  $\mathcal{Q}$ . The last argument is rather complicated.  $\square$

Since  $\mathcal{T}$  is an map of quadrangulations to well-labeled trees, and  $\mathcal{Q}$  is an map of trees to triangulations, and since they are inverses of each other, that means that there is a bijection between quadrangulations and well-labeled trees and this concludes the proof of Theorem 8.1.1

□

### 8.1.2 Embedded trees



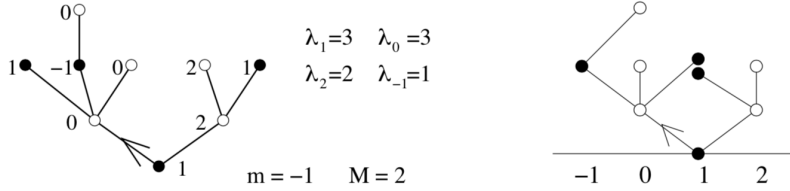
**Figure 8.7:** An example of an embedded tree. The vector  $\lambda$  is the profile of the tree.  $[m, M]$  is the support of the distribution of labels.

How can we generate the well-labelled trees? One of the problems is that well-labelling assumes that the labels is positive and this is a stringent constraint. So we proceed in two steps. The first step is to relax this constraint and consider a more general class of trees, which is called *embedded trees*. The second step that every well-labelled tree belongs to a specific class of embedded trees. Every such class has the same cardinality and the same number of representatives from well-labelled trees. So we will be able to generate a well-labelled tree provided that we are able to generate embedded trees and convert each embedded tree to a corresponding well-labelled tree in its class. The second step will require introduction of yet another class of the trees, the *blossom trees*.

So, for the first step of this program, consider the class of labelled rooted planar trees, where labels on the neighboring vertices can be different by no more than 1 unit but there is no condition of positivity. (The root is labelled 1, as before.) An example is shown in Figure 8.7.

Another way to describe this tree is to imagine that not vertices, but edges are labelled by labels from the set  $\{-1, 0, 1\}$ , which describe how the labels on vertices change when we move away from the root.

We call these trees *embedded trees* for the following reason. These trees are a special type of trees that were considered by David Aldous to describe a random distribution of mass in space. He accomplished this by embedding the trees marked by a random vector  $v_e$  on the edge  $e$  in the space. Every vertex  $v$  in the tree corresponds to the point in space given by the sum  $\sum_{e \in [r, v]} v_e$ , where the summation is over edges in the path from the vertex  $v$  to the root  $r$ .



**Figure 8.8:** An example of an embedded tree and the corresponding mass distribution.

In our special case the “space” is  $\mathbb{Z}$ , the labels on edges are  $\{-1, 0, 1\}$  and the distribution of mass is simply the profile of the tree. See Figure 8.8.

It is easy to generate the embedded tree. Simply generate a planar tree and put the labels  $\{-1, 0, 1\}$  on the edges randomly.

The relation of embedded trees and well-labelled trees is described in the following theorem. Let  $\mathcal{E}_n$  and  $\mathcal{W}_n$  denote the sets of embedded and well-labelled trees with  $n$  edges, respectively. We know that  $\mathcal{W}_n \subset \mathcal{E}_n$  since the well-labelled trees are simply embedded trees for which an additional constraint on labels is satisfied. (The labels must be non-negative.)

Recall also that  $\lambda_k(T)$  denotes the number of vertices with label  $k$ . We also define the cumulated label distribution as

$$\Lambda_k(T) = \sum_{l=1}^k \lambda_{m+l-1}(T),$$

where  $m$  is the minimal label. Clearly  $\Lambda_{M-m+1} = n + 1$ , where  $n$  is the number of edges in  $T$ . For the example in Figure 8.8, we have  $\Lambda_1 = 1$ ,  $\Lambda_2 = 4$ ,  $\Lambda_3 = 7$ ,  $\Lambda_4 = 9$ .

**Theorem 8.1.6** (Chassaing-Schaeffer). *There exists a partition of embedded trees  $\mathcal{E}_n$  into disjoint classes  $C_i$ , such that for every class  $C_i$ ,*

1.  $|C_i| \leq n + 2$ ;
2. *well-labelled trees are fairly represented,*

$$|C_i \cap \mathcal{W}_n| = \frac{2}{n+2} |C_i|,$$

and

3. *there is a relation between label profiles. For every tree  $W \in C_i \cap \mathcal{W}_n$  and every tree  $T \in C_i$ , and all  $k \geq 1$ ,*

$$\Lambda_{k-2}(T) \leq \Lambda_k(W) \leq \Lambda_{k+2}(T).$$

A consequence of this theorem is the enumeration of quadrangulations with  $n$  faces (originally proved by another method in Cori - Vauquelin, 1981).

**Corollary 8.1.7.** *The number of quadrangulations with  $n$  faces is*

$$\frac{2}{n+2} \frac{3^n}{n+1} \binom{2n}{n}.$$

*Proof.* By Schaeffer's theorem (Theorem 8.1.1), we know that the number of quadrangulations with  $n$  faces equal to the number of well-labelled maps with  $n$  edges,  $|\mathcal{W}_n|$ . Theorem 8.1.6 implies that

$$|\mathcal{W}_n| = \frac{2}{n+2} |\mathcal{E}_n|,$$

where  $|\mathcal{E}_n|$  is the number of embedded trees. The total number of rooted planar trees is the Catalan number  $C_n = \frac{1}{n+1} \binom{2n}{n}$ , and the total number of labellings is  $3^n$ . Hence  $|\mathcal{E}_n| = \frac{3^n}{n+1} \binom{2n}{n}$ . Altogether, this implies the statement of the corollary. □

Theorem 8.1.6 can be restated in the following form. Let  $W_n, T_n$  be random trees uniformly distributed on  $\mathcal{W}_n$  and  $\mathcal{E}_n$ , respectively. Also let  $\mu(W_n)$  denote the maximal label of tree  $W_n$ , and  $M(T_n), m(T_n)$  denote the maximum and minimum labels of tree  $T_n$ , respectively.

**Theorem 8.1.8.** *There is a coupling  $(W_n, T_n)$  such that the induced random variables  $(\lambda(W_n), \lambda(T_n))$  for all  $k$  satisfy inequalities*

$$\Lambda_{k-2}(T_n) \leq \Lambda_k(W_n) \leq \Lambda_{k+2}(T_n),$$

and in particular, for this coupling,

$$|\mu(W_n) - (M(T_n) - m(T_n))| \leq 3.$$

*Proof.* The joint distribution of the coupling is given by the following formula

$$\mathbb{P}[(W_n, T_n) = (W, T)] = \begin{cases} \frac{1}{2|\mathcal{E}_n|}, & \text{if } W \text{ and } T \text{ are both in } C \text{ with } |C \cap \mathcal{W}_n| = 2, \\ \frac{1}{|\mathcal{E}_n|}, & \text{if } W \text{ and } T \text{ are both in } C \text{ with } |C \cap \mathcal{W}_n| = 1, \\ 0, & \text{otherwise .} \end{cases}$$

The first inequality follows from the inequality in Theorem 8.1.6. One particular case of this inequality is

$$\Lambda_{\mu-2}(T_n) \leq \Lambda_{\mu}(W_n) = n+1 \leq \Lambda_{\mu+2}(T_n),$$

which implies that

$$\mu - 2 \leq M - m + 1 \leq \mu + 2,$$

and this implies the second inequality of the Theorem. □

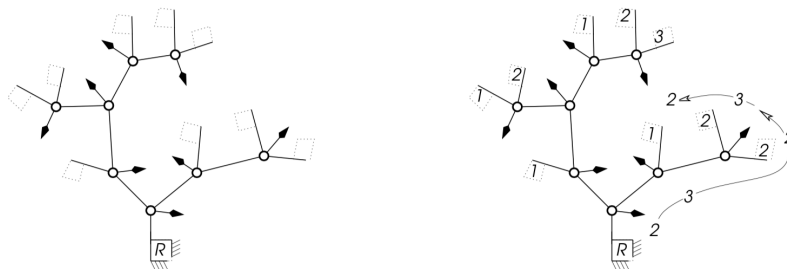


Figure 8.9: Blossom Tree and its labelling

### 8.1.3 Blossom trees

The proof of Theorem 8.1.6 uses *blossom trees*. These are rooted planar trees with the following properties.

1. The leaves of a blossom tree are of two types: *arrows* and *flags*.
2. The root is a *special* leaf, and it is a *special flag*.
3. The inner nodes of the tree (i.e., those which are not leaves) have degree 4, and adjacent to exactly one arrow.

The set of blossom trees with  $n$  inner nodes is denoted  $\mathcal{B}_n$ . Since the degree of inner nodes is 4, it is easy to check that the number of leaves is  $2n + 2$  and, in particular, there are  $n$  arrows and  $n + 2$  leaves (including root).

For the following construction we need to introduce some labels on the flag leaves of the blossom tree. This labelling is completely determined by the tree and given by the following *labelling process*:

- Start with current label 2 just after the root.
- Go around the border of the tree in counter-clockwise direction. If an arrow is reached, increase the *current label* by 1, otherwise, when a non-special flag is reached, decrease the current label by one and write it on the flag.
- Stop when the special (root) flag is reached again.

See an example in Figure 8.9/

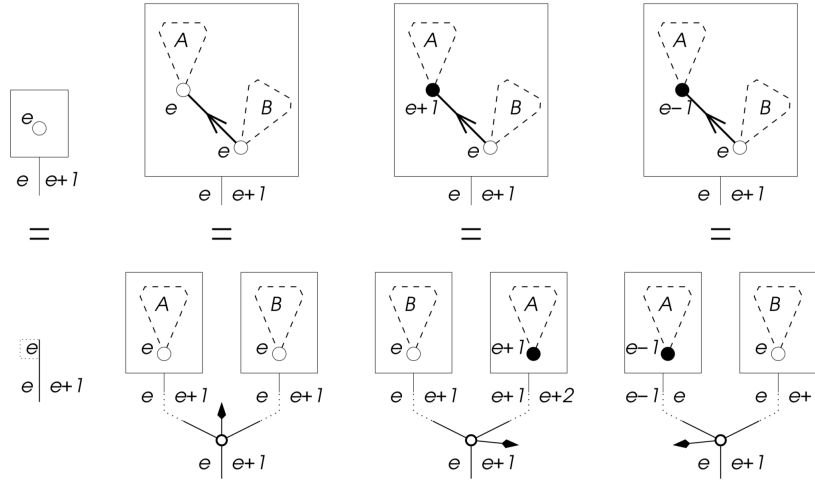
**Theorem 8.1.9.** *There is a bijection between embedded trees with  $n$  edges and blossom trees with  $n$  inner nodes that preserves the label distribution.*

The bijection is build by a recursion and we also need “*decorated*” blossom trees. These are blossom trees in which flags contains not only labels but also an embedded tree, which can be empty. If the embedded tree in the flag is not empty, then *its root has the label equal to the label of the flag*.

In addition, for every embedded tree we define the *root edge* as the left-most edge incident to the root vertex.



*Proof.* The rules for the bijection are illustrated in Figure 8.10.



**Figure 8.10:** Rules for the bijection from embedded to blossom trees

The first step of the encoding of an embedded tree is special and consists in writing it on the normal flag of the unique blossom tree with two flags and no inner node.

The next step is to apply one of the rules in Figure 8.10. The first rule says that if the tree in the flag consist of only one vertex with label  $e$ , then this flag is replaced by the undecorated flag with the label  $e$ .

Three remaining rule pertain to the situation when the embedded tree inside the flag has at least one edge. They prescribe splitting the embedded tree by cutting it over its root edge. The two resulting trees are put then into two flags which connected to a new inner node that replaces the flag that we process. An arrow is also added to the new inner node in a specific direction.

For example in Figure 8.11, we apply first the rule II.

Then the procedure is repeated recursively until no flags with embedded trees remains.

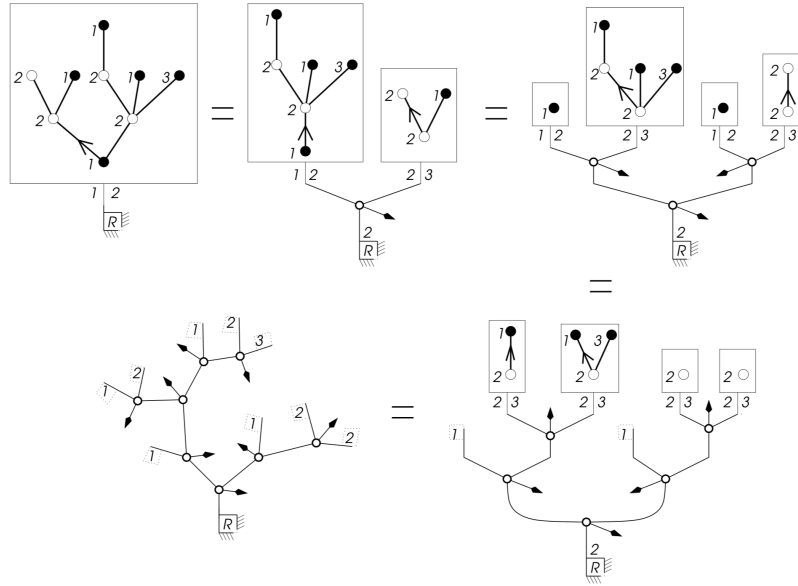
The rules are local so the result do not depend on the order in which the rules are applied.

Note that every time a new inner node is added, an edge is removed from the union of the embedded trees. It follows that if the embedded tree had  $n$  edges, the resulting blossom tree will have  $n$  inner nodes.

It is also obvious that the inner nodes have out-degree 4.

Next, the choice of the position in which the arrow is placed ensures that the labels on the new blossom tree will be consistent with the labels on the embedded trees.

Finally, it is obvious that the procedure is reversible and therefore define a valid bijection.



**Figure 8.11:** An example for the bijection from embedded to blossom trees

□

As a result of this bijection, applied to embedded trees we obtain an *unlabelled* trees, in which, however, the leaves can be of two types, arrows and flags. (The labels on the blossom trees are determined by the graph itself.)

On blossom trees, there is a natural operation of choosing a different root. More precisely a *cyclic shift* of a blossom tree  $A$  is obtained by replacing the special flag by the normal flag and choosing a new special flag.

Given a blossom tree  $B$  with  $n$  arrows and  $n + 2$  flags, the evolution of the current label, in the labelling process is a walk with  $n$  increments  $+1$  and  $n + 2$  increments of  $-1$ , whose last step is a negative increment.

A cyclic shift of a blossom tree corresponds to a cyclic shift of the corresponding walk.

The walks with  $n$  values of  $+1$  and  $n + k$  values of  $-1$  have some interesting properties. They were uncovered in the study of the ballot-problem in the study of election voting and are concerned with “low records” in such walks. We use  $x_i$ ,  $i = 1, \dots, 2n + k$  to denote increments, and  $S_j$ ,  $j = 0, \dots, 2n + k$  to denote the walk values,  $S_0 = 0$ ,  $S_j = \sum_{i=1}^j x_i$ , for  $j > 0$ . A (*low*) *record*  $r$  is a (non-zero) step  $j \geq 1$ , at which a minimum is reached for a first time: for all  $j < r$ ,  $S_j > S_r$ .

Since the walks with  $n$  increments of  $+1$  and  $n + k$  increments of  $-1$  start from 0 and end at  $-k$ , there must be at least  $k$  low records. (Potentially, there can be  $n + k$  low records if the walk goes down for the first  $n + k$  steps.) Let  $r_1 < r_2 < \dots < r_k$  denote the  $k$  *lowest* records. Note that  $r_k$  is the location of

the first occurrence of the absolute minimum of the walk.

An interesting property of the lowest  $k$  records is that under the cyclic shift they are also shifted cyclically.

**Lemma 8.1.10** (Cycle Lemma II). *Suppose that  $x = (x_1, x_2, \dots, x_{2n+k})$ ,  $x_i \in \{-1, +1\}$  is a sequence of increments such that  $\sum_{i=1}^{2n+k} x_i = -k$ , and let  $x' = (x_{1+s}, x_{2+s}, \dots, x_{2n+k+s})$  is its cyclic shift by  $s$ . (Indices are calculated modulo  $2n+k$ .) If  $\{r_1, \dots, r_k\}$  is the set of the low records of  $x$ , then  $\{r_1 - s, \dots, r_k - s\}$  is the set of the low records of  $x'$  (with calculations done modulo  $2n+k$ ).*

Remark: In fact, the proof shows that the order of the low records in  $x'$  is the same as in  $x$  except for the cyclic permutation. So if  $r_i \leq s < r_{i+1}$ , then the new low record sequence is

$$\begin{aligned} r'_1 &= r_{i+1} - s < r'_2 = r_{i+2} - s \leq \dots \leq r'_{k-i} = r_k - s \\ &< r'_{k-i+1} = r_1 - s + 2n + k \leq \dots \leq r'_k = r_i - s + 2n + k. \end{aligned}$$

*Proof.* It is enough to check the validity of the lemma for  $s = 1$ . We consider two cases.

If  $r_1 \geq 2$  then  $S'_{i-1} = S_i - x_1$  for  $i = 2, \dots, 2n+k$  so all the sum were changed by the same amount and therefore the sums  $S_{r_j-1}$  will be the  $k$  smallest low records among these numbers. In addition,  $S'_{2n+k} = -k$  and we know that the minimal of  $S_{r_j}$  was at least  $-k$ . Hence  $S'_{2n+k}$  cannot be a low record. This proves the statement of the lemma in this case.

The second case is when  $r_1 = 1$ . Then we know that  $S_{r_1} = -1$ , which implies that  $S_{r_i} = -i$  for all  $i = 1, \dots, k$ . After the shift we find that  $S'_{i-1} = S_i - x_1 = S_i + 1$  for  $i = 2, \dots, 2n+k$ . This shows that  $S_{r_j-1}$  with  $j = 2, \dots, k$  are still the low records for these numbers. In addition, the new low record will be  $2n+k$  with the sum equal to  $-k$ . This completes the proof of the lemma.  $\square$

Let  $B_{n,k}$  denote the set of walks with  $n$  increments  $+1$  and  $n+k$  increments  $-1$  that ends with a negative increment. The elements of  $B_{n,k}$  are called conjugate if they can be transformed one to another by a cyclic shift.

Note that the number of elements in a conjugacy class cannot exceed  $n+k$ . This is because the walk must end with  $-1$ , so there are only  $n+k$  choices for a possible value of the shift.

Let also  $D_{n,k}$  be a subset of  $B_{n,k}$  such that the value of the walk  $S_i$  is greater than  $-k$  for all but the last step. We call these walks *excursions*. Obviously, a walk from  $B_{n,k}$  belongs to  $D_{n,k}$  if and only if the lowest record  $r_k = 2n+k$  and  $x_{r_i} = -i$ .

**Lemma 8.1.11.** *If  $C$  is a conjugacy class of  $B_{n,k}$ , then the number of excursions (i.e., elements of  $D_{n,k}$  in this class is*

$$|C \cap D_{n,k}| = \frac{k}{n+k} |C|.$$

*Example 8.1.12.* If  $k = 1$  then it means that there is exactly one excursion in each conjugacy class. If  $k = 2$  then it either two or one excursions depending on whether  $C$  has  $n + k$  or  $(n + k)/2$  elements.

*Proof of Lemma.* Note that all walks in  $D_{n,k}$  have a low record at  $2n + k$ . Mark also a downstep in these walks. The total number of such marked walks in a class  $C$  is  $(n + k)|C \cap D_{n,k}|$ . Send this walk by a cyclic shift in such a way that the marked down-step is in the last position. By Lemma the low record at  $2n + k$  will go to a low record. So the result is a walk in  $C$  in which a low record is marked. The number of such objects is  $k|C|$ .

The statement of the lemma follows from the claim that the map that we described is a bijection.

Indeed, the inverse transformation is the cyclic transformation of a walk in  $C$  that sends the marked low record to the position  $2n + k$ .  $\square$

*Proof of Theorem 8.1.6.* We will say that two embedded trees are equivalent if their blossom trees representatives are related by a cyclic shift. This is an equivalence relation and it partitions the set of embedded trees into classes  $C_i$ . Since the number of flags is  $n + 2$ , the number of elements in each class is  $\leq C_i$ . (Can be smaller because a cyclic shift can lead to an isomorphic embedded tree.)

Next, the embedded tree is well-labelled if and only if all its labels are positive. If the corresponding walk started at zero, it should be always non-negative except for two last steps.  $\square$

# Appendix A

## Various Useful Facts

### A.1 Proof of Caratheodory theorem

Since this theorem belongs mainly to the measure theory we will not give the full proof. For the full proof see Durrett's or Shiryaev's books. However we indicate some ideas. In particular we will prove uniqueness.

Before we start proving this result, we are going to develop some useful machinery.

**Definition A.1.1.** A system of sets  $\mathcal{A}$  is called a  $\pi$ -system if for any  $I_1, I_2 \in \mathcal{A}$ , we have  $A_1 \cap A_2 \in \mathcal{A}$ .

**Definition A.1.2.** A collection of subsets  $\mathcal{D}$  of set  $\Omega$  is called a  $\lambda$ -system (or  $d$ -system, or Dynkin system) if

1.  $\Omega \in \mathcal{D}$
2. If  $A \in \mathcal{D}$  and  $B \in \mathcal{D}$ ,  $A \subset B \Rightarrow B - A \in \mathcal{D}$
3. If  $A_n \in \mathcal{D}$  and  $A_n \uparrow A \Rightarrow A \in \mathcal{D}$

Here  $A_n \uparrow A$  means that  $A_1 \subset A_2 \subset \dots$  and  $\bigcup_n A_n = A$ .

Note that every  $\lambda$ -system also satisfy this property:

If  $A_n \in \mathcal{D}$  and  $A_n \downarrow A \Rightarrow A \in \mathcal{D}$ . This is a consequence of the de Morgan law.

Now trivially, any  $\sigma$ -algebra is a  $\lambda$ -system. The converse is not true.

*Ex.* A.1.3. Let  $\Omega = \{1, 2, 3, 4\}$ . Give an example of a  $\lambda$ -system of subsets of  $\Omega$  which is not an algebra.

The importance of  $\lambda$ -systems comes from the following observation.

**Lemma A.1.4.** *If an algebra  $\mathcal{A}$  is a  $\lambda$ -system, then it is a  $\sigma$ -algebra.*

*Proof.* Consider  $A_n \in \mathcal{A}$ ,  $n = 1, 2, \dots$ . Define  $B_n := \bigcup_{i=1}^n A_i \in \mathcal{A}$ . It is clear that  $B_n \subset B_{n+1}$ . Consequently, by the property (3) of a  $\lambda$ -system,  $B_n \uparrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ . Similarly, one can show that  $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$ . Therefore,  $\mathcal{A}$  is a  $\sigma$ -algebra.  $\square$

**Theorem A.1.5** (Dynkin's  $\pi$ - $\lambda$  Theorem). *Suppose  $\mathcal{A}$  is a  $\pi$ -system. If  $\mathcal{D}$  is a  $\lambda$ -system and  $\mathcal{A} \subset \mathcal{D}$ , then  $\sigma(\mathcal{A}) \subset \mathcal{D}$ .*

*Proof.* (of the Dynkin Theorem A.1.5) The key here is to show that  $\mathcal{D}$  is not only a  $\lambda$ -system but also an algebra. Then, the statement of the theorem will hold by Lemma A.1.4.

Without loss of generality we can assume that  $\mathcal{D}$  is the smallest  $\lambda$ -system that contains  $\mathcal{A}$ . Every  $\sigma$ -algebra is a  $\lambda$ -system, and by minimality of  $\mathcal{D}$ , we have that  $\mathcal{D} \subset \sigma(\mathcal{A})$ . We aim to prove that  $\mathcal{D}$  is a  $\sigma$ -algebra and therefore  $\mathcal{D} = \sigma(\mathcal{A})$  by minimality of  $\sigma(\mathcal{A})$ .

First, let us show that  $\mathcal{D}$  is closed under intersections. Here we will use the fact that  $\mathcal{D}$  contains  $\pi$ -system  $\mathcal{A}$ .

Let

$$\mathcal{A}_1 = \{B \in \mathcal{D} : A \cap B \in \mathcal{D} \text{ for all } A \in \mathcal{A}\}.$$

Since  $\mathcal{A}$  is closed under intersections,  $\mathcal{A} \in \mathcal{A}_1$ . We can check that  $\mathcal{A}_1$  is a  $\lambda$ -system. By minimality of  $\mathcal{D}$ ,  $\mathcal{A}_1 = \mathcal{D}$ .

Now let

$$\mathcal{A}_2 = \{B \in \mathcal{D} : A \cap B \in \mathcal{D} \text{ for all } A \in \mathcal{D}\}.$$

Again, one can check that  $\mathcal{A}_2$  is a  $\lambda$ -system. In addition, if  $B \in \mathcal{A}$  then  $B \cap A \in \mathcal{D}$  for all  $A \in \mathcal{A}_1 = \mathcal{D}$ . Hence  $\mathcal{A} \subset \mathcal{A}_2$ . By minimality of  $\mathcal{D}$ ,  $\mathcal{A}_2 = \mathcal{D}$ . This shows that  $\mathcal{D}$  is closed under intersections.

From the definition of  $\lambda$ -system it follows that  $\mathcal{D}$  is also closed under unions, and so it is an algebra. As we have seen in the beginning of the proof, this implies the statement of the Dynkin theorem. □

**Lemma A.1.6** (Identification Lemma for Probabilities). *Let  $P$  and  $Q$  be two probability measures on  $\sigma(\mathcal{A})$  where  $\mathcal{A}$  is a  $\pi$ -system. If  $P(A) = Q(A)$  for  $A \in \mathcal{A}$ , then  $P(A) = Q(A)$  for all  $A \in \sigma(\mathcal{A})$ .*

*Ex.* A.1.7. Give an example of two probability measures  $\mu \neq \nu$  on  $\mathcal{F} =$  all subsets of  $\{1, 2, 3, 4\}$  that agree on a collection of sets  $\mathcal{C}$  with  $\sigma(\mathcal{C}) = \mathcal{F}$ , i.e., the smallest  $\sigma$ -algebra containing  $\mathcal{C}$  is  $\mathcal{F}$ .

*Proof of Lemma A.1.6.* Consider class  $\mathcal{C}$  of sets  $A \in \sigma(\mathcal{A})$ , for which it is true that  $P(A) = Q(A)$ . Note that  $\mathcal{A} \subset \mathcal{C} \subset \sigma(\mathcal{A})$ . We need to prove that  $\mathcal{C} = \sigma(\mathcal{A})$ .

By Theorem A.1.5, in order to prove Lemma A.1.6, it is enough to show that  $\mathcal{C}$  is a  $\lambda$ -system.

Indeed

1.  $\Omega \in \mathcal{C}$  because  $P(\Omega) = Q(\Omega) = 1$ .
2. Let  $A, B \in \mathcal{C}$ . Then  $A \subset B$  implies  $B - A \in \mathcal{C}$ . This is because  $P(B) = Q(B) \Rightarrow P(B - A) + P(A) = Q(B - A) + Q(A) \Rightarrow P(B - A) = Q(B - A)$ .
3.  $A_n \in \mathcal{C}$  and  $A_n \uparrow A$  implies  $A \in \mathcal{C}$ . This is because  $P(A_n) = Q(A_n) \Rightarrow P(A) = Q(A)$  and we can use the result that the countable additivity of a probability measure implies that it is continuous from below.

□

Now we are able to prove the uniqueness part of the Carathéodory theorem.

*Proof of the uniqueness part of the Carathéodory theorem.* Lemma A.1.6 directly implies the uniqueness part of the Carathéodory theorem. Indeed, every algebra is trivially a  $\pi$ -system. Hence if a measure can be extended from an algebra  $\mathcal{F}$  to  $\sigma$ -algebra  $\sigma(\mathcal{F})$ , then this extension is unique. □

*Proof of the existence part of the Carathéodory theorem.* We will give only a sketch of the existence proof:

*Step 1.* Define a set function on all subsets of  $\Omega$ , which is called the *outer measure*:

$$\mu^*(A) = \inf_{A \subset \bigcup_j A_j} \sum_j \mu(A_j), \quad (\text{A.1})$$

where the infimum is taken over all countable collections  $A_j$  of sets from  $\mathcal{A}$  that cover  $A$ . Without loss of generality we can assume that  $A_j$  are disjoint. (Replace  $A_j$  by  $(\bigcup_{i=1}^{j-1} A_i^c) \cap A_j$ .)

*Step 2.* Show that  $\mu^*$  has the following properties:

1. The set function  $\mu^*$  is countably subadditive, that is,

$$\mu^* \left( \bigcup_j A_j \right) \leq \sum_j \mu^*(A_j).$$

2. For  $A \in \mathcal{A}$ ,  $\mu^*(A) \leq \mu(A)$ . (Trivial)
3. For  $A \in \mathcal{A}$ ,  $\mu^*(A) \geq \mu(A)$ . (Here we need to use the countable additivity of  $\mu$  on  $(A)$ .)

*Step 3.* Define a set  $E$  to be measurable if

$$\mu^*(A) = \mu^*(A \cap E) + \mu^*(A \cap E^c)$$

holds for all sets  $A$ , and establish the following properties for the class  $\mathcal{M}$  of measurable sets: The class of measurable sets  $\mathcal{M}$  is a  $\sigma$ -algebra and  $\mu^*$  is countably additive measure on it.

*Step 4.* Finally, show that  $\mathcal{A} \subset \mathcal{M}$ . This implies that  $\sigma(\mathcal{A}) \subset \mathcal{M}$  and  $\mu^*$  is an extension of  $\mu$  from  $\mathcal{A}$  to  $\sigma(\mathcal{A})$ . □

## A.2 Function spaces



Probability theory is closely related to functional analysis. In particular, we

can often use the fact that random variables with a given number of moments can be thought of as belonging to some functional spaces. The most important for us will be Banach and Hilbert spaces.

**Definition A.2.1.** (i) Let  $X$  be a normed linear space with norm  $\|\cdot\|_X$ . If  $X$  is complete with respect to the induced metric  $d(x, y) := \|x - y\|_X$ , it is called a *Banach space*.

(ii) If in addition norm  $\|\cdot\|_X$  arises from an inner product  $(\cdot, \cdot)_X$ , then  $X$  is called a *Hilbert space*.

An example of Banach spaces are  $\mathcal{L}^p(\Omega, \mu)$  spaces. Let  $(\Omega, \mathcal{F}, \mu)$  be a measure space.

Then  $\mathcal{L}^p(\Omega, \mu)$  is a space of measurable functions  $f$ , which have a finite  $p$ -moment,  $\int |f|^p < \infty$ , factored by the following equivalence relation  $f \sim g$  in  $\mathcal{L}^p \iff f = g$   $\mu$ -everywhere.

Let  $\|X\|_p := (\int \|X\|^p)^{1/p}$  be the  $p$  norm of  $X$ . Define convergence in  $\mathcal{L}^p$  as follows:

$$X_n \xrightarrow{\mathcal{L}^p} X \iff \|X_n - X\|_p \rightarrow 0 \quad (\text{A.2})$$

It can be shown that  $\mathcal{L}^p$  is complete for  $p \geq 1$ , i.e. if

$$\lim_{n, m \rightarrow \infty} \|X_n - X_m\|_p = 0 \Rightarrow \exists \text{ a r.v. } X \text{ s.t. } X_n \xrightarrow{\mathcal{L}^p} X. \quad (\text{A.3})$$

Therefore  $\mathcal{L}^p$  spaces are Banach spaces for  $p \geq 1$ . (For  $p < 1$ ,  $\|\cdot\|_p$  is not a norm.)

For  $p = 2$ , the space  $\mathcal{L}^2$  is a Hilbert space with the inner product  $(f, g) = \int fg \, d\mu$ .

For  $p = 1$ , the space  $\mathcal{L}^1$  is the space of all integrable functions.

For  $p = \infty$ ,  $\mathcal{L}^\infty$  is the space of essentially bounded functions with the norm,

$$\|X\|_\infty = \inf\{M : \mathbb{P}(|X| > M) = 0\}.$$

In general, the  $\mathcal{L}^p$  spaces with larger  $p$  are more restrictive and are easier to handle.

### A.3 Convergence of Functions and Integration

**Theorem A.3.1** (Monotone Convergence Theorem).  If  $0 \leq X_n \uparrow X$  then  $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$ .

**Theorem A.3.2** (Fatou's Lemma). If  $X_n \geq 0$  then

$$\liminf_{n \rightarrow \infty} \mathbb{E}X_n \geq \mathbb{E}(\liminf_{n \rightarrow \infty} X_n).$$

*Proof.* Define  $g_k = \inf_{n \geq k} f_n$  and note that  $f_n \geq g_k$  for all  $n \geq k$ . Hence,

$$\mathbb{E}(f_n) \geq \mathbb{E}(g_k) \text{ for all } n \geq k.$$

□



Often we know that the sequence of functions  $X_n$  converges almost surely. In this case the inequality is used in the following form:

$$\liminf_{n \rightarrow \infty} \mathbb{E}X_n \geq \mathbb{E}(\lim_{n \rightarrow \infty} X_n).$$

That is, the expectation of the limit of non-negative functions can only be less than the limit (or limit infimum) of the expectations of these functions.

*Example A.3.3.* Define  $X_n$  on  $[0, 1]$  as  $X_n = n\mathbb{1}_{(0, 1/n)}$ .

$$\lim_{n \rightarrow \infty} \mathbb{E}(X_n) = \lim_{n \rightarrow \infty} 1 = 1 \geq 0 = \mathbb{E}(0) = \mathbb{E}\left(\lim_{n \rightarrow \infty} X_n\right) \quad (\text{A.4})$$

**Remark:** Remember this example to get the right sign in the inequality. Here is a variant of the Fatou Lemma.

**Theorem A.3.4** (Reverse Fatou's Lemma). *If  $X_n \leq X$ , where  $X \geq 0$  and  $\mathbb{E}X < \infty$ , then*

$$\limsup_{n \rightarrow \infty} \mathbb{E}X_n \leq \mathbb{E}(\limsup_{n \rightarrow \infty} X_n).$$

**Theorem A.3.5** (Dominated Convergence Theorem). *If  $X_n \rightarrow X$  a.s.,  $|X_n| \leq Y$  for all  $n$ , and  $\mathbb{E}(Y) < \infty$ , then  $\mathbb{E}(X_n) \rightarrow \mathbb{E}(X)$ .*

The simplest bound  $Y$  in the dominated convergence theorem is a constant. (This works because we are in a finite measure space!)


## A.4 Convergence in $L^1$ and uniform integrability

A class  $\mathcal{C}$  of random variables is called *uniformly integrable* (UI) if given  $\varepsilon > 0$ , there exists  $K \in [0, \infty)$  such that

$$\mathbb{E}(|X|I_{|X| \geq K}) \leq \varepsilon$$

for all  $X \in \mathcal{C}$ .

Uniform integrability is a sufficient condition to ensure that  $X_n \rightarrow X$  implies  $\mathbb{E}X_n \rightarrow \mathbb{E}X$ .

**Theorem A.4.1.**  Let  $\{X_n\}$  be a sequence in  $L^1$ , and let  $X \in L^1$ . Then  $X_n \rightarrow X$  in  $L^1$ , equivalently  $\mathbb{E}(|X_n - X|) \rightarrow 0$ , if and only if the following two conditions are satisfied:

1.  $X_n \rightarrow X$  in probability,
2. the sequence  $\{X_n\}$  is uniformly integrable, (UI).

*Proof of "if" part.* Suppose the conditions (1) and (2) are satisfied. For  $K \geq 0$ , define a function  $\varphi_K$  as follows:

$$\varphi_K(x) = \begin{cases} K, & \text{if } x > K, \\ x, & \text{if } |x| < K, \\ -K, & \text{if } x < -K. \end{cases}$$

Let  $\epsilon > 0$  be given. Then

$$\mathbb{E}|\varphi_K(X_n) - X_n| = \mathbb{E}(|X_n| - K)^+ \leq \mathbb{E}(|X_n|\mathbb{1}_{(X_n > K)}) < \epsilon$$

for sufficiently large  $K$  uniformly in  $n$ .

In addition, uniform integrability implies that  $\sup \mathbb{E}|X_n| < \infty$ , that is, the sequence is  $L^1$ -bounded. By Fatou's lemma we can conclude that  $\mathbb{E}|X| < \infty$  and therefore we can ensure that  $\mathbb{E}|\varphi_K(X) - X| < \epsilon$  for sufficiently large  $K$ . Finally,  $\mathbb{E}|\varphi_K(X_n) - \varphi_K(X)| < \epsilon$  for sufficiently large  $n$  because  $X_n \xrightarrow{\mathbb{P}} X$ , therefore  $\varphi_K(X_n) \xrightarrow{\mathbb{P}} \varphi_K(X)$  (by uniform continuity of  $\varphi_K(x)$ ), and the following fact holds:

*Ex. A.4.2.* If  $Y_n \xrightarrow{\mathbb{P}} Y$  and  $|Y_n| < K < \infty$ , then  $\mathbb{E}Y_n \rightarrow \mathbb{E}Y$ .

By triangle inequality,

$$\mathbb{E}|X_n - X| \leq \mathbb{E}|\varphi_K(X_n) - \varphi_K(X)| + \mathbb{E}|\varphi_K(X_n) - X_n| + \mathbb{E}|\varphi_K(X) - X| < 3\epsilon$$

for all  $n > n_0$ , and the proof is complete.  $\square$

*Ex. A.4.3.* Let  $f \geq 0$  be a Borel function such that  $f(r)/r \rightarrow \infty$  as  $r \rightarrow \infty$ . Suppose  $\mathbb{E}f(|X_\alpha|) \leq C$  for some finite non-random constant  $C$  and all  $\alpha \in \mathcal{I}$ . Show that then  $\{X_\alpha : \alpha \in \mathcal{I}\}$  is a uniformly integrable collection of random variables.

**Solution:** Let  $\epsilon > 0$  and  $a = C/\epsilon$ . Take  $K$  so large that  $f(r)/r \geq a$  for  $r \geq K$ . Then

$$\mathbb{E}[X_n \mathbb{1}_{|X_n| \geq K}] \leq \frac{1}{a} \mathbb{E}[f(|X_n|) \mathbb{1}_{|X_n| \geq K}] \leq \frac{1}{a} \mathbb{E}[f(|X_n|)] \leq \frac{C}{a} = \epsilon.$$

## A.5 Inequalities



Let  $X, Y$  etc. be real r.v.'s defined on  $(\Omega, \mathcal{F}, \mathbb{P})$ .

**Theorem A.5.1** (Jensen's Inequality). *Let  $\varphi$  be convex,  $\mathbb{E}(|X|) < \infty$ ,  $\mathbb{E}(|\varphi(X)|) < \infty$ . Then*

$$\varphi(\mathbb{E}(X)) \leq \mathbb{E}(\varphi(X)) \tag{A.5}$$

*Sketch of proof.* As  $\varphi$  is convex,  $\varphi$  is the supremum of a countable collection of lines.

$$\begin{aligned} \varphi(x) &= \sup_n L_n(x), \quad L_n(x) = a_n x + b_n \\ L_n(\mathbb{E}X) &\stackrel{(1)}{=} \mathbb{E}(L_n(X)) \\ &\stackrel{(2)}{\leq} \mathbb{E}(\varphi(X)) \end{aligned}$$

Take sup on  $n$ .

(1) used linearity, (2) used monotonicity.  $\square$

Keep the following example in mind to remember the direction of the inequality.

*Example A.5.2.*

$$\mathbb{E}X^2 \geq (\mathbb{E}X)^2. \quad (\text{A.6})$$

In other words,  $\text{Var}(X) \geq 0$ .

Other noteworthy facts can be derived as corollaries of Jensen's Inequality.

*Example A.5.3.*

$$\|X\|_p \uparrow \text{ as } p \uparrow \quad (\text{A.7})$$

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|) \quad (\text{A.8})$$

**Theorem A.5.4** (Markov's Inequality). *If  $X \geq 0$ ,  $a > 0$ , then*

$$\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a \quad (\text{A.9})$$

*Proof.* Integrate  $\mathbb{1}_{X \geq a} \leq X/a$ . The stated result follows by monotonicity and linearity.  $\square$

**Theorem A.5.5** (Generalized Chebyshev's Inequality). *Let  $\psi : \mathbb{R} \rightarrow \mathbb{R}^+$  be increasing. Then*

$$\mathbb{P}(Y > b) \leq \mathbb{E}(\psi(Y))/\psi(b) \quad (\text{A.10})$$

*Proof.*

$$\mathbb{P}(Y > b) \stackrel{(1)}{=} \mathbb{P}(\psi(Y) > \psi(b)) \stackrel{(2)}{\leq} \mathbb{E}(\psi(Y))/\psi(b)$$

(1) used that  $\psi$  is increasing, and (2) used Markov's inequality.  $\square$

*Example A.5.6.* Note important examples  $\psi(x) = x^p$ ,  $\exp(x)$ , etc.

$$\begin{aligned} \psi(x) = x^2 &\implies \mathbb{P}(|Y| > b) \leq \mathbb{E}(Y^2)/b^2 \\ X = Y - \mathbb{E}(Y) &\implies \mathbb{P}(|Y - \mathbb{E}(Y)| > b) \leq \mathbb{E}((Y - \mathbb{E}(Y))^2) / b^2 \end{aligned}$$

Here is an application.

**Theorem A.5.7.** *Suppose  $X_i$  are independent random variables with*

$$\mathbb{P}(X_i = 1) = P(X_i = -1) = \frac{1}{2},$$

*and set  $S_n = X_1 + \dots + X_n$ . Then, for each  $a > 0$ ,*

$$\mathbb{P}(S_n > a) < e^{-\frac{a^2}{2n}}.$$

*Proof.* By (A.10),

$$\begin{aligned}\mathbb{P}(S_n > a) &\leq \mathbb{E}e^{\lambda S_n} e^{-\lambda a} \\ &= \prod_{i=1}^n \mathbb{E}e^{\lambda X_i} e^{-\lambda a} \\ &= \cosh(\lambda)^n e^{-\lambda a} \\ &< e^{\lambda^2 n/2 - \lambda a}.\end{aligned}$$

To optimize the inequality, we set  $\lambda = a/n$  and get  $\mathbb{P}(S_n > a) < e^{-a^2/(2n)}$  as claimed.  $\square$

For example, if take  $a = tn$ , then we find that  $\mathbb{P}(|S_n| > tn) < 2e^{-\frac{t^2 n}{2}}$ .

**Theorem A.5.8** (Hölder's Inequality). *If  $p, q \in [1, \infty]$  with  $1/p + 1/q = 1$  then*

$$\mathbb{E}(|XY|) \leq \|X\|_p \|Y\|_q \quad (\text{A.11})$$

Here  $\|X\|_r = (\mathbb{E}(|X|^r))^{1/r}$  for  $x \in [1, \infty)$ ; and  $\|X\|_\infty = \inf\{M : \mathbb{P}(|X| > M) = 0\}$ .

*Proof.* See the proof of (5.2) in the Appendix of Durrett.  $\square$

*Example A.5.9.* If  $|Y| \leq b$  then

$$\mathbb{E}(|XY|) \leq b\mathbb{E}(|X|)$$

**Theorem A.5.10** (Cauchy-Schwarz Inequality). *The special case  $p = q = 2$  is the Cauchy-Schwarz inequality.*


$$\mathbb{E}(|XY|) \leq (\mathbb{E}(X^2)\mathbb{E}(Y^2))^{1/2} \quad (\text{A.12})$$

*Proof.* Apply Hölder's inequality for  $p = q = 2$ .  $\square$

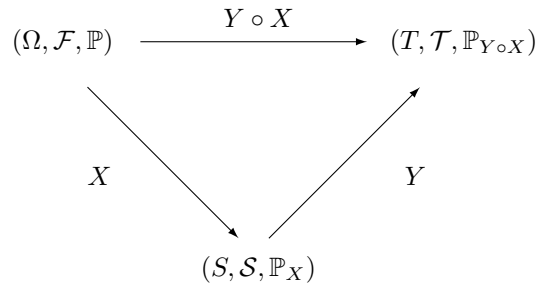
**Theorem A.5.11** (Minkowski's Inequality (Triangle inequality for  $L^p$ )).

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$$

## A.6 Change of Variable

 Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $X : \Omega \rightarrow S$  a  $(\mathcal{F} \setminus \mathcal{S})$ -measurable random variable.  $X$  induces a new probability measure  $\mathbb{P}_X$  on  $(S, \mathcal{S})$ .

**Definition A.6.1.**  $\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$  is called the  $\mathbb{P}$  law of  $X$  or the  $\mathbb{P}$  distribution of  $X$ .



**Figure A.1:** An illustration of the transitivity of the image laws.

Let  $(T, \mathcal{T})$  be another measurable space and  $Y : S \rightarrow T$  a measurable map. Then we have transitivity of the image laws.

**Theorem A.6.2** (Transitivity of the image laws). *The  $\mathbb{P}$  distribution of  $Y \circ X$  is equal to the  $\mathbb{P}_X$  distribution of  $Y$ .*

**Theorem A.6.3** (Change of variable formula). *Let  $Y$  be a real-valued r.v. on  $(S, \mathcal{S})$ .  $Y$  is  $\mathbb{P}_X$ -integrable iff  $Y \circ X$  is  $\mathbb{P}$ -integrable, and then*

$$\int_S Y d\mathbb{P}_X = \int_{\Omega} (Y \circ X) d\mathbb{P}$$

*Proof.* Fix  $X$  and vary  $Y$ . For indicators  $Y$  the identity is the transitivity of image laws, and this passes to simple r.v.'s  $Y$ , then all r.v.'s  $Y$ . See Durrett [1.3, pp. 17]  $\square$

## A.7 Types of Convergence of Random Variables

**W**e learned about the almost sure convergence and convergence in probability. The weak and strong laws of large numbers state that  $\frac{1}{n} \sum_{k=1}^n X_k$  converges in probability (respectively, almost surely) to the expectation of  $X_k$  provided that  $X_k$  are i.i.d and  $\mathbb{E}|X_k| < \infty$ .

Here we discuss some other types of convergence of random variables.

**Convergence in  $L^p$  ( $p \geq 1$ ):** We say  $X_n \xrightarrow{p} X$  if  $\|X_n - X\|_p \rightarrow 0$ , i.e.  $\lim_{n \rightarrow \infty} \mathbb{E}|X_n - X|^p = 0$ .

**Convergence in Distribution:** We say  $X_n \xrightarrow{d} X$  if  $\mathbb{P}(X_n \leq x) \rightarrow \mathbb{P}(X \leq x)$  for all  $x$  at which the RHS is continuous.

The convergence in distribution is also called the weak convergence for the following reason.

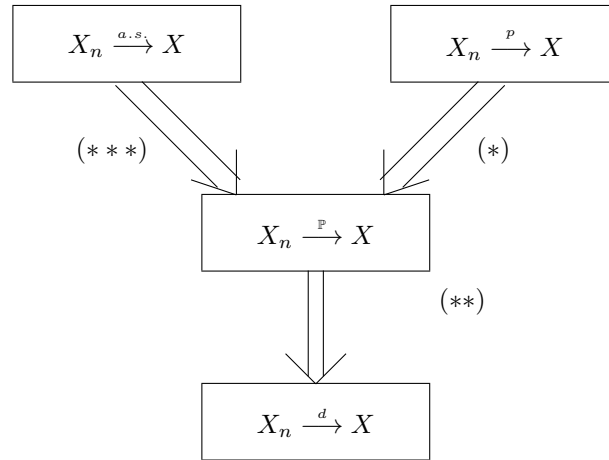
**Theorem A.7.1** (Portmanteau Theorem).  $X_n \xrightarrow{d} X \iff \mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X)$  for all bounded and continuous function  $f$ .

For the proof see Theorem 3.9.1 in Durrett. As for the name, portmanteau is a large leather suitcase that opens up in two hinged compartments. This name for the theorem was apparently introduced by Billingsley.

**Properties in Common for**  $\xrightarrow{\mathbb{P}}, \xrightarrow{a.s.}, \xrightarrow{p}$ :

- a)  $X_n \rightarrow X, Y_n \rightarrow Y \implies X_n + Y_n \rightarrow X + Y, X_n Y_n \rightarrow XY$ .
- b)  $X_n \rightarrow X \iff (X_n - X) \rightarrow 0$  (useful and common reduction).
- c) For all of  $\xrightarrow{\mathbb{P}}, \xrightarrow{a.s.},$  and  $\xrightarrow{p}$  the limit  $X$  is unique up to a.s. equivalence.
- d) Cauchy sequences are convergent (completeness). (Need a metric to metrize  $\xrightarrow{\mathbb{P}}$ , but that is easily provided. See text.)

**Theorem A.7.2.** *The following property holds among the types of convergence.*



*Proof.* (\*) can be proved by Chebyshev's inequality:

$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\epsilon^p}$$

(\*\*): Observe that  $X_n \xrightarrow{\mathbb{P}} X$  implies that  $f(X_n) \xrightarrow{\mathbb{P}} f(X)$  for every bounded and continuous  $f$ . By dominated convergence,  $\mathbb{E}f(X_n) \rightarrow \mathbb{E}f(X)$ , and this implies that  $f(X_n) \xrightarrow{d} f(X)$  by the portmanteau theorem.

(\*\*\*)

**Lemma A.7.3.** *The necessary and sufficient condition for  $X_n \xrightarrow{a.s.} X$  is that*

$$\mathbb{P}\left\{\sup_{k \geq n} |X_k - X| \geq \varepsilon\right\} \rightarrow 0,$$

as  $n \rightarrow \infty$  for every  $\varepsilon > 0$ .

*Proof of Lemma.* Indeed, let  $A_n^\varepsilon = \{\omega : |X_n - X| \geq \varepsilon\}$ , and let

$$A^\varepsilon = \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k^\varepsilon,$$

the set of  $\omega$  where the sequence  $X_n$  deviates from  $X$  by more than  $\varepsilon$  infinitely often. Then,

$$\{\omega : X_n \not\rightarrow X\} = \bigcup_{\varepsilon > 0} A^\varepsilon.$$

In fact in this union it is enough to take a countable union over  $\varepsilon_k = 1/k$ . It follows that

$$\mathbb{P}\{\omega : X_n \not\rightarrow X\} = 0 \Leftrightarrow \mathbb{P}(A^\varepsilon) = 0,$$

for every  $\varepsilon > 0$ . By continuity of probability function this is equivalent to the requirement that

$$\mathbb{P}\left(\bigcup_{k \geq n} A_k^\varepsilon\right) \rightarrow 0,$$

as  $n \rightarrow 0$ , which means that

$$\mathbb{P}\left(\{\omega : \sup_{k \geq n} |X_k(\omega) - X(\omega)| > \varepsilon\}\right) \rightarrow 0,$$

as  $n \rightarrow \infty$ , for every  $\varepsilon > 0$ . □

Since  $\sup_{k \geq n} |X_k - X| > \varepsilon$  implies that  $|X_n - X| > \varepsilon$ , the almost sure convergence implies that

$$\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0,$$

for every  $\varepsilon > 0$ , which is exactly the definition of the convergence in probability. □

Here is an example that shows how the convergence of a sequence in  $L_2$  can be used to prove the convergence of this sequence in probability.

*Example A.7.4.* Suppose that  $X_1, X_2, \dots$  are r.v.'s that have mean 0, have finite variances, and are uncorrelated. Let  $S_n = X_1 + \dots + X_n$ . If  $\sum_{k=1}^{\infty} \mathbb{E}(X_k^2) < \infty$ , then one can show that  $S_n$  converges in  $L^2$  to a limit  $S_\infty$ . This implies that hence  $S_n \xrightarrow{\mathbb{P}} S_\infty$ , i.e.  $\lim_{n \rightarrow \infty} \mathbb{P}(|S_n - S_\infty| > \varepsilon) = 0$  for all  $\varepsilon > 0$ .

*Proof.* Look at the Cauchy criterion. Take  $m > n$ :

$$\mathbb{E}(S_m - S_n)^2 = \mathbb{E} \left( \sum_{k=n+1}^m X_k \right)^2 = \sum_{k=n+1}^m \mathbb{E}(X_k^2) \rightarrow 0$$


as  $m, n \rightarrow \infty$ . Here we used the condition that

$$\sum_{k=1}^{\infty} \mathbb{E}(X_k^2) < \infty.$$

Therefore, the sequence of  $S_m$  converges in  $L^2$ .  $\square$

**Fact:** If the  $X_n$  are independent (or more generally, martingale distributions), then one can prove a stronger statement that  $S_n \xrightarrow{a.s.} S_\infty$ . This is the Kolmogorov-Khinchin theorem about convergence of series, and its extension to the case of martingales. However, there are examples of uncorrelated (but *dependent*) sequences with  $\sum_n X_n^2 < \infty$  where almost sure convergence fails. (See Stout's book "Almost Sure Convergence".)

## A.8 SLLN with with finite 2-nd moment

 It is an insight due to Kolmogorov, that the method of subsequences is still useful for the proof of the almost sure convergence. The idea is to choose a convergent subsequence and to prove that fluctuations of partial sums  $S_n$  between the elements of this subsequence converge to zero almost surely.

**Theorem A.8.1.** *If  $X, X_1, X_2, \dots$  are i.i.d. random variables with  $E(X^2) = \sigma^2 < \infty$ , and  $S_n := X_1 + X_2 + \dots + X_n$ , then*

$$\frac{S_n}{n} \xrightarrow{a.s.} E(X). \tag{A.13}$$

*Proof.* Without loss of generality we can assume that  $E(X) = 0$ . Then, as we have seen,  $S_{k^2}/k^2 \xrightarrow{a.s.} 0$ . Indeed,

$$\mathbb{P} \left( \left| \frac{S_{k^2}}{k^2} \right| > \epsilon \right) < \frac{\sigma^2}{k^2 \epsilon^2},$$

and the convergence holds by the Borel-Cantelli lemma, as in our previous theorem.

Now, let us define

$$M_k = \max_{k^2 \leq n < (k+1)^2} \left| \frac{S_n}{n} - \frac{S_{k^2}}{k^2} \right|.$$



Since for  $k^2 \leq n < (k+1)^2$ ,

$$\left| \frac{S_n}{n} \right| \leq \left| \frac{S_{k^2}}{k^2} \right| + \left| \frac{S_n}{n} - \frac{S_{k^2}}{k^2} \right| \leq \left| \frac{S_{k^2}}{k^2} \right| + |M_k|,$$

and we know that  $\frac{S_{k^2}}{k^2} \xrightarrow{a.s.} 0$ , therefore it is enough to prove that  $M_k \xrightarrow{a.s.} 0$ . (Since if  $X_n \xrightarrow{a.s.} 0$  and  $Y_n \xrightarrow{a.s.} 0$ , then  $X_n + Y_n \xrightarrow{a.s.} 0$ .)

For convenience we define

$$D_k := \max_{k^2 \leq n < (k+1)^2} |S_n - S_{k^2}|.$$

Then, we have

$$\begin{aligned} M_k &= \max_{k^2 \leq n < (k+1)^2} \left| \frac{S_n - S_{k^2}}{n} + \frac{S_{k^2}}{n} - \frac{S_{k^2}}{k^2} \right| \\ &\leq \left| \frac{D_k}{k^2} \right| + 2 \left| \frac{S_{k^2}}{k^2} \right|. \end{aligned}$$

It follows that it is enough to prove that  $D_k/k^2 \xrightarrow{a.s.} 0$ .

We have

$$\begin{aligned} D_k^2 &= \max_{1 \leq m \leq 2n} (S_{k^2+m} - S_{k^2})^2 \\ &\leq \sum_{m=1}^{2n} (S_{k^2+m} - S_{k^2})^2. \end{aligned}$$

Taking expectations on both sides, we get that

$$\begin{aligned} E(D_k^2) &\leq \sum_{m=1}^{2k} m\sigma^2 = k(2k+1)\sigma^2 \\ &\leq 4k^2\sigma^2, \end{aligned}$$

Hence we get that

$$\begin{aligned} \mathbb{P} \left( \left| \frac{D_k}{k^2} \right| > \epsilon \right) &\leq \frac{E \left( \left( \frac{D_k}{k^2} \right)^2 \right)}{\epsilon^2} \\ &\leq \frac{4\sigma^2}{k^2\epsilon^2}. \end{aligned}$$

Hence,

$$\sum_{k=1}^{\infty} \mathbb{P} \left( \left| \frac{D_k}{k^2} \right| > \epsilon \right) < \infty$$

By applying the Borel-Cantelli lemma (BC I), we get that  $D_k/k^2 \xrightarrow{a.s.} 0$ , which completes the proof.  $\square$

## A.9 Devroye's method for generation random variables

In this section we present how to generate random variables from a log-concave distribution.

We assume that we have a random generator that can produce random numbers uniformly distributed on the unit interval  $[0, 1]$ . The basis of Devroye's approach is the rejection method.

Suppose that the target density is  $f$  and we can bound it by a multiple of another function  $g$ , and that we can generate easily random variables from  $g$ .

$$f(x) \leq h(x) = cg(x).$$

For any nonnegative integrable function  $h$  on  $\mathbb{R}^d$ , define the body of  $h$  as  $B_h = \{(x, y) : x \in \mathbb{R}^d, 0 \leq y \leq h(x)\}$ . Note that if  $(X, Y)$  is uniformly distributed on  $B_h$ , then  $X$  has density proportional to  $h$ . Vice versa, if  $X$  has density proportional to  $h$ , then  $(X, Uh(X))$ , where  $U$  is uniform  $[0, 1]$  and independent of  $X$ , is uniformly distributed on  $B_h$ . These facts can be used to show the validity of the rejection method in the following algorithm is as follows.

**Result:**  $X$  is distributed according to density  $f(x)$

**repeat**

    | Generate  $U$  uniformly on  $[0, 1]$  ;  
    | Generate  $X$  with density  $g$ ;

**until**  $U \leq \frac{f(X)}{cg(X)}$ ;

**Algorithm 1:** Rejection algorithm for generation from density  $f(x)$

*Example A.9.1* (Normal random variable). See Devroye's paper.

This can be used for random variables with log-concave density with universal function  $g(x)$ .

In this case, it is useful to define a modified random variable

$$Y = f(m)(X - m),$$

where  $m$  is the mode of the original density  $f(x)$ . Then the new variable has a log-concave density,  $g(x)$ , with a mode at  $x = 0$  and  $g(0) = 1$ .

It is enough to generate  $Y$  since the original random variable can be recovered as  $X = m + Y/f(m)$ .

Devroye showed that the normalized log-concave density  $g(y)$  satisfies the following inequality:

$$g(y) \leq \min\left(1, e^{1-|y|}\right).$$

Hence the function  $h(y) = \min(1, e^{1-|y|})$  can be used in the rejection algorithm above. A random variable with density proportional to  $h(y)$  can be generated by first flipping a perfect coin. If it is "heads", then generate  $(1 + E)S$ , where

$E$  is exponential, and  $S$  is another perfect coin. Otherwise generate  $US$ , where  $U$  is uniform  $[0, 1]$  and independent of  $S$ . Formally, the algorithm is as follows.

**Result:**  $Y$  is distributed according to density  $g(y)$

```

repeat
  Generate  $U$  (a uniform on  $[0, 1]$ ),  $E$  (exponential), and  $S$  (a fair
  random bit);
  Generate a random sign  $S'$ ;
  if  $S == 0$  then
    |  $Y = 1 + E$ ;
  else
    |  $Y = V$ , where  $V$  is uniform on  $[0, 1]$  ;
  end
  Set  $Y = YS'$ ;
until  $U \min(1, e^{1-|y|}) \leq g(Y)$ ;

```

**Algorithm 2:** Algorithm for generation of a random variable with normalized log-concave density  $g(y)$

Devroye generalized this algorithm to cover the case of discrete random variables with log-concave pmf, that is, with such  $p_n = \mathbb{P}\{X = n\}$ , that

$$p_n^2 \geq p_{n-1}p_{n+1}.$$

If the mode of the log-concave pmf is at  $m$ , then one has

$$p_{m+k} \leq p_m \min\left(1, e^{1-p_m|k|}\right),$$

for all  $k$ . Hence, one can develop an appropriate rejection algorithm.

For discrete distributions, the general rejection algorithm is as follows. Suppose that  $p_{m+k} \leq g(x)$  for all  $k - 1/2 \leq x \leq k + 1/2$  and all  $x \in \mathbb{R}$ . Then the a random variable with pmf  $\{p_k\}$  can be generated as follows:

**Result:**  $m + X$  is distributed according to pmf  $p_n$

```

repeat
  Generate  $U$  uniformly on  $[0, 1]$ ;
  Generate  $Y$  with density proportional to  $g$ ;
   $X = \text{round}(Y)$ ;
until  $Ug(Y) \leq p_{m+X}$ ;

```

**Algorithm 3:** Algorithm for generation of a discrete random variable with pmf  $p_n$ .

Devroye applies it with  $g(x) = \min(p_m, p_m e^{1-p_m(|x|-1/2)})$ , and observes that  $g(x)$  is a mixture of a rectangular function on  $[-w/p_m, w/p_m]$  (of integral  $2w$  where  $w = 1 + p_m/2$ ) and two exponential tails outside of this interval (of integral 2).

So the rejection algorithm can be reformulated as follows.

**Result:**  $m + X$  is distributed according to pmf  $p_n$

Compute  $w = 1 + p_m/2$ ;

**repeat**

    Generate  $U, V, W$  uniformly on  $[0, 1]$ ;

    Generate a random sign  $S$ ;

**if**  $U \leq \frac{w}{1+w}$  **then**

        |  $Y = Vw/p_m$ ;

**else**

        |  $Y = (w + E)/p_m$  (where  $E$  is exponential) ;

**end**

    Set  $X = S \times \text{round}(Y)$ ;

**until**  $W \min(1, e^{w-p_m|Y|}) \leq p_{m+X}/p_m$ ;

**Algorithm 4:** Algorithm for generation of a discrete random variable with log-concave pmf  $p_n$  and mode at  $m$

(Note that the exponential  $E$  can be generated simply as  $-\log V$ .)

## A.10 Statistics of Random Structures

In this section, when we call an object random, we mean that it is selected from the uniform distribution on the complete set of these objects.

### A.10.1 Random permutations

The uniform distribution on permutations favors the permutations that have long cycles if we compare it with the random set partitions and its blocks, which we discuss in the next section.

Large cycles are prevalent in a random permutation and the total number of cycles is smaller. In particular, for the number of cycles  $\xi_n$  in random permutation of  $n$ , we have formulas

$$\mathbb{E}\xi_n = \sum_{j=1}^n \frac{1}{j} \sim \log n,$$

$$\mathbb{E}\xi_n = \sum_{j=1}^n \frac{1}{j} - \sum_{j=1}^n \frac{1}{j^2} \sim \log n.$$

Goncharov ... proved that after the appropriate normalization, the random variable  $\xi_n$  becomes asymptotically normal (Theorem 5.1.1 in [?]).

As an intuitive consequence we can expect that the small cycles are relatively rare in a random permutation. This is quantified by another theorem by Goncharov (Theorem 5.1.2 in [?]). Namely, if  $\kappa_n(l)$  denotes the number of cycles of length  $l$  in a random permutation of  $n$ , then for a fixed  $l$  and  $n \rightarrow \infty$ , the distribution of the random variable  $\kappa_n(l)$  converges to the distribution of a Poisson r.v. with the mean  $\lambda_l = 1/l$ .

For a fixed  $s$ -tuple  $l_1, \dots, l_s$ , the random variables  $\kappa_n(l_i)$  become asymptotically independent, as  $n \rightarrow \infty$ .

The length of the largest cycle,  $l_n^{max}$  is asymptotically proportional to  $n$ . The limiting distribution of  $l_n^{max}$  is complicated. Its moments were determined by Shepp and Lloyd in [?] as certain integrals. Lloyd and Shepp have also studied the length of the smallest cycle.

Another interesting statistics on random permutations, the largest increasing sequence, has been a subject of recent research, which led to the discovery of connections with random matrices and Airy distributions.

### A.10.2 Random set partitions

We rely here on the book [?] by V. N. Sachkov.

Let  $\xi_n$  is the number of blocks in a random partition of  $[n]$ . Then, for large  $n$ ,

$$\begin{aligned}\mathbb{E}\xi_n &= \frac{n}{\log n}(1 + o(1)), \\ \text{Var}(\xi_n) &= \frac{n}{(\log n)^2}(1 + o(1)),\end{aligned}$$

and the distribution of the normalized random variable

$$\eta_n = \frac{\xi_n - \mathbb{E}(\xi_n)}{\sqrt{\text{Var}(\xi_n)}}$$

converges to the standard normal distribution as  $n \rightarrow \infty$  (Theorem 4.1.1 in [?]).

Now, let the random variables  $\kappa_n(l)$ ,  $l = 1, \dots, n$ , denote the number of blocks that have size  $l$  in a random partition. Then, the distribution of  $\kappa_n(l)$  has the expectation and the variance both equal to  $\lambda_n = \frac{(r_n)^l}{l!}$ , where  $r_n$  is the solution of the equation  $re^r = n$ .

If  $l$  is fixed and  $n$  is growing then the variances of random variables  $\kappa_n(l)$ ,  $\lambda_n$  are also growing. We can define the normalized random variables

$$\hat{\kappa}_n(l) = \frac{\kappa_n(l) - \mathbb{E}\kappa_n(l)}{\sqrt{\text{Var}(\kappa_n(l))}}.$$

For a fixed  $s$ -tuple  $l_1 < \dots < l_s$ , the joint distribution of normalized random variables  $\hat{\kappa}_n(l_i)$  converges to the standard multivariate normal distribution (Theorem 4.2.1 in [?]).

V. N. Sachkov discusses the distribution of the size of the maximum block, and shows that it is concentrated within a neighborhood of the point

$$er_n - \log \sqrt{2\pi er_n} - \log(e - 1),$$

and that in this domain it is close to the double exponential distribution (without any additional normalization). (For a more precise statement, see Theorem 4.5.2 in [?].) Note that  $r_n$  is asymptotically close to  $\log n$ , hence in the first approximation, the size of the largest block is  $e \log(n)$ .