Exchangeability and a Model of Biological Evolution

Renee Haddad, Undergraduate Thesis

University of Connecticut, 2024

Abstract

A sequence of random variables (RVs) is exchangeable if its distribution is invariant under permutations. For example, every sequence of independent and identically distributed (IID) RVs is exchangeable. The main result on exchangeable sequences of random variables is de Finentti's theorem, which identifies exchangeable sequences as conditionally IID. In this thesis, we explore exchangeability, provide an elementary proof of de Finetti's theorem, and present two applications: the classical Polya's urn model and a toy model for biological evolution.

Exchangeability and a Model of Biological Evolution

Honors Thesis Submitted by

Renee Haddad,

B.S. Mathematics, University of Connecticut, 2024

Thesis & Honors Advisor: Iddo Ben-Ari

University of Connecticut

2024

Table of Contents

Chapter 1

Introduction

1.1 Motivation: Law of Large Numbers (LLN)

The law of large numbers [4, Theorem 2.3.5., p. 69] is one of the foundations of probability theory. It asserts that if $\mathbf{X} = (X_n : n \in \mathbb{N} = \{1, 2, \dots\})$ is a sequence of independent and identically distributed (IID) random variables (RVs) with finite expectation μ and $S_n := X_1 + \cdots + X_n$, then the empirical averages S_n/n converge both in probability, almost surely, and in L^1 to μ. Convergence in probability is the statement that for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0.$$

Convergence almost surely is the statement that there exists an event E, where $P(E) = 1$, such that on E,

$$\lim_{n\to\infty} \frac{S_n}{n} = \mu,$$

and convergence in L^1 is the statement

$$\lim_{n\to\infty} E\left[\left|\frac{S_n}{n} - \mu\right|\right] = 0$$

[4, Theorem 4.6.3., p. 245]. Convergence in L^1 implies convergence in probability through Markov's inequality. Convergence almost surely also implies convergence in probability, and

convergence in probability implies convergence almost surely along some (deterministic) subsequence [4, Section 4.6, p. 244-249].

1.2 LLN for Conditional IID

1.2.1 IID Bernoulli

Consider the simplest non-trivial case corresponding to the RVs being Bernoulli distributed with parameter Θ, a distribution denoted by Bern(Θ). That is, Θ is a constant in $[0, 1]$ and $P(X_n = 1) = \Theta = 1 - P(X_n = 0)$. In this case, the sequence $\mathbf{X}$ can be viewed as representing the results in a sequence of experiments or trials and X_n is the indicator of "success" in the n-th trial, where the probability of success in each trial is equal to the constant Θ, independently of all other trials. This means that the event $X_n = 1$ represents a success in the n-th trial and the event $X_n = 0$ represents a failure in the n-th trial. The RV S_n counts the number of successes in the first n trials and the empirical average S_n/n is the proportion of successes in the first n trials. The law of large numbers states that as $n \to \infty$, these random proportions approach the deterministic constant $\mu = E[X_1] = 1 * P(X_1 = 1) + 0 * P(X_1 = 0) = \Theta$ in probability, almost surely, and in L^1.

1.2.2 Conditional IID Bernoulli

We continue the discussion from the last section, making things a little more complicated by assuming that the probability of success, Θ, is itself a RV and that, conditioned on Θ, $\mathbf{X}$ is an IID sequence of Bern(Θ). As we will now show, this implies that the empirical averages converge to the random variable Θ in probability. More precisely, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \Theta\right| > \epsilon\right) = 0.$$

Indeed, if we fix $\epsilon > 0$ and let $f_n(\theta) = P(|\frac{S_n}{n} - \theta| > \epsilon \mid \Theta = \theta)$, the fact that the sequence $\mathbf{X}$ is IID under the conditional measure $P(\ \cdot\ |\Theta = \theta)$ coupled with the law of large numbers

applied to this probability measure yields

$$\lim_{n\to\infty} f_n(\theta) = 0. \tag{1.1}$$

By definition of conditional probability, $P(|\frac{S_n}{n} - \Theta| > \epsilon) = E[f_n(\Theta)]$. Using (1.1) and the fact that $|f_n(\Theta)| \leq 1$, the Bounded Convergence Theorem [4, Theorem 1.5.3., p. 26] yields that $\lim_{n\to\infty} E[f_n(\Theta)] = E[\lim_{n\to\infty} f_n(\Theta)] = 0$, or:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \Theta\right| > \epsilon\right) = \lim_{n\to\infty} E[f_n(\Theta)] = E\left[\lim_{n\to\infty} f_n(\Theta)\right] = 0.$$

Minor adaptions of the argument allow us to conclude that the empirical averages also converge in L^1. A more involved argument can be used to show that the convergence also holds almost surely.

To make this more concrete, consider the following example.

Example 1.2.1. In our pocket, we have two coins: one fair and one biased, landing Heads with probability 2/3 and Tails with probability 1/3. We randomly pick a coin from our pocket and start tossing it repeatedly. We implicitly assumed the following: each coin is equally likely to be picked, and once chosen, the sequence of tosses is IID. Let Θ be the probability that the coin picked is fair and for $n \in \mathbb{N}$, let X_n be the indicator that the n-th toss lands Heads. Then, in the event that the fair coin was picked, $\Theta = \frac{1}{2}$ and in the event that the biased coin was picked, $\Theta = \frac{2}{3}$. Each of these events has probability $\frac{1}{2}$. Conditioned on Θ, the results of our tosses are IID Bernoulli with parameter Θ. Therefore, as we have just shown above, the empirical averages converge in probability to the random variable Θ.

We continue by examining the distribution of $\mathbf{X}$ through analysis of its covariance structure. Recall that the variance of a square-integrable RV, U, $\mathrm{Var}(U)$ is defined as $\mathrm{Var}(U) = E[(U - E[U])^2] = E[U^2] - E[U]^2$. The variance is nonnegative and is equal to 0 if and only if the RV is constant. The covariance of two square-integrable RVs U and V,

denoted by $\mathrm{Cov}(U, V)$, is defined as

$$\mathrm{Cov}(U, V) = E[(U - E[U])(V - E[V])] = E[UV] - E[U]E[V].$$

Clearly $\mathrm{Cov}(U, U) = \mathrm{Var}(U)$, while if U and V are independent, then $\mathrm{Cov}(U, V) = 0$.

In our case, for every $i \in \mathbb{N}$, we have $E[X_i] = E[P(X_i = 1|\Theta)] = E[\Theta]$. In addition, $E[X_i X_i] = E[X_i^2] = E[X_i] = \Theta$ because $X_i \in \{0, 1\}$. When $j \neq i$, we use the conditional independence to obtain

$$\begin{aligned} E[X_i X_j] &= E\Big[E[X_i X_j|\Theta]\Big] \\ &= E\Big[E[X_i|\Theta]E[X_j|\Theta]\Big] \\ &= E[\Theta^2], \end{aligned}$$

Putting these together we have

$$\mathrm{Cov}(X_i, X_j) = \begin{cases} E[\Theta](1 - E[\Theta]) & i = j \\ \mathrm{Var}(\Theta) = E[\Theta^2] - E[\Theta]^2 & i \neq j. \end{cases}$$

When is the conditionally IID sequence $\mathbf{X}$ actually an IID sequence? As noted above, a necessary condition is $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, which in our setting is equivalent to $\mathrm{Var}(\Theta) = 0$, which is equivalent to the existence of some $\theta \in [0, 1]$ such that $P(\Theta = \theta) = 1$. On the other hand, if the latter condition holds, then the distribution of $\mathbf{X}$ coincides with its distribution conditioned on $\Theta = \theta$, and therefore both are IID $\mathrm{Bern}(\theta)$. In other words, our conditionally IID sequence $\mathbf{X}$ is IID if and only if Θ is a constant RV.

In this section we introduced one way of generating a conditionally IID sequence through some auxiliary RV Θ. In the next sections, we will characterize conditional IID sequences in terms of properties intrinsic to the sequence. The main result, de Finetti's Theorem, will show that conditionally IID sequences are exactly those sequences whose distributions are not affected by relabeling the RVs, also known as exchangeable sequences, and that these sequences can be generated by the very same procedure described in this section, with Θ

defined as the limit of the empirical means, an object intrinsic to the sequence.

Chapter 2

Exchangeability

2.1 Basic Notions

To develop the discussion, we need a few more notions.

Definition 2.1.1. Let $N \in \mathbb{N}$. A permutation on $\{1, \ldots, N\}$ is a bijection on $\{1, \ldots, N\}$, a one-to-one and onto function from $\{1, \ldots, N\}$ to itself.

Any permutation σ has a unique inverse which is also a permutation, denoted by σ^{-1}, satisfying $j = \sigma(i)$ if and only if $i = \sigma^{-1}(j)$ or, equivalently, $(\sigma^{-1} \circ \sigma)(i) = i$, where $\circ$ denotes composition of functions. One trivial but important example of a permutation is the identity mapping $\sigma(i) = i$. Another example is $\sigma(i) = N + 1 - i$. An easy calculation shows that there are exactly $N!$ permutations of $\{1, \ldots, N\}$. It is also easy to show that the set of permutations on $\{1, \ldots, N\}$ is group with respect to composition of functions.

Definition 2.1.2. A sequence of RVs, $\mathbf{X} = (X_n : n \in \mathbb{N})$, is called exchangeable if, for every $N \in \mathbb{N}$ and every permutation σ of $\{1, \ldots, N\}$, the joint distribution of $(X_1, \ldots, X_N)$ is the same as the joint distribution of $(X_{\sigma(1)}, \ldots, X_{\sigma(N)})$.

The simplest example of an exchangeable sequence is an IID sequence. We now present a very important example which is not IID.

Example 2.1.1 (Polya's Urn, [4], Section 4.3.2, p. 226). Consider an urn initially containing b black marbles and w white marbles where $b, w \in \mathbb{N}$. We begin a sequence of trials where, during each, a marble is drawn from the urn at random (meaning uniformly), independently of the past. This marble is then replaced in the urn, along with another marble of the same color (we assume that we have an infinite supply of marbles of each color). For $n \in \mathbb{N}$, let X_n be the indicator that in the n-th trial, a black marble is drawn.

Let's examine the probabilities of drawing one white and two black marbles in a total of three trials. There are three different ways that this sequence can be realized, being $X_1 = 0, X_2 = 1, X_3 = 1$, $X_1 = 1, X_2 = 0, X_3 = 1$, and $X_1 = 1, X_2 = 1, X_3 = 0$. Clearly, the distinction between these ways is the order in which the marbles are drawn. The probability of each distinct sequence is calculated below.

$$\begin{aligned} P(X_1 = 0, X_2 = 1, X_3 = 1) &= P(X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_3 = 1|X_1 = 0, X_2 = 1) \\ &= \frac{w}{b+w} \times \frac{b}{b+w+1} \times \frac{b+1}{b+w+2} \\ &= \frac{w \times b \times (b+1)}{(b+w) \times (b+w+1) \times (b+w+2)}. \end{aligned}$$

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 1) &= P(X_1 = 1)P(X_2 = 0|X_1 = 1)P(X_3 = 1|X_1 = 1, X_2 = 0) \\ &= \frac{b}{b+w} \times \frac{w}{b+w+1} \times \frac{b+1}{b+w+2} \\ &= \frac{w \times b \times (b+1)}{(b+w) \times (b+w+1) \times (b+w+2)}. \end{aligned}$$

$$\begin{aligned} P(X_1 = 1, X_2 = 1, X_3 = 0) &= P(X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_3 = 0|X_1 = 1, X_2 = 1) \\ &= \frac{b}{b+w} \times \frac{b+1}{b+w+1} \times \frac{w}{b+w+2} \\ &= \frac{w \times b \times (b+1)}{(b+w) \times (b+w+1) \times (b+w+2)}. \end{aligned}$$

As we can see, these probabilities are all equal, despite the order in which the two black and one white marbles are drawn differing. This may suggest that the sequence is exchangeable. Now, let's extend this same logic to any finite number of trials.

Let us now prove a general formula for the probability of the event $\cap_{i=1}^{n}\{X_i = x_i\}$ for

any given $n \in \mathbb{N}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$. This is done by induction. To do so, we need some notation. Let $b_n = \sum_{i=1}^{n} x_i$. For simplicity, let $w_n = n - b_n$. Of course, b_n and w_n represent the number of black and white balls sampled respectively.

Our goal is to prove the following:

$$P\left(\bigcap_{i=1}^{n}\{X_i = x_i\}\right) = \frac{(b+w-1)!}{(b-1)!(w-1)!} \times \frac{(b+b_n-1)!(w+w_n-1)!}{(b+w+n-1)!} \tag{2.1}$$

The base case for the proof is $n_0 = 1$. In this case, (2.1) reads

$$\frac{(b+w-1)!}{(b-1)!(w-1)!} \times \frac{(b+b_1-1)!(w+w_n-1)!}{(b+w)!} = \frac{(b+b_1-1)!}{(b-1)!} \times \frac{(w+w_1-1)!}{(w-1)!} \times \frac{1}{b+w}.$$

Since the marble drawn must either be black or white, we now encounter two cases: either $b_1 = 1$ and $w_1 = 0$ or $b_1 = 0$ and $w_1 = 1$. We examine these two cases below:

- If the marble sampled in the first trial is black, meaning $b_1 = 1$ and $w_1 = 1 - 1 = 0$, then the right-hand side becomes

$$\begin{aligned}\frac{(b+1-1)!}{(b-1)!} \times \frac{(w+0-1)!}{(w-1)!} \times \frac{1}{b+w} &= \frac{b!}{(b-1)!} \times \frac{(w-1)!}{(w-1)!} \times \frac{1}{b+w} = b \times 1 \times \frac{1}{b+w} \\ &= \frac{b}{b+w}\end{aligned}$$

  Thus, equation (2.1) gives $\frac{b}{b+w}$.

- If the marble sampled in the first trial is white, meaning $b_1 = 0$ and $w_1 = 1 - 0 = 1$, then the right-hand side becomes

$$\begin{aligned}\frac{(b+0-1)!}{(b-1)!} \times \frac{(w+1-1)!}{(w-1)!} \times \frac{1}{b+w} &= \frac{(b-1)!}{(b-1)!} \times \frac{w!}{(w-1)!} \times \frac{1}{b+w} = 1 \times w \times \frac{1}{b+w} \\ &= \frac{w}{b+w}\end{aligned}$$

  Thus, in this case, equation (2.1) gives $\frac{w}{b+w}$.

We continue to the induction step. Using conditional probability,

$$\begin{aligned}P(X_{n+1} = x_{n+1}, X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) \\ \times P(X_1 = x_1, \ldots, X_n = x_n).\end{aligned} \tag{2.2}$$

By the induction hypothesis, we assume that equation (2.1) is true for all positive integers $n = k$ where $k \geq n_0 = 1$. Namely,

$$P(X_1 = x_1, \ldots, X_n = x_n) = \frac{(b+w-1)!}{(b-1)!(w-1)!} \frac{(b+b_n-1)!(w+w_n-1)!}{(b+w+n-1)!}.$$

At the $n+1$-th trial, we randomly draw a ball from the urn which now contains a total of $b+w+n$ balls, of which exactly $b+b_n$ are black and $w+w_n$ are white. This gives:

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = \begin{cases} (b+b_n) \times \frac{1}{b+w+n} & \text{if } x_n = 1 \\ (w+w_n) \times \frac{1}{b+w+n} & \text{if } x_n = 0 \end{cases}.$$

Let's continue case by case. If the $n+1$-th marble is black, $x_{n+1} = 1$, and thus $b_{n+1} = b_n + 1$ and $w_{n+1} = w_n$. The conditional probability above can be written as $(b+b_{n+1}-1)/(b+w+(n+1)-1)$. Plugging this into (2.2) with the induction hypothesis then gives (2.1). Similarly, if the $n+1$-th marble is white, $x_{n+1} = 0$, then $w_{n+1} = w_n + 1$ and $b_{n+1} = b_n$. Therefore, the conditional probability is $(w + w_{n+1} - 1)/(b + w + (n+1) - 1)$, and (2.1) follows.

We observe that when freezing b, w, and n, the probability is a function of the number of black marbles sampled, a quantity invariant under permutations. Therefore, the sequence $\mathbf{X}$ is exchangeable.

In Section 2.6 we will generalize the model.

An important feature of exchangeable sequences is that the RVs are identically distributed, or, in other words, that being identically distributed is a necessary condition for exchangeability. Suppose that $\mathbf{X}$ is exchangeable. Then, for $x \in \mathbb{R}$ and $n \geq 2$ we have

$$P(X_1 \leq x) = P(X_1 \leq x, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}).$$

Let σ be a permutation on $\{1, \ldots, n\}$ satisfying $\sigma(1) = n, \sigma(n) = 1, \sigma(j) = j$ for $j = 2, \ldots, n-1$. Then,

$$P(X_1 \leq x, X_2 \in \mathbb{R}, \ldots, X_n \in \mathbb{R}) = P(X_{\sigma(1)} \leq x, X_\sigma(2) \in \mathbb{R}, \ldots, X_{\sigma(n)} \in \mathbb{R}) = P(X_n \leq x).$$

Therefore, the distribution of X_1 and X_n is the same for all $n \in \mathbb{N}$.

We now give a very simple example of a sequence $\mathbf{X}$ of identically distributed RVs which is not exchangeable.

Example 2.1.2. Let $X_1 \sim \text{Bern}(\frac{1}{2})$. For $n \in \mathbb{N}$, let $X_{2n} = 1 - X_1$ and $X_{2n+1} = X_1$. As can be easily seen, $X_n \sim \text{Bern}(\frac{1}{2})$ for all $n \in \mathbb{N}$.

Let σ be the permutation on $\{1, 2, 3\}$ given by $\sigma(1) = 1, \sigma(2) = 3$, and $\sigma(3) = 2$. Then

$$P(X_1 = 1, X_2 = 0, X_3 = 1) = \frac{1}{2},$$

while

$$P(X_{\sigma(1)} = 1, X_{\sigma(2)} = 1, X_{\sigma(3)} = 0) = P(X_1 = 1, X_3 = 0, X_2 = 0) = 0.$$

Thus, $\mathbf{X}$ is not exchangeable.

Proposition 2.1.3. *Let $\boldsymbol{X} = (X_n : n \in \mathbb{N})$ be a sequence of random variables. If there exists a random variable Θ such that the distribution of $\boldsymbol{X}$, when conditioned on Θ, is IID, then $\boldsymbol{X}$ is exchangeable. In particular, any IID sequence is exchangeable. In this case, Θ is called the mixing RV.*

This proposition describes how to construct some examples of exchangeable sequences. The sequences discussed in Section 1.2.2 are of this type. To explain the term "mixing," consider the case where Θ is discrete, say $P(\Theta = \theta_i) = p_i$ where $p_i \geq 0$ and $\sum p_i = 1$. Now, $P(\mathbf{X} \in \cdot) = \sum_i p_i P(\mathbf{X} \in \cdot | \Theta_i = \theta_i)$, that is, we are mixing the conditional IID distributions, assigning the one corresponding to $\Theta = \theta_i$ probability (or weight) p_i.

Proof. For simplicity, we will assume that the RVs in the sequence $\mathbf{X}$ are discrete and that Θ is a discrete random variable. We write p_θ for the probability mass function of X_1 conditioned on θ. That is, $p_\theta(x) = P(X_1 = x | \Theta = \theta)$. Since $\mathbf{X}$ conditioned on Θ is IID, it follows that for every θ, $N \in \mathbb{N}$, and $x_1, \ldots, x_N$, we have

$$P(X_1 = x_1, \ldots, X_N = x_N | \Theta = \theta) = \prod_{n=1}^{N} p_\theta(x_n).$$

Now, let σ be a permutation of $\{1, \ldots, N\}$. Then, from the definition of σ^{-1}, we have that

$$\{X_{\sigma(1)} = x_1, \ldots, X_{\sigma(N)} = x_N\} = \{X_1 = x_{\sigma^{-1}(1)}, X_2 = x_{\sigma^{-1}(2)}, \ldots, X_N = x_{\sigma^{-1}(N)}\}.$$

Therefore,

$$\begin{aligned} P(X_{\sigma(1)} = x_1, \ldots, X_{\sigma(N)} = x_N | \Theta = \theta) &= \prod_{n=1}^{N} p_\theta(x_{\sigma^{-1}(x_n)}) = \prod_{n=1}^{N} p_\theta(x_n) \\ &= P(X_1 = x_1, \ldots, X_N = x_N | \Theta = \theta). \end{aligned} \tag{2.3}$$

To complete the proof, we observe that these joint probability functions are equal to each other despite the permutation having reordered the events. Thus, by definition, these IID sequences are exchangeable. □

This leads to the question of whether the converse of Proposition 2.1.3 holds. That is, can *every* exchangeable sequence be expressed as a sequence of conditionally IID RVs? De Finetti's theorem, Theorem 2.2.1, provides an affirmative answer in the case that the sequence takes values in $\{0, 1\}$.

2.2 de Finetti's Theorem

In this section we present de Finetti's Theorem. We will state it here and provide a proof in Section 2.4, after we develop some necessary tools.

Theorem 2.2.1 (de Finetti's Theorem [6], Theorem 2)**.** *Let $\boldsymbol{X} = (X_n : n \in \mathbb{N})$ be a sequence of $\{0, 1\}$-valued RVs which is exchangeable. Then,*

(1) $\Theta := \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k$ *almost surely.*

(2) *Conditioned on Θ, the random variables in $\boldsymbol{X}$ are IID Bern(Θ).*

(3) *For every $m \in \mathbb{N}$, $E[\Theta^m] = E[X_1 \cdots X_m] = P(X_1 = \cdots = X_m = 1)$.*

We make the following comments:

- As Θ is a $[0,1]$-valued RV, its distribution is determined by its moments which are given by the third part of the theorem. This is because the moments determine the expectation of $f(\Theta)$ for all polynomials Θ, and since the polynomials form a dense subset of the continuous functions on $[0,1]$, the result follows. See [2, Example 9.1., p. 146] for details.

- This theorem is a partial converse of Proposition 2.1.3 in the sense that we limit the RVs to be $\{0,1\}$-valued. The mixing RV Θ is derived from the sequence.

- The assumption that $\mathbf{X}$ is an infinite sequence of RVs is necessary for the conclusion. Below is a concrete example of a finite sequence of RVs which is exchangeable, yet the conclusion of the theorem fails.

Example 2.2.1. Suppose that X is Bern($\frac{1}{2}$) distributed. Consider the vector $\mathbf{X} := (X, 1-X)$. That is, we are looking at the first two variables in the sequence from Example 2.1.2. Consider the Kronecker delta defined as follows:

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Then, $\mathbf{X}$ is exchangeable, indeed,

$$\begin{aligned} P(X=i, 1-X=j) = P(X=i, X=1-j) = \frac{1}{2}\delta_{i,1-j} = \frac{1}{2}\delta_{j,1-i} = P(X=j, 1-X=i) \\ = P(1-X=i, X=j). \end{aligned}$$

In particular, take $i=j=1$, $P(X=1, 1-X=1)=0$. Next, we argue by contradiction assuming that $\mathbf{X}$ satisfies the conclusion of de Finetti's theorem, namely, that there exists a random variable Θ such that, conditioned on Θ, X and $1-X$ are IID Bernoulli with parameter Θ. In particular,

$$\begin{cases} \frac{1}{2} = P(X=1) = E[P(X=1|\Theta)] = E[\Theta] \text{ and} \\ 0 = P(X=1, 1-X=1) = E[P(X=1, 1-X=1|\Theta)] = E[\Theta^2] = 0 \end{cases}$$

The first equality implies $P(\Theta > 0) > 0$ and the second implies $P(\Theta > 0) = 0$, a contradiction.

In light of the above, a description of finitely exchangeable sequences is another object of interest. This topic was thoroughly studied in [3].

2.3 Convergence Result

In this section we present a lemma which leads to the proof of de Finetti's theorem. Specifically, we will show that exchangeable sequences - much like IID sequences - satisfy a law of large numbers: the empirical averages converge to some limit. The difference from the IID case is that the limit is in general and not deterministic. Notably and similarly to the IID case, the limit is attained along any increasing subset of the RVs and is the same regardless of the choice of the subset. Before we state our result, we review the notion of convergence in L^2.

Definition 2.3.1. (1) A RV, X, is said to be in L^2, or square integrable, if $E[X^2] < \infty$. The L^2-norm of X, $\|X\|$, is defined as $\sqrt{E[X^2]}$.

(2) A sequence, $(X_n : n \in \mathbb{N})$, of RVs in L^2 converges in L^2 if there exists some RV, X, such that $\lim_{n\to\infty} \|X_n - X\| = 0$. In this case, X is referred to as the limit of the sequence.

(3) A sequence of RVs in L^2, $(X_n : n \in \mathbb{N})$, is a Cauchy sequence in L^2 if $\lim_{n\to\infty} \sup_{m\geq 0} \|X_{n+m} - X_n\| = 0$.

We note that the L^2 norm, $\|\cdot\|$, satisfies the triangle inequality: $\|X + Y\| \leq \|X\| + \|Y\|$. Therefore, any convergent sequence is necessarily Cauchy. Moreover, $\|X\| = 0$ if and only if $X = 0$ almost surely. This and the triangle inequality imply that if X and X' are two limits of the same convergent sequence, then $\|X - X'\| = 0$, equivalently, $X = X'$ almost surely. In other words, limits are unique (up to an additive RV which is equal to 0 with probability 1). Finally, we have the following important completeness result [8, Theorem 25, p. 282]:

Theorem 2.3.2. *Any Cauchy sequence in L^2 is convergent. In this case, the limit is in L^2.*

We comment that convergence in L^2 implies convergence in measure. Indeed, if $(X_n : n \in \mathbb{N})$ converges to X in L^2, then Markov's inequality gives:

$$P(|X_n - X| > \epsilon) \leq E[|X_n - X|^2]/\epsilon^2.$$

Lemma 2.3.3. *Let* $\boldsymbol{X}$ *be an exchangeable sequence of* $\{0,1\}$*-valued RVs. Then*

(1) *The limit* $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n X_k$ *converges in* L^2 *to a limit* Θ*.*

(2) *Let* I *be an infinite subset of* $\mathbb{N}$ *with an infinite complement. Let* $K(n) = \min\{m \in \mathbb{N} : |I \cap \{1,\ldots,m\}| = n\}$*. Let* $A_n^I = \frac{1}{n}\sum_{i\in I, i\leq K(n)} X_i$*. Then,* $A_n^I \to \Theta$ *in* L^2*.*

We can think of I as an infinite sequence of positive integers, $I = \{i_k : k \in \mathbb{N}\}$, where $1 \leq i_1 < i_2 < \ldots$. With this interpretation, $K(n) = i_n$.

Proof. Let $A_n = \frac{1}{n}\sum_{k=1}^n X_k$. To prove the first part, it is enough to show that $(A_n : n \in \mathbb{N})$ is Cauchy in L^2, and therefore converges in L^2.

Fix some $n, l \in \mathbb{N}$. Now,

$$A_{n+l} - A_n = \left(\frac{1}{n+l} - \frac{1}{n}\right)\sum_{k=1}^n X_k + \frac{1}{n+l}\sum_{k=n+1}^{n+l} X_k.$$

Therefore, the square of the left hand side is equal to

$$\frac{l^2}{(n(n+l))^2}\left(\sum_{k=1}^n X_k\right)^2 + \frac{1}{(n+l)^2}\left(\sum_{k=n+1}^{n+l} X_k\right)^2 - 2\frac{l}{n(n+l)(n+l)}\left(\sum_{k=1}^n X_k\right)\left(\sum_{k=n+1}^{n+l} X_k\right).$$

The expectation of this expression reduces to a combinatorial calculation. Write $\mu = E[X_1] = E[X_1^2]$ and $\rho = E[X_1X_2]$. It follows that the expectation is equal to

$$\begin{aligned}
\|A_{n+l} - A_n\|^2 &= E[(A_{n+l} - A_n)^2] \\
&= \frac{l^2}{(n(n+l))^2}(n\mu + n(n-1)\rho) + \frac{1}{(n+l)^2}(l\mu + l(l-1)\rho) - 2\frac{l}{n(n+l)^2}nl\rho \\
&= \left(\frac{l}{n} + \frac{1}{n^2}\right)\frac{l}{(n+l)^2}(\mu - \rho) \\
&\leq \left(\frac{1}{n} + \frac{1}{n^3}\right)(\mu - \rho) \underset{n\to\infty}{\to} 0. \qquad (2.4)
\end{aligned}$$

This proves that $(A_n : n \in \mathbb{N})$ is Cauchy in L^2, and is therefore convergent in L^2, which completes the proof of the first part.

Clearly, the same argument proves that $\lim_{n\to\infty} A_n^I$ exists in L^2. What remains to be shown is that this limit, which we denote by θ^I, is independent of I. Although it may not be immediately clear, the argument above already contains the proof of this statement.

We examine $\|A_n^I - A_{K(n)}\|^2 = E[(A_n^I - A_{K(n)})^2]$. As a consequence of exchangeability, this expectation is equal to $E[(A_n - A_{K(n)})^2]$. Since $K(n) \geq n$, it follows from (2.4) that $\lim_{n\to\infty} E[(A_n^I - A_{K(n)})^2] = 0$. Therefore, as a result of the triangle inequality, it follows that

$$\|\Theta^I - \Theta\| \leq \|\Theta^I - A_n^I\| + \|A_n^I - A_{K(n)}\| + \|A_{K(n)} - \Theta\| \underset{n\to\infty}{\to} 0,$$

completing the proof. □

2.4 Proof of de Finetti's Theorem

In this section, we'll use Lemma 2.3.3 to prove de Finetti's theorem 2.2.1. In order to prove this, we will show that the variable Θ that the distribution of $\mathbf{X}$ is conditioned on coincides with the mixing RV introduced in Proposition 2.1.3, making $\mathbf{X}$ exchangeable. As the distribution of $\mathbf{X}$ is determined by the finite-dimensional distributions, it is enough to show that for every $n \in \mathbb{N}$ and $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$, the probability of $\cap_{i=1}^n \{X_i = x_i\}$, a quantity we denote by $p_{\mathbf{x}}$, is the one obtained by a sequence which, conditioned on Θ, is IID Bern(Θ). Conditioning on Θ, the probability that Bern(Θ) is x for some $x \in \{0,1\}$ is given by $\Theta^x(1-\Theta)^{1-x}$. Therefore, we need to prove that

$$p_{\mathbf{x}} = E\left[\Theta^{\sum_{i=1}^n x_i}(1-\Theta)^{n-\sum_{i=1}^n x_i}\right]. \tag{2.5}$$

Define a finite set of arithmetic progressions $I_1, \ldots, I_n$ as follows:

$$I_j = \{j + nk : k \in \mathbb{N}\} = \{j+n, j+2n, j+3n, \ldots\}.$$

By construction, $I_1, \ldots, I_n$ are all disjoint. Next, using the exchangeability property, we can replace X_1 with X_ℓ for $\ell \in I_1$ as follows:

$$\begin{aligned} p_{\mathbf{x}} &= P(X_{1+n} = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P(X_{1+2n} = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= ... \\ &= P(X_{1+kn} = x_1, X_2 = x_2, \ldots, X_n = x_n). \end{aligned} \tag{2.6}$$

A visual representation of this shuffling can be seen within Figure 2.1. Each of the k rows above can be written as $E[\mathbf{1}_{\{x_1\}}(X_\ell) \prod_{j=2}^n \mathbf{1}_{\{x_j\}}(X_j)]$ by equation (2.5). Because of the linearity of the expectation, we have

$$p_{\mathbf{x}} = E\left[\underbrace{\frac{1}{k} \sum_{i=1}^{k} \mathbf{1}_{\{x_1\}}(X_{1+i\times n})}_{=(*)_k} \prod_{j=2}^{n} \mathbf{1}_{\{x_j\}}(X_j) \right].$$

Note that $(*)_k$ is either $A_k^{I_1}$ if $x_1 = 1$ or $1 - A_k^{I_1}$ if $x_1 = 0$ and that Lemma 2.3.3 guarantees that as $k \to \infty$, it converges in L^2 to Θ or $1 - \Theta$, respectively. Therefore, the limit can be written as $\Theta^{x_1}(1-\Theta)^{x_1}$. As the products of elements in a sequence convergent in L^2 and a bounded RV Z converge in L^2 to the product of the limit and Z, by taking $k \to \infty$ we obtain

$$p_{\mathbf{x}} = E[\Theta^{x_1}(1-\Theta)^{1-x_1} \prod_{j=2}^{n} \mathbf{1}_{\{x_j\}}(X_j)].$$

We then successively repeat the argument, replacing X_2 by the first k elements in I_2, etc., shuffling all individual X_i, $1 \le i \le n$, with values of X_{i+kn}. At the end of this process we obtain (2.5).

2.5 Application: Polya's Urn

In this section, we will describe the distribution of the mixing RV Θ for Polya's urn from Example 2.1.1. First, we must introduce some notation. Let $P_{b,w}$ and $E_{b,w}$ denote the

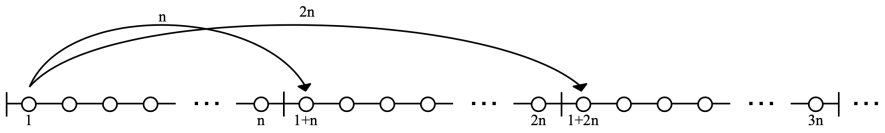

Figure 2.1: This figure demonstrates the shuffling taking place for the X_i values of the $\mathbf{X}$ sequence. Here, we see a demonstration of the first two rows of equation (2.6). Within the arithmetic progression of I_1, we may replace X_1 with X_{1+n}, or further with X_{1+2n}.

distribution of $\mathbf{X}$ when the urn initially contains $b \geq 1$ black marbles and $w \geq 1$ white marbles. The reason we do not suppress the dependence on b and w is because we will vary them.

Before we continue to the details, we introduce two important functions that we will utilize in the sequel. The Gamma function, Γ, defined on $(0, \infty)$, is given by

$$\Gamma(s) = \int_0^\infty e^{-t} t^{s-1} ds.$$

Clearly, $\Gamma(1) = 1$ and integration by parts gives us the equation

$$\Gamma(s+1) = s\Gamma(s), \ s > 0. \tag{2.7}$$

Note then, that as a result, we have that for any $s > 0$ and $k \in \mathbb{N}$,

$$\begin{aligned} s \times (s+1) \times \cdots \times (s+(k-1)) &= \frac{\Gamma(s+1)}{\Gamma(s)} \times \frac{\Gamma(s+2)}{\Gamma(s+1)} \times \cdots \times \frac{\Gamma(s+k)}{\Gamma(s+(k-1))} \\ &= \frac{\Gamma(s+k)}{\Gamma(s)}. \end{aligned} \tag{2.8}$$

Let $n \in \mathbb{N}$. Choosing $k = n+1$ and $s = 1$, (2.8) gives $\Gamma(n+1) = n!$. A closely related function is the Beta function, defined for $\alpha, \beta > 0$ as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{2.9}$$

For $\alpha, \beta > 0$ and $\theta \in [0, 1]$, let

$$f_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}. \tag{2.10}$$

Recall that in Polya's urn model, the RV X_n is the indicator of the n-th draw being a black marble.

Theorem 2.5.1. *Consider Polya's urn from Example 2.1.1 that initially contains b black and w white marbles. Then,*

(1) *The sequence of RVs, $\boldsymbol{X} = (X_n : n \in \mathbb{N})$ is exchangeable.*

(2) *The mixing random variable, Θ, has a density given by $f_{b,w}(\theta)$.*

Two comments are in place:

- As part of the proof, we show that for $b, w \in \mathbb{N}$, $f_{b,w}(\cdot)$ is a density.
- We will extend the result to a slightly more general version of the model in Section 2.6.

Proof. The proof rests on the following simple observation. Fix $b', w' \in \mathbb{N}$. Suppose we run our urn with one marble of each color and sample as follows: the first $b' - 1$ samples are black and the following $w' - 1$ samples are white. Therefore, after $b' + w' - 2 = (b' - 1) + (w' - 1)$ samples, our urn has a total of $b' + w' = 1 + (b' - 1) + 1 + (w' - 1)$ marbles, b' of which are black and w' of which are white. Thus, for all samples from the urn from that point onwards, the sampling will have the same distribution as the sampling under $P_{b',w'}$. In other words, the distribution of $(X_k : k \geq b' + w' - 1)$ under $P_{1,1}$ conditioned on $\{X_1 = \cdots = X_{b'-1} = 1, X_{b'} = \cdots = X_{b'+w'-2} = 0\}$ is the same as the distribution of $\mathbf{X}$ under $P_{b',w'}$. This simple observation allows us to reduce all calculations to the case where $b = w = 1$. In particular,

$$\begin{aligned} E_{b',w'}[\Theta^m] &= E_{1,1}\left[X_{b'+w'-1} = \cdots = X_{b'+w'-m-1} = 1 \;\middle|\; \begin{array}{l} X_1 = \cdots = X_{b'-1} = 1 \\ X_{b'} = \ldots X_{b'+w'-1} = 0 \end{array} \right] \\ &= \frac{E_{1,1}[\Theta^{b'-1+m}(1-\Theta)^{w'-1}]}{P_{1,1}(X_1 = \cdots = X_{b'-1} = 1, X_{b'} = \ldots X_{b'+w'-1} = 0)}. \end{aligned} \tag{2.11}$$

Using (2.1) with $b = w = 1$, the denominator of (2.11) is given by

$$\begin{aligned} P_{1,1}(X_1 = \cdots = X_{b'-1} = 1, X_{b'} = \ldots X_{b'+w'-1} = 0) &= \frac{(b'-1)!(w'-1)!}{(b'+w'-1)!} = \frac{\Gamma(b')\Gamma(w')}{\Gamma(b'+w')} \\ &= B(b', w'). \end{aligned} \tag{2.12}$$

As for the numerator, we first identify the distribution of Θ under $P_{1,1}$. For $m \in \mathbb{Z}_+$, we have

$$E_{1,1}[\Theta^m] = P_{1,1}(X_1 = \cdots = X_m = 1) = \frac{m!}{(m+1)!} = \frac{1}{m+1} = \int_0^1 \theta^m d\theta.$$

This equality clearly holds for $m = 0$. Recall that a random variable is uniformly distributed on $[0, 1]$ if it has a density equal to 1 on $[0, 1]$ and equal to zero elsewhere. As the moments of Θ determine its distribution and the moments of Θ under $P_{1,1}$ coincide with those of a uniformly distributed RV on $[0, 1]$, we conclude that the distribution of Θ under $P_{1,1}$ is uniform on $[0, 1]$. In particular, the numerator of (2.11) is equal to $\int_0^1 \theta^{b'-1+m}(1-\theta)^{w'-1}\theta^m d\theta$. Using this and (2.12), (2.11) can be rewritten as

$$E_{b',w'}[\Theta^m] = \frac{1}{B(b', w')} \int_0^1 \theta^{b'-1}(1-\theta)^{w'-1}\theta^m d\theta. \tag{2.13}$$

In view of (2.10), this can be rewritten as

$$E_{b',w'}[\Theta^m] = \int_0^1 \theta^m f_{b',w'}(\theta) d\theta, \tag{2.14}$$

and the equality also holds for $m = 0$, so $f_{b,w}$ is a density function because it is nonnegative and integrates to 1. As moments determine the distribution of Θ (see comment below Theorem 2.2.1), we conclude that under $P_{b,w}$, Θ has density $f_{b,w}$. □

2.6 Generalization of Polya's Urn

In Example 2.1.1 and in the previous Section 2.5, we looked at the case of Polya's urn where either a black or white marble is drawn from an urn at random, then put back in the urn with an additional marble of the same color. Thus, in that model, we were working with integer-valued b and w as well as $C = 1$ replacement marbles. Here, we expand upon that example and look at a case where the three parameters b, w and C are all strictly positive real numbers. We will construct the model similarly, but instead of thinking of the number of black or white marbles, we will consider the (not necessarily integer-valued) "quantities" of

black and white "material" in the urn, with the probability of drawing each color in a given trial being proportional to the quantity of that material present. After each trial, we add to the urn C amount of the same color material retrieved in that trial, where C again is not necessarily integer-valued. The RV X_n is defined as the indicator of the n-th trial resulting in black material being drawn from the urn. Our goal is to prove a generalization of Theorem 2.5.1. We begin with the following result:

Lemma 2.6.1. *Let $\alpha, \beta > 0$. Then the function $f_{\alpha,\beta}$ defined in* (2.10) *is a probability density on* $[0, 1]$.

Proof. We use a simple change of variables formula for double integrals. Write

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty e^{-t_1} t_1^{\alpha-1} dt_1 \int_0^\infty e^{-t_2} t_2^{\beta-1} dt_2 \\ &= \int_0^\infty \int_0^\infty e^{-(t_1+t_2)} t_1^{\alpha-1} t_2^{\beta-1} dt_1 dt_2.\end{aligned}$$

We change variables from (t_1, t_2) to $(t = t_1 + t_2, \theta = t_1/(t_1 + t_2))$. Clearly, $t \in (0, \infty)$. Since $t_1 = \theta t$ and $t_2 = (1 - \theta)t$, the inequalities $0 < t_1 < \infty$ and $0 < t_2 < \infty$ are equivalent to $0 < \theta t < \infty$ and $0 < (1 - \theta)t < \infty$. As $0 < \theta < 1$, the two inequalities are then equivalent to $0 < t < \infty$ and no additional constraint on θ is given. Finally, the Jacobian matrix is

$$\left|\frac{\partial J(t_1, t_2)}{\partial J(t, \theta)}\right| = \begin{vmatrix} \theta & t \\ 1 - \theta & -t \end{vmatrix} = |-t| = t$$

and so

$$\begin{aligned}\Gamma(\alpha)\Gamma(\beta) &= \int_0^1 \int_0^\infty e^{-t} \theta^{\alpha-1} (1 - \theta)^{\beta-1} t^{\alpha+\beta-2} t dt d\theta \\ &= \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta \int_0^\infty e^{-t} t^{\alpha+\beta-1} dt \\ &= \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} dt \times \Gamma(\alpha + \beta).\end{aligned}$$

Dividing both sides by $\Gamma(\alpha+\beta)$ and using the definition of the Beta function (2.9), we obtain

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta, \tag{2.15}$$

and the result now follows by plugging this into the definition of $f_{\alpha,\beta}$, (2.10). □

Thus, initially we have an amount of b black material and w white material and the probability of drawing black is

$$P(X_1 = 1) = \frac{b}{b+w}.$$

The probability of the first trial being white is

$$P(X_1 = 0) = \frac{w}{b+w}.$$

As before, for $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$, we define $b_n = \sum_{i=1}^n x_i$ and, accordingly, w_n as $w_n = \sum_{i=1}^n (1 - x_i)$. Of course, $n = b_n + w_n$. Using this, we define the probability of drawing black in the $(n+1)$-th trial as

$$P(X_{n+1} = 1 | X_n = x_n, \ldots X_1 = x_1) = \frac{b + C \times b_n}{b + w + C \times n}.$$

Likewise, we define the probability of drawing white in that trial as

$$P(X_{n+1} = 0 | X_n = x_n, \ldots, X_1 = x_1) = \frac{w + C \times w_n}{b + w + C \times n}.$$

A derivation identical to the one leading to (2.1) yields the following

$$P(\bigcap_{i=1}^n \{X_i = x_i\}) = \prod_{k=0}^n \frac{1}{b + w + C \times k} \times \prod_{k=0}^{b_n} (b + C \times k) \times \prod_{k=0}^{w_n} (w + C \times k). \tag{2.16}$$

To simplify this expression, we introduce some notation. For $r, C > 0$ and $j \in \mathbb{N}$, define

$$r^{(C,j)} = r(r + C)(r + 2C) \cdots (r + (j-1)C). \tag{2.17}$$

With this notation, (2.16) can be rewritten as

$$P(\cap_{i=1}^n \{X_i = x_i\}) = \frac{b^{(C,b_n)} w^{(C,w_n)}}{(b+w)^{(C,n)}}, \tag{2.18}$$

We have the following generalization of Theorem 2.5.1:

Theorem 2.6.2. *Consider the generalized Polya's urn as described above with parameters*

$b, w, C \in (0, \infty)$. For each $n \in \mathbb{N}$, let X_n be the indicator that the material drawn from the urn in the n-th trial is black.

(1) *The sequence of RVs, $\boldsymbol{X} = (X_n : n \in \mathbb{N})$, is exchangeable.*

(2) *The mixing random variable, Θ, has a density given by $f_{b/C,w/C}(\theta)$*

Proof. With n, b, w, C all fixed, the expression for the joint distributions (2.18) is a function of b_n, a quantity invariant under permutations. Thus, $\mathbf{X}$ is exchangeable.

Now, consider the following manipulation of the generalized permutation formula presented in (2.17), wherein we reorganize and regroup the terms so that they align with the equation of the gamma function:

$$r^{(C,j)} = C^j \frac{r}{C}\left(\frac{r}{C}+1\right)\cdots\left(\frac{r}{C}+j-1\right) \overset{(2.8)}{=} C^j \frac{\Gamma(\frac{r}{C}+j)}{\Gamma(\frac{r}{C})}.$$

Plugging this expression into (2.18), we obtain

$$\begin{aligned} P(\cap_{i=1}^n \{X_i = x_i\}) &= \frac{\Gamma(\frac{b}{C}+b_n)}{\Gamma(\frac{b}{C})} \times \frac{\Gamma(\frac{w}{C}+w_n)}{\Gamma(\frac{w}{C})} \times \frac{\Gamma(\frac{b+w}{C})}{\Gamma(\frac{b+w}{C}+n)} \\ &= \frac{B(\frac{b}{C}+b_n, \frac{w}{C}+w_n)}{B(\frac{b}{C}, \frac{w}{C})}. \end{aligned}$$

We have now represented the joint distributions for the generalized Polya's urn through the Beta function. It therefore follows that

$$E[\Theta^m] = \frac{B(\frac{b}{C}+m, \frac{w}{C})}{B(\frac{b}{C}, \frac{w}{C})} \overset{(2.15)}{=} \frac{\int_0^1 \theta^{\frac{b}{C}-1}(1-\theta)^{\frac{w}{C}-1}\theta^m d\theta}{B(\frac{b}{C}, \frac{w}{C})} \overset{(2.13)}{=} \int_0^1 \theta^m f_{b/C,w/C}(\theta) d\theta.$$

As moments of a random variable taking values in a bounded interval determine its distribution, we conclude that the density of the mixing random variable Θ is indeed $f_{b/C,w/C}$, completing the proof. □

Chapter 3

Model of Biological Evolution

3.1 The Model

In this section we present an application of de Finetti's theorem in a study of a toy model of biological evolution which results in a self-similar structure. This model is introduced and studied in [1] and [9].

Suppose that a population consists of infinitely many sites labeled by the set of positive integers $\mathbb{N}$. Each site within the population is assigned a "fitness" value, defined here as a number in $[0, 1]$. The system evolves in discrete time as follows.

(1) At each unit of time, we sample one environment value and one newly proposed fitness value for each site within the population.

   (a) The proposed fitness values are IID uniform on $[0, 1]$, independent of the past; and

   (b) the environment value is a Bernoulli random variable independent of the past and is of one of two types: "good" with probability p or "bad" with probability $1 - p$.

(2) The fitness of each member of the population is then updated to be the maximum or the minimum between its current fitness and the newly proposed fitness according to whether the environment is good or bad.

Figure 3.1 provides a visual representation of the evolution dynamics.

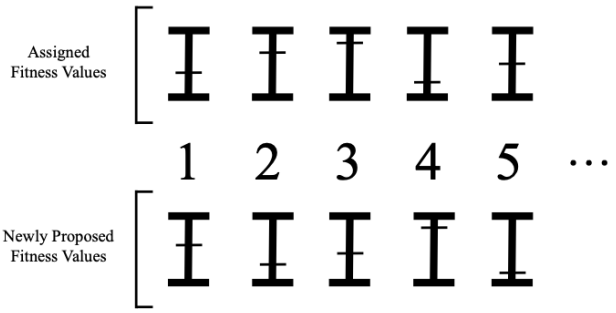

Figure 3.1: Here, the integers represent the first five of the infinitely many sites. The markers in the top row represent the present fitness values and those in the lower row represent the newly proposed fitness values. If the environment is good, then site 1 will assume the newly proposed fitness value in the bottom row, site 2 will retain its assigned fitness value in the top row, and so on, as these values are higher than their alternatives.

Let $\eta = (\eta_t(n) : t \in \mathbb{Z}_+, n \in \mathbb{N})$ denote the process described above, where $\eta_t(n) \in [0,1]$ represents the fitness of the site labeled by $n \in \mathbb{N}$ at time $t \in \mathbb{Z}_+$. Then, $\eta_0(\cdot)$ is the initial assignment of the fitness values and its distribution will therefore be referred to as the initial distribution of the process. To describe the evolution, let $(U_t(n) : t, n \in \mathbb{N})$ be IID uniformly distributed on $[0,1]$ and let $(B_t : t \in \mathbb{N})$ be IID Bernoulli with parameter $p \in (0,1)$. For each $t \in \mathbb{N}$, we view B_t as the indicator of a good environment at time t and $U_t(\cdot)$ as the proposed fitness values. With this notation, we have

$$\eta_{t+1}(n) = \begin{cases} \max(\eta_t(n), U_{t+1}(n)) & \text{if } B_{t+1} = 1 \\ \min(\eta_t(n), U_{t+1}(n)) & \text{if } B_{t+1} = 0 \end{cases}, \tag{3.1}$$

which describes the evolution of the sites during the successive trials according to the sampled random variables. Notably, the evolution at each individual site can be viewed as a Markov chain and, further, this entire evolutionary process is a system of infinitely many Markov chains.

3.2 First Observations

An immediate observation and key to the analysis of the process is the following:

Proposition 3.2.1. *Suppose that the initial distribution of η is exchangeable. Then, for every $t \in \mathbb{N}$, $(\eta_t(n) : n \in \mathbb{N})$ is exchangeable.*

Proof. Let $N \in \mathbb{N}$ and let σ be a permutation on $\{1, \ldots, N\}$. Since both $\eta_t(\cdot)$ and $U_{t+1}(\cdot)$ are exchangeable and independent, a quick calculation shows that the distribution of the vector $(\max(\eta_t(\sigma(n)), U_{t+1}(\sigma(n)) : n = 1, \ldots, N)$ is the same as the distribution of the vector $(\max(\eta_t(n), U_{t+1}(n)) : n \in \mathbb{N})$. The same holds when taking the minima rather than the maxima. The result now follows by conditioning on B_{t+1}. □

We now wish to utilize de Finetti's Theorem on this process, but, to do so, the sequence must be composed of $\{0, 1\}$-valued random variables. To meet this requirement, we impose the following:

(1) Let $u \in [0, 1]$ and let

$$\mathbf{I}_t(n, u) = \mathbf{1}_{\{\eta_t(n) \leq u\}}.$$

We can think of u as a "cut-off" value.

(2) Heuristically, since the labeling of the RVs is arbitrary, the distribution of $(\mathbf{I}_t(n, u) : n \in \mathbb{N})$ is exchangeable and, as such, amenable to de Fintetti's theorem.

Observe that $\eta_t(n)$ can be recovered from $\mathbf{I}_t(n, \cdot)$ through the formula

$$\eta_t(n) = \sup_u \{\mathbf{I}_t(n, u) = 1\}.$$

Corollary 3.2.2. *Suppose that the initial distribution of η is exchangeable. Then, for every $t \in \mathbb{N}$ and $u \in [0, 1]$,*

$$\Theta_t(u) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{I}_k(n, u)$$

exists a.s. and, conditioned on $\Theta_t(u)$*, the random variables* $(\boldsymbol{I}_t(n,u) : n \in \mathbb{N})$ *are IID Bern*$(\Theta_t(u))$.

Proof. For the first statement, use induction. The remaining statements follow from de Finetti's theorem, Theorem 2.2.1. □

Thus, $\Theta_t(u)$ is the proportion of the population of sites that have a fitness value that is less than or equal to u at time t, an object whose existence hinges on de Finetti's theorem. The corollary also allows us to shift our attention from the "complex" process η to the "simpler" $[0,1]$-valued processes $(\Theta_t(u) : t \in \mathbb{N})$. Additionally, we note that for every fixed $t \in \mathbb{Z}_+$, the random function $u \to \Theta_t(u)$ is:

- Non-decreasing;
- Equal to 0 for $u < 0$ and equal to 1 for $u \geq 1$; and
- Right-continuous (this requires a proof, which we omit).

Therefore, it is a (random) cumulative distribution function, or CDF. In other words, the function-valued process $t \to \Theta_t(\cdot)$ takes it values in the set of cumulative distribution functions for probability measures supported on $[0,1]$. Recall that a CDF gives the probability that a RV, X, is less than or equal to some value x. In other words, $P(X \leq x)$.

3.3 Limit Result

For a given site $n \in \mathbb{N}$, the evolution of the $\{0,1\}$-valued process $(\mathbf{I}_t(n,u) : t \in \mathbb{Z}_+)$ is obtained as follows:

(1) If $B_{t+1} = 1$, then

$$\mathbf{I}_{t+1}(n,u) = \mathbf{I}_t(n,u)\mathbf{1}_{\{U_{t+1}(n) \leq u\}}. \tag{3.2}$$

Indeed, the site's fitness is updated to the maximum between its current value and the newly proposed fitness $U_{t+1}(n)$, and this maximum will be less than or equal to u if and only if both $I_t(n,u) = 1$ and $U_{t+1}(n) \leq u$.

(2) If $B_{t+1} = 0$, then

$$\mathbf{I}_{t+1}(n,u) = \mathbf{I}_t(n,u) + (1 - \mathbf{I}_t(n,u))\mathbf{1}_{\{U_{t+1}(n) \leq u\}}. \tag{3.3}$$

Indeed, the site's fitness is updated to the minimum between its current fitness value and the newly proposed fitness $U_{t+1}(n)$. This minimum will be less than or equal to u if and only if $I_t(n,u) = 1$ or $I_t(n,u) = 0$ but $U_{t+1}(n) \leq u$.

We wish to pass from the individual site level dynamics to the bulk. To do that, condition first on $\eta_t(\cdot)$ and on $B_{t+1} = 1$, and consider the proportion among those sites n satisfying $I_t(n,u) = 1$ which also satisfy $U_{t+1}(n) \leq u$. As the RVs $U_{t+1}(\cdot)$ are IID and independent of both $\eta_t(\cdot)$ and B_{t+1}, the law of large numbers tells us that this proportion is u. As the proportion of sites n satisfying $I_t(n,u) = 1$ is by definition $\Theta_t(u)$, it follows from this argument that in the event that $\{B_{t+1} = 1\}$, $\Theta_{t+1}(u) = \Theta_t(u)u$. This and a similar argument for the case that $B_{t+1} = 0$ yield

$$\Theta_{t+1}(u) = \begin{cases} \Theta_t(u)u & B_{t+1} = 1 \\ \Theta_t(u) + (1 - \Theta_t(u))u & B_{t+1} = 0 \end{cases} \tag{3.4}$$

For fixed $u \in [0,1]$, let $S_0, S_1 : [0,1] \to [0,1]$ be the affine mappings $S_1(x) = ux$ and $S_0(x) = u + (1-u)x$. In words, $\Theta_{t+1}(u)$ is either $S_1(\Theta_(u))$ or $S_0(\Theta_t(u))$ chosen according to whether $B_{t+1} = 1$ or $B_{t+1} = 0$. Let's write this as an equation:

$$\Theta_{t+1}(u) = S_{B_{t+1}}(\Theta_t(u)) = \begin{cases} S_1(\Theta_t(u)) & B_{t+1} = 1 \\ S_0(\Theta_t(u)) & B_t = 0 \end{cases} \tag{3.5}$$

In other words, the process $t \to \Theta_t(u)$ is obtained as a successive composition of the affine mappings S_0 and S_1.

Recall that a geometric distribution with parameter q counts the number of independent trials, each with probability q of success, until the first success. We denote this distribution by Geom(q). Equivalently, G is a Geom(q)-distributed RV if and only if

$$P(G = k) = (1 - q)^{k-1}q, \ k \in \mathbb{N}.$$

Theorem 3.3.1. *[1, Theorem 1, p. 3] Let $G_0, G_1, \ldots, G_k$ be a sequence of IID Geom$(1-p)$-distributed RVs. For $k \in \mathbb{Z}_+$, let $T_k = G_0 + G_1 + \cdots + G_k$. For every $u \in (0,1)$, the distribution of the random CDF $\Theta_t(u)$ converges to the distribution of the random CDF $\Theta_\infty(u)$ as $t \to \infty$. This distribution is given below:*

$$\Theta_\infty(u) = \sum_{k=0}^{\infty} u^{T_k} \left(\frac{1-u}{u}\right)^k. \tag{3.6}$$

The process $t \to \Theta_t(u)$ is a Markov process. We do not provide a definition of Markov processes, but heuristically, Markov processes are processes which have no memory: the distribution of the process in the future conditioned on the past is only a function of the present state.

Note that the distribution of the limit $\Theta_\infty(u)$, a probability distribution we denote by μ_u, is independent of the distribution of $\Theta_0(u)$, the initial distribution. A stationary distribution for a Markov process is a probability distribution that is invariant under the dynamics of the process: when taken as the initial distribution, the distribution at all times remains the same. As can be easily seen, the convergence result in the theorem implies that μ_u is in fact the unique stationary distribution of the process. Indeed, if μ is any stationary distribution and $\Theta_0(u)$ is μ-distributed, then for all $t \in \mathbb{Z}_+$, $\Theta_t(u)$ is μ distributed, but then the convergence result implies $\mu = \mu_u$.

The fact that our process is obtained through successive compositions of the affine functions S_0 and S_1 leads to a self-similar structure of μ_u which we now explain. Suppose that $\Theta_0(u)$ is μ_u-distributed. What would be the distribution at time 1? The answer is, of course, μ_u, but when conditioning on B_1 we obtain the following identity. Let $I \subseteq [0,1]$ be

an interval. Then,

$$\begin{aligned}\mu_u(I) = P(\Theta_1(u) \in I) &= P(\Theta_1(u) \in I|B_1 = 0)P(B_1 = 0) + P(\Theta_1(u) \in I|B_1 = 1)P(B_1 = 1) \\ &\overset{(3.5)}{=} pP(S_1(\Theta_0(u)) \in I) + (1-p)P(S_0(\Theta_0(u)) \in I) \\ &= pP(\Theta_0(u) \in S_1^{-1}(I)) + (1-p)P(\Theta_0(u) \in S_0^{-1}(I)) \\ &= p(\mu_u \circ S_0^{-1})(I) + (1-p)(\mu_u \circ S_1^{-1})(I).\end{aligned}$$

To understand what this equation means, recall that the image of S_1 is $[0, u]$ and that the image of S_0 is $[u, 1]$. The intersection of these two intervals is a single point, u, and μ_u is a continuous distribution $\mu_u(\{u\}) = 0$ ([1, Proposition 5]). We will use this below. Fix an interval $B \subseteq [0, 1]$ and let $A_0 = S_0(B) \subseteq [u, 1]$ and $A_1 = S_1(B) \subseteq [0, u]$. We have

$$\begin{aligned}\mu_u(A_0) &= p\mu_u(S_1^{-1}(A_0)) + (1-p)\mu_u(S_0^{-1}(A_0)) \\ &= 0 + (1-p)\mu_u(B),\end{aligned}$$

obtained because $A_0 \subseteq [u, 1]$ and $S_1^{-1}(A_0) \subseteq \{u\}$. After rearrangement, the above equation reads

$$\mu_u(B) = \frac{1}{1-p}\mu_u(S_0(B)). \tag{3.7}$$

The probability assigned to B is $\frac{1}{1-p}$ times the probability assigned to $S_0(B) = u + (1-u)B = \{u + (1-u)b : b \in B\}$, a translation of a dilation of B contained in $[u, 1]$. Or, for any interval B, dilating B by $(1-u)$ then shifting this by u yields an interval whose measure under μ_u is $1-p$ times the measure of B. This can be repeated ad-infinitum. An identical argument shows that

$$\mu_u(B) = \frac{1}{p}\mu_u(S_1(B)) \tag{3.8}$$

or: the measure of B is $\frac{1}{p}$ times the measure of its dilation $S_1(B) = \{ub : b \in B\}$. Again, this can be repeated ad-infinitum. The bottom line is that (3.7) and (3.8) tell us that μ_u is "mirrored," up to multiplicative constants in a sequence of arbitrarily small intervals obtained by successive applications of each of the affine functions S_0 and S_1. This is what self-similarity

is all about. A special case corresponds to $u = p$. In this case, (3.7) and (3.8) imply that the measure of any interval under μ_u is exactly its length, resulting in μ_u being the uniform distribution on $[0, 1]$.

Figure 3.2 provides a visual illustration of self-similarity. Each curve was obtained by looking at $N = 1000$ copies of the the process for a fixed value of u at time $t = 10^5$, starting from $\Theta_0(u) = 0$ and recording the proportion of those N copies, which at time t did not exceed the value x for each $x \in [0, 1]$. In other words, each curve is the empirical distribution of the N copies of $\Theta_t(u)$ we ran, and should be considered as an approximation of the CDF of μ_u. This is clearly visible:

- The bottom curve corresponds to $p = \frac{1}{2}$ and $u = \frac{5}{6}$ and, therefore, (3.8) suggests its restriction to $[0, u] = [0, \frac{5}{6}]$ is a rescaling of the entire curve with its width multiplied by $\frac{5}{6}$ and its height multiplied by $1 - p = \frac{1}{2}$, while (3.7) suggests that the restriction of the same curve to $[\frac{5}{6}, 1]$ is a shifted and rescaled copy of the entire curve.

- The middle curve corresponds to $p = u = \frac{1}{2}$ and, indeed, appears as an approximation of the CDF of the uniform distribution on $[0, 1]$.

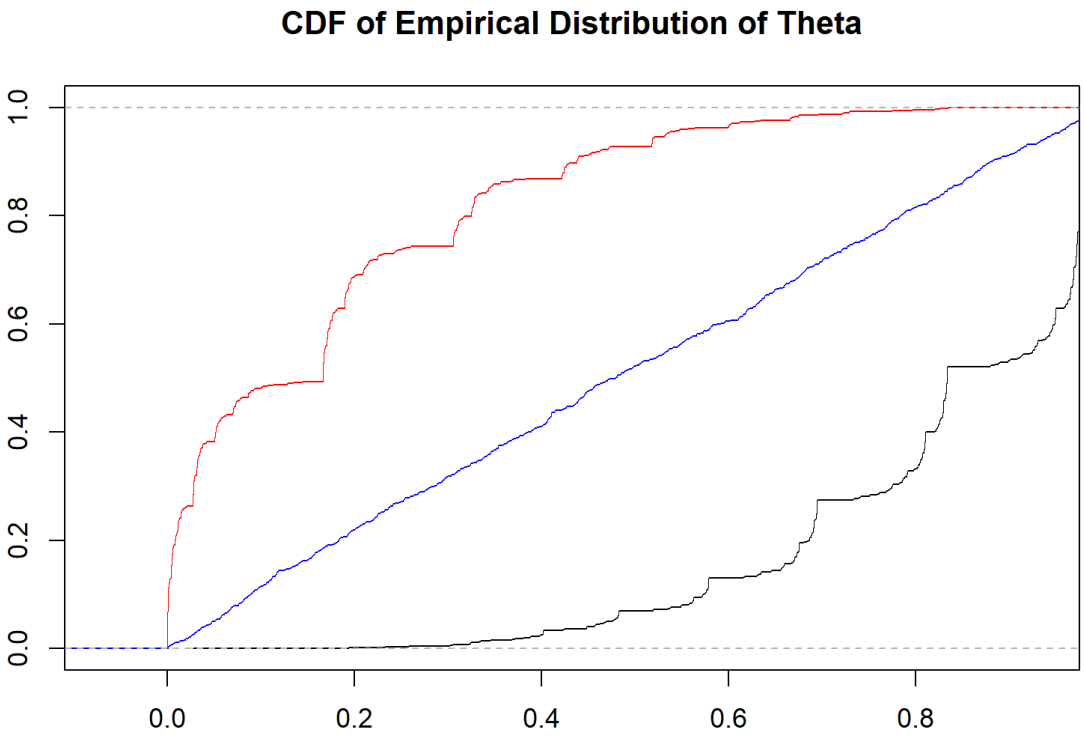

Figure 3.2: The empirical distributions of $\Theta_t(u)$, for $p = \frac{1}{2}$ and $t = 10^5$. The upper curve corresponds to $u = 1/6$, the middle curve to $u = \frac{1}{2}$ and the lower curve to $u = 5/6$. Each curve was obtained from 1000 copies of the process with $\Theta_0(u) = 0$.

Bibliography

[1] I. Ben-Ari and R. B. Schinazi. "Self-similarity in an exchangeable site-dynamics model". *J. Stat. Phys.* 188 (2022), Paper No. 17.

[2] P. Billingsley. *Probability and measure*. Third. Wiley Series in Probability and Mathematical Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1995, xiv+593.

[3] P. G. Bissiri. "Characterization of the law of a finite exchangeable sequence through the finite-dimensional distributions of the empirical measure". *Statist. Probab. Lett.* 80 (2010), 1306–1312.

[4] R. Durrett. *Probability—theory and examples*. Fifth. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019, xii+419.

[5] J. F. C. Kingman. "Uses of exchangeability". *Ann. Probability* 6 (1978), 183–197.

[6] W. Kirsch. "An elementary proof of de Finetti's theorem". *Statist. Probab. Lett.* 151 (2019), 84–88.

[7] J. W. Lewin. "The Teaching of Mathematics: A Truly Elementary Approach to the Bounded Convergence Theorem". *Amer. Math. Monthly* 93 (1986), 395–397.

[8] H. L. Royden. *Real analysis*. Third. Macmillan Publishing Company, New York, 1988, xx+444.

[9] R. B. Schinazi. "Collective evolution under catastrophes". *Amer. Math. Monthly* 131 (2024), 48–59.