An Introduction to Markov Chains in Music Composition and Analysis

Hugo Tremonte de Carvalho
Federal University of Rio de Janeiro (UFRJ)
hugo@dme.ufrj.br
Orcid: 0000-0003-0776-0400

Abstract: *Since the second half of the 20th. century, the use of probabilistic structures to compose and analyze music became even more popular, and nowadays, with the widespread and easiness of use of high-level structures such as neural networks due to very efficient and intuitive computational packages, the area of algorithmic composition became even more popular. This text is an expansion of a lecture named "A brief overview of algorithmic composition", presented at the Music School of the Federal University of Rio de Janeiro in October 2019, and its main goal is to present a gentle and music-oriented introduction to Markov chains, a very intuitive and interpretable tool widely used in algorithmic composition.*

Keywords*: Markov chain, algorithmic composition, probability, statistics*

I. Introduction

Our lives are surrounded by randomness, and at each moment we face decisions that must be made without knowing the exact outcome. In a macroscopic scale, we usually resort to randomness to model extremely complex phenomena where it is infeasible to have all the necessary information required to predict the observed result; for example, the number observed when rolling a die is anything but random, since it is subjected to Newton's laws of motion, a set of deterministic equations. However, since this system is very sensitive to the initial conditions (i.e., initial position and velocity of the die) as well as a very accurate modeling of the die's geometry, the corresponding table where it will fall, atmospheric conditions, among others, it is much easier to simply model the observed value as purely random. On the other hand, the microscopical quantum world provides us real randomness, as it can be seen, for example, in Heisenberg's uncertainty principle, which says, intuitively, that one cannot measure accurately both the position and momentum of a particle at the same time. Despite this formulation it seems to be only a fancy way of translating one's ignorance with respect to some quantities of interest. It can be proven that it "actually states a fundamental property of quantum systems and is not a statement about the observational success of current technology" [4].

Focusing on the musical aspect, the role of the composer is essentially to organize musical material in order to make a pleasing sound, evoke some emotion on the listener or simply innovate artistically, giving form and substance to an idea that lives on his mind. Naively, one may think that this process is as easy to perform as it was easy to state it, but as it is common in Mathematics, some of the hardest problems to solve are the easiest to formulate[1]. However, as previously

[1]Recall Fermat's Last Theorem, that was an open problem during 358 years and can be stated as: no three positive integers a, b, and c satisfy the equation $a^n + b^n = c^n$ for any integer value of $n > 2$. Fermat himself claimed that he had a

exemplified, complex procedures could be modeled by stochastic[2] phenomena. Indeed, given a fixed style of composition, there is some "common sense" in the transition of one musical material to another in order to maintain some degree of musical coherence, and this coherence could be translated into probabilistic laws. The knowledge of these laws, which can be stated beforehand or inferred from a *corpus*, could be used to create new sounds or extract information about the respective *corpus* itself.

This text is an expansion of a lecture named "A brief overview of algorithmic composition", presented at the Music School of the Federal University of Rio de Janeiro in October 2019 for an audience composed mainly of musicians. The main goal in writing this note is to present some basic aspects of Probability theory in a mostly intuitive fashion, as well as showing its capabilities in producing new musical pieces and analyzing musical *corpora*, in order to stimulate future investigation in this direction and encouraging researchers to go more deeply into the Mathematics of modern methods not covered in this text (e.g., neural networks [5] and hidden Markov models [29]). This exposition is far from being exhaustive, and several references are given along the text in order to direct the reader to more specific topics of their interest. Since I am a mathematician and only an amateur musician, I apologize in advance if someone is offended by some naive musical analogy. I promise I will do my best!

The text is organized as follows: in Section II we introduce some aspects of Probability theory, its history, and Markov chains in an intuitive way followed by a more technical discussion on the same topics in Sections III and IV, respectively; Section V reviews some applications of Probability and Markov chains in music composition, including a discussion about the probabilistic structures underlying the piece *Analogique A* from Iannis Xenakis; in Section VI three other musical examples are discussed, namely an excerpt of *Brazilian Landscapes No. 20* for bassoon and string quartet by Liduino Pitombeira and inspired by the Binomial distribution, and two excerpts generated from a Markov chain inferred from some of the chorales from J. S. Bach; some words on the relationship between interpretability and flexibility of statistical methods are presented in Section VII, and conclusions are drawn in Section VIII.

II. Elements of Probability theory and Markov chains

In this section we will introduce some basic concepts of Probability theory and Markov chains, one of the first probabilistic model used in algorithmic composition (see [2] for an extensive review). To strictly follow the historical path is usually not the best way to learn Mathematics, but on the other hand, the history of Probability theory is very rich and interesting, and also deeply linked with our habit of using gambling scenarios to gain intuition (or learn that our prior intuition was wrong!) about its fundamental concepts. Therefore, our presentation will be an interchange of historical and more technical information.

Since our scientific history is largely influenced by the development of western civilization, I will mainly refer to developments in Europe and America before the 16th century, even though it is known that advances in Probability and Statistics also occurred in Chinese, Indian, Arabian, Egyptian civilizations, sometimes much earlier than in Europe. For more details, see [6, 25, 13, 26], being [25] also available in Portuguese.

proof of this result in 1637, but wrote in his copy of *Arithmetica*, an ancient text on Mathematics by Diophantus, that it was too large to fit the margin. Nowadays, it is believed that whatever was his proof it was wrong, since the result was only proven in 1994 by Andrew Wiles using very sophisticated mathematical techniques. It is widely believed that it is impossible to prove this result without these tools.

[2] This term comes from the greek $\sigma\tau o\chi o\zeta$, (*stókhos*), which means "aim, guess".

i. Early days of Probability

Our history begins long ago, in the 16th century with Gerolamo Cardano, an Italian polymath and gambler. Motivated by his addiction and lowering funds at the end of his life, he investigated probabilities associated with simple dice rolling games and wrote in 1564 the *Liber de ludo aleae* ("Book on Games of Chance"), which contains the first known systematic treatment of probability. The text was posthumously published in 1663.

Some years later, Blaise Pascal, a french mathematician, was asked to solve a problem, that was noted to be very difficult to solve at that time: what is the fairest way of dividing the stake if a game of chance is interrupted? Pascal noted that the tools necessary to answer this question were not available, and started a sequence of correspondences with Pierre de Fermat, a french lawyer and amateur mathematician.

It is important to note that these early developments were made under the hypothesis that some fundamental probabilities where known *a priori*. For example, Cardano was assuming that the dice used during the game were fair, and the question posed to Pascal assumed the hypothesis that at each round of the game each player was equally likely to win. More generally, as stated in [27], in this scenario we are aware of the mechanisms of the data generating process and we wish to forecast information about the observed data. However, in some problems the real probabilities or other quantities of interest are not known beforehand and must be inferred from observed data: that is, given some observed data we wish to infer information about the data generating process, and this is what we call *Statistics*. A deeper study of Statistics only became possible some years later, after the development of the laws of the large numbers by J. Bernoulli, which will be presented in subsection iii of this Section.

ii. Interlude: A little bit of technicalities

The formal treatment of Probability theory as we know today was developed much later in 1933 by the Russian mathematician Andrey Kolmogorov. However, it is convenient to introduce some terminology and concepts in order to follow more easily the forthcoming discussions [21] (also available in Portuguese).

A *random variable* X is a numerical outcome of some random experiment. In other words, it is a numerical variable whose value depends on a random phenomena. For example, the number observed when rolling a fair die is a random variable that can assume each value in the set $\mathcal{C} = \{1, 2, \ldots, 6\}$ with probability equal to $1/6$, that is, $\mathbb{P}(X = x) = 1/6$, for $x \in \mathcal{C}$.

Formally, the probability of any event is a number between 0 and 1, and this protocol will be followed in this text, except in Section V, where some probabilities will be measured in % to avoid inappropriate rounding to zero when dealing with small numbers. Each time a probability is measured in %, we will be explicit.

Example 1 *One could use a very simple probabilistic model to generate a music. Despite this example being not musically interesting, I will use it later to introduce other probabilistic concepts. Suppose one wishes to compose an infinite-length monophonic piece for piano only with the notes C, C♯, D, D♯ and E*[3]*, chosen at random with equal probability at each time instant*[4]*. If a listener is aware of the procedure employed by the composer, it is easy to see that the probability of a given note be played in a black key is* $2/5$.

[3]The specific pitch of the note is irrelevant here, since only the color of the keys being played are considered.

[4]This restriction is not necessary: the musician can choose the note n_i with some probability $0 \leq p_i \leq 1$, with the single restriction that $p_1 + \cdots + p_5 = 1$. This probability assignment is called the *distribution* of the random variable. However, we will keep the equiprobable scenario for simplicity. Note that in this particular case the random variables are also *identically distributed*, since all of them have the same probability assignment.

Mathematically, one can think that this music is a realization of a *stochastic process*, a sequence $(X_t)_{t\in\mathcal{T}}$ of random variables, where $\mathcal{T}$ contains the onset time of each note, arbitrarily chosen by the composer in principle. This structure is very general and important in Mathematics, but we will restrict ourselves to this particular case, in order to gain more intuition. Each of these random variables X_t assume values on the set $\mathcal{C} = \{\text{C}, \text{C}\sharp, \text{D}, \text{D}\sharp, \text{E}\}$[5], and each note is assumed with equal probability $1/5$. Moreover, *notes played at distinct times are independent*. This is defined as shown in Equation 1:

$$\mathbb{P}(X_{t_1} = n_{t_1}, \ldots, X_{t_k} = n_{t_k}) = \mathbb{P}(X_{t_1} = n_{t_1}) \ldots \mathbb{P}(X_{t_k} = n_{t_k}), \tag{1}$$

and this equality should hold for all $k \geq 2$ and $n_{t_1}, \ldots, n_{t_k} \in \mathcal{C}$, if $t_1 < t_2 < \cdots < t_k$. That is, the probability of playing k specific notes $n_{t_1}, \ldots, n_{t_k}$ at specific time instants $t_1 < \cdots < t_k$ is the product of the individual probabilities, for all choices of notes, time instants and quantities of notes. However, one does not need to be a musician to know that the vast majority of musics are not composed with total randomness like this! Let us go back to history in order to introduce more structure in our model.

iii. Calculus, Probability and the laws of large numbers

The formal development of calculus in the late 17th century by Issac Newton and Gottfried Leibniz[6] allowed immense advances in several areas of science, being Physics perhaps the most notable one, mainly because of its intimacy with the ideas developed by Newton and Leibniz themselves. Obviously, Probability theory also took advantage of this new tool, mainly in two works: "The doctrine of chances: or, a method for calculating the probabilities of events in play" published in 1718 by the french mathematician Abraham de Moivre and *Ars Conjectandi* ("The Art of Conjecturing") by Jacob Bernoulli, posthumously published in 1713.

The work of J. Bernoulli contains the first step towards one of the greatest achievements of Probability theory, the *laws of the large numbers*, a result for which he devoted about 20 years of his life. Recall the composition in Example 1, but suppose now that a listener is not aware of the procedure employed by the composer and wishes to *infer* the probability of a given note being played in a black key. Still assuming that different notes are independent, a very careful listener can listen the piece for an arbitrarily long time and at each note write down if it comes from a black or white key. Mathematically, the listener is observing a realization of *another* stochastic process $(Y_t)_{t\in\mathcal{T}}$, where each random variable Y_t assumes values in the set $\mathcal{D} = \{\text{white}, \text{black}\}$ and are also independent. However, he does not know the probability of observing black or white outcomes. Intuitively, he can simply count the occurrences of black keys and divide it by the total of notes being played and hope that this number will be close to the true proportion of $2/5$. This is a prototypical example of an *statistical inference* procedure, a scenario where one is interested in estimating quantities from observed phenomena instead of studying the possible outcomes of random variables knowing its distribution *a priori*. More generally, as stated in [27], now one is interested in in inferring information about the data generating process from observed data.

It is quite intuitive that if more notes are being played, more accurate is the listener's estimate. Indeed, as J. Bernoulli itself stated in his book in a quite presumptuous manner: "[even] the most stupid of men [...] is convinced that the more observations have been made, the less danger there

[5] Note that this does not contradict our definition of a random variable assuming only numerical values, since one can easily map this set to a numerical one, for example, its respective MIDI numbers. For a moment, we will make this abuse of notation for the sake of clarity.

[6] "Standing on the shoulders of giants", as said by Newton himself, since ideas of limits and integrals existed since the Greek mathematicians.

is of wandering from one's aim" [17] (also available in Portuguese). Again, this is an example of a mathematical result that is easy to formulate, at least intuitively, but quite hard to prove.

J. Bernoulli called this result the *law of the large numbers* and nowadays this name is given to an entire class of results about asymptotic behavior of sequences of random variables. We will not dive too much into this topic, but it is important to note two forms of the law, which will be stated in the scenario of our example:

- Weak law of large numbers: with a sufficient large sample of observed notes, the proportion computed by the listener is very likely to be close to the true value or $2/5$;
- Strong law of large numbers: the probability of observing an infinite song that leads to an incorrect estimation of the true value of $2/5$ is null.

On his work, J. Bernoulli only have proved the weak law of large numbers for a specific kind of random variable, and most important for us, *under the hypothesis of independence*, crucial to its proof. Nowadays, it is known that a very large class of random variables satisfy the laws, but the majority of them also assuming independence, being the Markov chains the first class which breaks this hypothesis [24].

The history of the laws of the large numbers is very rich and intimately linked to the development of the formal theory of Probability in the beginning of the 20th. century. For more details see [26, 17].

iv. Markov chains

In order to introduce the specific kind of dependence between random variables in a Markov chain, let us change somewhat the composition in Example 1.

Example 2 *Let us go back again to the composition we made in Example 1 and try to improve it imposing more structure. Instead of choosing randomly a note from the set $\mathcal{C}$, our notes now could be taken from the sets $\mathcal{C}_1 = \{G, G\sharp\}$ and $\mathcal{C}_2 = \{C, C\sharp, D, D\sharp, E\}$, using the following rules:*

1) The first note n_1 is chosen at random from set $\mathcal{C}_2$
2) For $k = 2, \ldots, N$, where N is the number of notes the composer desires in his song:
   i) If the previous note n_{k-1} comes from a black key, the note n_k is chosen at random from the set $\mathcal{C}_1$;
   ii) Else, if the previous note n_{k-1} comes from a white key, the note n_k is chosen at random from the set $\mathcal{C}_2$.

Assuming the same onset times as before, denote this song as a realization of another stochastic process $(X'_t)_{t\in\mathcal{T}}$. From the procedure above, it is clear that these random variables are not independent anymore, since a note depends on a feature of the previous one.

But assume that the listener also wishes to estimate the probability of a given note be played in a black key. He could repeat the same procedure as before, listening an arbitrarily long excerpt of the music, writing down if each note being played is black or white and computing the proportion of black keys being played. However, since the random variables being observed are not independent anymore, there is no guarantee that the listener's procedure will be close to the true proportion being estimated, for the result of J. Bernoulli needed the independence assumption.

In 1902, the Russian theologian and mathematician Pavel Nekrasov claimed that independence was a *necessary* condition for the laws of the large numbers to hold, that is, there is no possibility that our listener's estimate will be accurate in this new scenario. Being his work grounded not on

Mathematics itself but only in religious principles of predestination and free will, there was a lot of margin to more rigorous discussion on his claims.

Indeed, motivated by its mathematical accuracy and personal disputes with Nekrasov, the Russian mathematician Andrey Markov initiated a detailed study on dependent sequences of random variables, aiming the extension of J. Bernoulli's work to this new scenario. In 1906 he published his work which title can be loosely translated as "Extension of the law of large numbers to quantities that depend on each other", containing the beginning of an entire new theory in the field of Probability theory and proving that under some hypothesis, sequences of dependent random variables *can* satisfy the laws of the large numbers [24] . It is curious to note that even though Markov created its chains merely as a counterexample to Nekrasov's claim, its importance in applied fields is enormous [22].

Intuitively, Markov chains are stochastic processes such that at a particular time instant t, the random variable X_t is purely determined by its immediate preceding observed value plus a random effect, independent X_t. This is exactly the scenario our listener is faced with: in order to determine a given note, it is necessary only to know if the previous note came from set $\mathcal{C}_1$ or $\mathcal{C}_2$, plus a random choice which only depends on the set. We devote now some time formalizing some important concepts used from here on.

III. Some technical aspects of Probability theory

Recall that a random variable X was defined as a numerical outcome of some random experiment and probabilities associated with X are denoted using the letter $\mathbb{P}$. A particular assignment of probabilities to subsets of real numbers *via* X is called the *distribution* of the random variable. Some distributions of random variables are remarkably important and appears in many distinct and apparently uncorrelated scenarios that special names are given to them. We will now clarify this definition on random variables and present some examples.

Firstly, note that we can classify a random variable in two main classes: *discrete* or *continuous*. Discrete random variables assumes its values in finite or countable sets[7] of real numbers, and we can loosely say that continuous random variables assumes their values in uncountable sets. Despite this not being the formal definition of continuous random variables, the precise definition is quite involved and we will keep this intuition for a moment.

i. Discrete random variables

Discrete random variables are simplest than continuous one, since its manipulation requires only, in general, the basic arithmetic operations.

Example 3 *Some examples of important discrete distributions are:*

- *Bernoulli distribution: This is the simplest distribution of a random variable, and receives the name of J. Bernoulli since it was very important on his studies in the laws of the large numbers. We say that X is a Bernoulli random variable with parameter p if it assumes only the values 0 or 1, with respective probabilities $1-p$ and p, that is,*

$$\mathbb{P}(X=0) = 1-p \tag{2}$$

$$\mathbb{P}(X=1) = p. \tag{3}$$

[7] A set A of real numbers is said to be *countable* if there is an one-to-one correspondence between A and $\mathbb{N}$, the set natural numbers. Intuitively, A is countable if its elements can be counted. The sets of integer numbers $\mathbb{Z}$ and rational numbers $\mathbb{Q}$ are examples of countable sets, whereas the interval $[0,1]$, the set of irrational numbers $\mathbb{R}\setminus\mathbb{Q}$ and the set of real numbers itself $\mathbb{R}$ are examples of non-countable sets.

It is used to model the outcome of binary experiments, and the values 0 and 1 are usually denoted as "failure" and "success", respectively. The sentence "X has a Bernoulli distribution with parameter p" is abbreviated as $X \sim Bern(p)$.
Note that we already came across a Bernoulli random variable: recall Example 1 where each random variable Y_t indicates if the note X_t is played in a black key or not. Assuming that observing a black key is a success, we can state now, more formally, that $Y_t \sim Bern(2/5)$.

- *Binomial distribution: It is a generalization of the Bernoulli distribution, where a sequence of independent trials of a binary experiment with a probability of success equal to p is repeated n times and the quantity of successes are computed. Its probability function is given by Equation 4:*

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, \ldots, n, \tag{4}$$

where $\binom{n}{k}$ is given by $\frac{n!}{k!(n-k)!}$. Intuitively, this formula says that $p^x(1-p)^{n-x}$ is the probability of observing a particular sequence of x successes and $n-x$ failures, and the term $\binom{n}{k}$ accounts for the distinct ways we can organize these successes and failures. Finally, since n repetitions of the experiment are performed, one can obviously observe only 0 to n successes. We denote then $X \sim Bin(n, p)$.

- *Poisson distribution: Named after the French mathematician Siméon Poisson, it models the number of events of interest occurring in a fixed interval of time or space, assuming that these events occur with a known constant rate and are independent of the time since the last observation. Its probability function is given by Equation 5:*

$$\mathbb{P}(X = x) = \frac{\lambda^x e^{-x}}{x!}, \text{ for } x = 0, 1, 2, \ldots, \tag{5}$$

where e is the Euler number[8]. The parameter λ is the average numbers of occurrence of the event of interest in a fixed interval of time. We abbreviate $X \sim Poisson(\lambda)$.
This distribution appears quite naturally when one tries, for example, to model the telephone calls arriving in a system per time unit, the number of mutations on a strand of DNA per unit length, number of decays in a given time interval in a radioactive sample, among others.

More generally, a discrete random variable X assumes its values on a finite or countable set, here denoted by $\{x_1, x_2, \ldots\}$. The *distribution* of X is the particular assignment of probabilities of X to these numbers, that is, the values $\mathbb{P}(X = x_i) = p_i$, with the single restrictions that $p_i \geq 0$ for all $i = 1, 2, \ldots$ and $\sum_{i=1}^{\infty} p_i = 1$. The function $f_X(x) = \mathbb{P}(X = x)$ is called the *probability function* of X.

ii. Continuous random variables

A continuous random variable X, on the other hand, posses a *probability density function*, that is, a function $f_X : \mathbb{R} \to \mathbb{R}$ such that $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and satisfying the Equation 6:

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1. \tag{6}$$

Therefore, the value $f_X(x)$ does not represent the probability of observing X assume the value x, but the *density* of the probability on the value x, and the probability of observing X between

[8]The appearance of the Euler number in this probability function is not obscure, since it can be proven that the Poisson distribution is approximately the same as the Binomial distribution with n large and small p, with $\lambda = np$, and it appears quite naturally in the derivation. For more details, see [21].

values a and b is given by the area below the graph of f_X, from a to b (Equation 7):

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)\, dx. \tag{7}$$

It is important to note now that individual values have null probability of occurrence: that is, $\mathbb{P}(X = a) = \int_a^a f_X(x)\, dx = 0$, for all $a \in \mathbb{R}$. This is an apparent paradox, since we *can* observe $X = \pi$, for example, but this event has null probability of occurrence! It can be easily solved once we recall that a real number possess *infinite* decimal places, and any measurement device we have invented have a finite precision, that is, detects only a finite amount of decimal places: even if $X = \pi$ we may only observe its rounding on the first two decimal places $X \approx 3,14$, that is the same as saying $3,135 < X < 3,144$, an event with positive probability of occurrence.

Contrarily to discrete random variables, operating with continuous random variables it more involved, since it requires tools from Calculus, such as derivation and integration.

Example 4 *Some important examples of continuous random variables are:*

- *Normal distribution: Perhaps the most important probability distribution of all! Its first appearance was in the work of Abraham de Moivre, but it gained more visibility after Carl Gauss' work in 1809 about the movement of the planets around the Sun. He wished to estimate the orbit of Ceres, a dwarf planet between Mars and Jupiter, from a few observations and needed a way to model inaccuracies on its observations. He claimed that the "fairest" way of modeling errors in experiments is via the Normal distribution, whose probability density function is given by Equation 8:*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for all } x \in \mathbb{R}. \tag{8}$$

  The parameters μ and σ^2 are called its mean and variance, respectively, and controls where the distribution is centered and its spreadness, respectively. We denote normality as $X \sim \mathcal{N}(\mu, \sigma^2)$. For more details about the history of the Normal distribution see [26].
- *Exponential distribution: The Exponential distribution is intimately related with the Poisson distribution we saw before. Consider the particular example when one uses the Poisson distribution with parameter λ to model the number of decays in a given time interval in a radioactive sample, and supposes that one is interested in the distribution of the time interval between consecutive radioactive emissions. Obviously this random variable must be continuous, and one can prove [21] that it has the Exponential distribution, whose probability density function is given by Equation 9:*

$$f_X(x) = \lambda e^{-\lambda x}, \text{ for all } x > 0, \tag{9}$$

  this fact being denoted as $X \sim Exp(\lambda)$.
- *Maxwell-Boltzmann distribution: This is a very important probability distribution in Statistical Mechanics, which describes the distribution of speeds of molecules from a gas at a certain temperature. Its probability density function is given by Equation 10:*

$$f_X(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{a^3} e^{-\frac{x^2}{2a^2}}, \text{ for all } x > 0. \tag{10}$$

iii. Expected values

The distribution of a random variable X provides us all the probabilistic information we may need. However, sometimes all we need is some *numerical summary* of it. For example, if X models the

number of radioactive particles emitted by some source in an interval of one hour, it is reasonable to model it *via* a Poisson distribution. Recall that parameter λ refers to the rate of emission per unit of time (one hour, in this example). The whole information about the random variable allows us to compute, for instance, $\mathbb{P}(X > x_0)$, assuming that the value x_0 is important for heath security reasons. But recalling the interpretation of parameter λ, instead of informing the whole probability distribution of X, it may be enough to compare the value of λ with x_0.

Indeed, our former character Blaise Pascal noted the importance of single numerical summaries of random variables in November 23rd. of 1654, when he wrote a small note that he kept in his pocket for the last 8 years of life describing how "God came to him and set him free from the corrupted ways" [17]. When considering the pros and cons of his duties with God, he created a way of computing these quantities, which is now called the *expectation*, *expected value* or *mean* of a random variable. These names are used interchangeably in the literature, but here we will use only *expected value*. We define this quantity and many others only for discrete random variables, for simplicity. The interested reader may refer to [21] for more details.

Indeed, if X is a discrete random variable, its expected value is defined as shown in Equation 11:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i \mathbb{P}(X = x_i), \tag{11}$$

assuming that this summation is finite. The intuition behind this formula is quite simple: The value x_i is weighted by its probability of observance, and the final result is an weighted average of the values X can assume, interpreted as its "central value", again making an analogy with Physics, more specifically, the concept of center of mass.

Example 5 *Let us verify that some intuitive facts about expected values indeed hold:*

- *One can prove that if $X \sim Poi(\lambda)$, then $\mathbb{E}[X] = \lambda$, which formalize our claims on the example in the beginning of this Section.*
- *Recall Example 1. Let X denote the amount of black keys being played in n notes of our composition. Since the probability of a single note be played in a black key is $2/5$, we have that $X \sim Bin(n, 2/5)$, and one expects that approximately $2n/5$ notes played in black keys will be observed. Indeed, it can be show that if $X \sim Bin(n, p)$, then $\mathbb{E}[X] = np$.*

However, one does not always want to compute the expected value of X, but of some function $g(X)$ of X. For example, if X is a random variable denoting the radius of a circle, its area will also be a random variable, given by $g(X) = \pi X^2$. The *law of the unconscious statistician* [21] allows us to compute such expectations, without knowing the distribution of $g(X)$, which can be quite involved of obtaining, *via* the simple formula (Equation 12):

$$\mathbb{E}[g(X)] = \sum_{i=1}^{\infty} g(x_i) \mathbb{P}(X = x_i). \tag{12}$$

This Theorem allows us to define another important summaries of random variables, like the *variance*, denoted by $\mathbb{V}(X)$, which measures the spreadness of a random variable around its expected value (Equation 13):

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{i=1}^{\infty} (x_i - \mathbb{E}[X])^2 \mathbb{P}(X = x_i). \tag{13}$$

iv. Conditional probability

A necessary concept to properly define the Markov chains is the *conditional probability*, which essentially measures the probability of some event given that another event has occurred. More specifically, the probability of occurrence of the event A given that event B has occurred is called the *conditional probability of A given B*, denoted by $\mathbb{P}(A|B)$ and computed by Equation 14:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}. \tag{14}$$

This definition can be understood essentially as the restriction of the possible outcomes of the experiment, given that event B has occurred.

In order to clarify this concept, let us see it in the formerly presented examples.

Example 6 *In Example 1, since the notes being played are independent, we have that*

$$\begin{aligned}
\mathbb{P}(X_t = C\sharp | X_{t-1} = E) &= \frac{\mathbb{P}(X_t = C\sharp, X_{t-1} = E)}{\mathbb{P}(X_{t-1} = E)} && (15)\\
&= \frac{\mathbb{P}(X_t = C\sharp)\mathbb{P}(X_{t-1} = E)}{\mathbb{P}(X_{t-1} = E)} && (16)\\
&= \mathbb{P}(X_t = C\sharp) && (17)\\
&= 1/5, && (18)
\end{aligned}$$

that is, the knowledge of the previous note being an E does not change the probability of the next note be a C♯. Note that the particular choice of C♯ and E is not relevant and the same result will be obtained for any two notes in set $\mathcal{C}$.

However, in Example 2, the quantity $\mathbb{P}(X_t = C\sharp | X_{t-1} = x)$ depends on the particular choice of note x: if $x = E$ this probability is $1/5$, since E is played in a white key and this implies that the note X_t is chosen from set $\mathcal{C}_2$; on the other hand, if $x = D\sharp$ this probability is zero, because in this scenario the note X_t will be chosen from set $\mathcal{C}_1$, which does not contain note C♯.

IV. Markov chains

As stated beforehand, the study of stochastic processes, i.e. sequences of random variables, is quite important in Probability theory, and we already came across the Markov chains in Example 2. We introduce it now more formally and then returns to analyze in more details this introductory example.

i. Basic definitions

Intuitively, a Markov chain is like a Monopoly match: a random walk on a finite set $\mathcal{C}$, where the next step depends only on where we are now. More formally, a sequence of random variables $(X_t)_{t\in\mathbb{N}}$ is a *Markov chain* if it assume values on a common set $\mathcal{C}$ and satisfies the Equation 19:

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_2 = x_2, X_1 = x_1) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n), \tag{19}$$

for all $n \in \mathbb{N}$ and $x_{n+1}, \ldots, x_1 \in \mathcal{C}$. The elements of set $\mathcal{C}$ are called *states* of the chain and we denote the *transition probabilities* $\mathbb{P}(X_{n+1} = x_j | X_n = x_i)$ as $p_{ij}(n)$. These probabilities must be interpreted as the probability of going to state x_j in time instant $n+1$ given that in time instant n the position is state x_i.

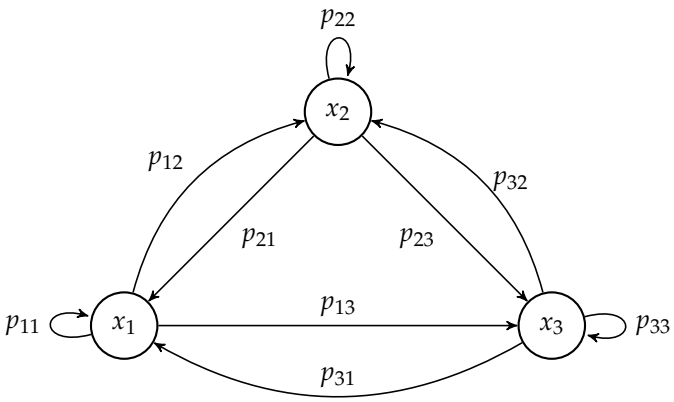

Figure 1: *Oriented graph associated with a Markov chain with states* $\mathcal{C} = \{x_1, x_2, x_3\}$ *and transition matrix in Equation 21.*

Note that in the previous definition we allowed the transition probabilities to depend on the time instant n, as if in the game of Monopoly we roll dice with different number of faces along the game. Chains with this characteristic will not be treated here, and we consider only the *homogeneous* ones, where the transition probabilities does not depend on time instant n, and we can simply write (Equation 20):

$$\mathbb{P}(X_{n+1} = x_j | X_n = x_i) = p_{ij}. \tag{20}$$

Since we will not deal with non-homogeneous chains, this adjective will be omitted from now on.

It is very instructive to represent Markov chains visually as an oriented graph, where the nodes are the states and the arrows indicates the allowed transitions, together with its respective probability of occurrence, as can be seen in Figure 1. The transition matrix of this chain, a concept to be shortly introduced in the beginning of next subsection, is given by

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}. \tag{21}$$

ii. Convergence of a Markov chain

A very important object associated to a Markov chain is its *transition matrix*: a square matrix $\mathbf{P}$ of size $M \times M$, where M is the cardinality of $\mathcal{C}$, whose element on the i-th row and j-th column is p_{ij}. The transition matrix uniquely determines the chain, and contains important information about its asymptotic behavior, which will lead us to a version of the laws of the large numbers for Markov chains. We will develop the reasoning to naturally arrive at the analogous result in this scenario.

Let $\boldsymbol{\pi}_1$ be an $M \times 1$ vector, containing the initial distribution probability of the chain, that is, the i-th entry of $\boldsymbol{\pi}_1$ is $\mathbb{P}(X_1 = x_i)$, the probability that the chain starts at state x_i, for all $i = 1, \ldots, M$. Note that in a Monopoly match the initial location of the players on the board is precisely determined by the game rules', but recalling Example 2, out first note could come from a random choice between sets $\mathcal{C}_1$ and $\mathcal{C}_2$. The distribution of the second step of the chain, X_2, is

then given by Equation 22:

$$\mathbb{P}(X_2 = x_j) = \sum_{i=1}^{M} \mathbb{P}(X_2 = x_j | X_1 = x_i)\mathbb{P}(X_i = x_i), \tag{22}$$

for all $j = 1, \ldots, M$. Denoting the vector containing this information by $\boldsymbol{\pi}_2$, this is exactly the vector-matrix multiplication (Equation 23):

$$\boldsymbol{\pi}_2 = \boldsymbol{\pi}_1 \mathbf{P}. \tag{23}$$

It is easy to see, *via* recursion, that the distribution of the n-th step of the chain, X_n, is given by Equation 24:

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_{n-1}\mathbf{P} = \boldsymbol{\pi}_1 \mathbf{P}^{n-1}, \tag{24}$$

for all $n = 1, 2, \ldots$.

Recalling the intuition about Monopoly again, assume that some player decided to start not on the standard place, but somewhere on the opposite side of the board, for example. In the long-term it will be possible to determine where this player have started, or this information will be "forgotten" in the course of time? More formally, the distribution of X_n will eventually "forget" how the chain has started? Markov proved in 1906 that if the chain is *ergodic*, then the initial distribution will be little by little forgotten, that is,

$$\lim_{n \to \infty} \boldsymbol{\pi}_n = \boldsymbol{\pi}, \tag{25}$$

and this limit is independent of the initial distribution $\boldsymbol{\pi}_1$ of the chain.

Vector $\boldsymbol{\pi}$ is called the *stationary distribution* of the chain, and its i-th entry represents the proportion of time that the chain spends on state x_i in the long-term, and Markov also proved that it can be estimated as the average of time spent in this state in a realization of the chain, regardless of the initial state $\boldsymbol{\pi}_1$ being considered, disproving once and for all the non-rigorous argument of Nekrasov.

Ergodicity is another name borrowed from Physics, and it means, intuitively, that the chain is sufficiently connected, and its definition is beyond the scope of this text [22]. It is possible to prove that the chain is ergodic if some power of its transition matrix $\mathbf{P}$ has only positive entries. Following a reasoning analogous to the one in Equation 24, it can be proven that the entry in line i and column j of $\mathbf{P}^n$ is given by Equation 26:

$$[\mathbf{P}^n]_{ij} = \mathbb{P}(X_{n+1} = x_j | X_1 = x_i), \tag{26}$$

that is, the probability of going from state x_i to x_j in exactly n steps.

Example 7 *In order to clarify these new definitions, let us return to Example 2, building a Markov chain and studying its asymptotic properties.*

First note that we have several elements to choose as the states of the chain: the notes itself, the color of the key being played or the sets $\mathcal{C}_1$ and $\mathcal{C}_2$, abbreviated via its indices for simplicity. For this example we choose the latter. From the compositional rule therein stated, it is easy to verify that

$$\mathbb{P}(X_{n+1} = 1 | X_n = 1) = \mathbb{P}(\textit{playing a note from a black key in set } \mathcal{C}_1) = 1/2 = p_{11} \tag{27}$$
$$\mathbb{P}(X_{n+1} = 2 | X_n = 1) = \mathbb{P}(\textit{playing a note from a white key in set } \mathcal{C}_1) = 1/2 = p_{12} \tag{28}$$
$$\mathbb{P}(X_{n+1} = 1 | X_n = 2) = \mathbb{P}(\textit{playing a note from a black key in set } \mathcal{C}_2) = 2/5 = p_{21} \tag{29}$$
$$\mathbb{P}(X_{n+1} = 2 | X_n = 2) = \mathbb{P}(\textit{playing a note from a white key in set } \mathcal{C}_2) = 3/5 = p_{22}, \tag{30}$$

and therefore, the corresponding transition matrix is given by

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/2 \\ 2/5 & 3/5 \end{bmatrix}. \tag{31}$$

Since all of its entries are strictly positive, the chain is ergodic and Markov's theorem guarantee the existence of the stationary distribution. It can be proved that as n increases, matrix $\mathbf{P}^n$ becomes closer to

$$\begin{bmatrix} 4/9 & 5/9 \\ 4/9 & 5/9 \end{bmatrix}, \tag{32}$$

regardless of the initial distribution of the chain, that for completeness of the example, is given by $\boldsymbol{\pi}_1 = [0\ 1]$, since our first note surely comes from set $\mathcal{C}_2$.

Therefore, independently on where our chain begins, in the long-term, our note has a probability of $4/9$ of being chosen from set $\mathcal{C}_1$ and of $5/9$ of being chosen from set $\mathcal{C}_2$. Since this chain is ergodic, the procedure employed by the listener indeed works, and his estimate will be, with high probability, quite close to $5/9$, provided that he listens the song for enough time.

iii. Learning the transition probabilities

The first widely known practical use Markov chains was in 1913, where Markov itself analyzed a sequence of 20,000 characters of the poem *Eugene Onegin* from Aleksandr Pushkin, inferring the transition probabilities between symbols by merely counting these transitions. Assuming that the chain is indeed homogeneous, this is known in Statistics as the *maximum likelihood estimator* [7]. For example, the probability that a vowel precedes another vowel is the ratio between the observed number of this transition and the total number of occurrences of vowels along the text.

Despite being quite intuitive this procedure possess some drawbacks, in particular when the observed sequence is not sufficient long and a big number of states is being considered, since in this scenario some transition probabilities could be quite underestimated. Techniques to overcome this and other difficulties exist, and are called *smoothing* procedures, that are beyond the scope of this text [16].

iv. Higher order Markov chains

Markov chains of higher order $N > 1$ can also be considered, that is, chains where the dependence of the observation in time instant n depends not only on X_{n-1} but also on $X_{n-2}, \ldots, X_{n-N}$.

Perhaps Markov preferred to use a higher-order chain to obtain more realistic information about Pushkin's poem, but in the absence of digital computers, estimating the transition probabilities became rapidly infeasible. For example, when considering a second-order chain, the transition matrix is not a matrix anymore, but a *tensor or order three*, a three-dimensional matrix $\mathbf{P}$ whose entry $\mathbf{P}_{ijk}$ represents the probability of the transition $x_i \to x_j \to x_k$. Whereas in usual Markov chains one needs to estimate M^2 transition probabilities, in second order chains this quantity increases to M^3 and it is easy to see that for a N-order chain, its respective tensor of order $N+1$ has M^{N+1} entries to be estimated!

V. Markov chains in music composition and analysis

Since its first use in algorithmic composition in the decade of 1950 [18][9], several applications and extensive well-written reviews can be found in the literature [2], and I specially recommend [16].

[9] Despite the paper being dated from 1961, the development of their work began in the early 1950, being published only a decade later.

In this section I will only briefly recall some remarkable applications and present in more detail the structure of *Analogique A*, by Iannis Xenakis [28].

i. A (very) small literature review

The first important aspect to note is that the notion of "state" is quite flexible to accommodate several musical aspects of interest. For instance, one can use a Markov chain to model transitions of order N between pitches, intensities, intervals, chords, vectors containing combinations of these features, among many others. Assuming that this modeling was performed intending the creation of a new musical piece[10], we came to the second point: will the transition probabilities be estimated from a *corpus* of interest or fixed beforehand, aiming some desired behavior? Combinations between these two aspects essentially classifies the works using Markov chains to create new musical material.

In the beginning of the decade of 1950, Harry Olson and Hebert Belar were the first researchers that used Markov models in algorithmic composition, developing the first known machine called a "synthesizer" [18]. Basically, by analyzing a *corpus* of 11 melodies by Stephen Foster, transposed to D major, they estimated transition probabilities of order 0[11], 1 and 2 between pitches of notes and of order 0 between rhythmic patterns in time signatures of $\frac{3}{4}$ and $\frac{4}{4}$. The notes of the new composition were generated accordingly to both models for pitch and rhythm simultaneously, reproduced by a speaker and recorded in a magnetic tape.

One of the main disadvantages of the Markov model in algorithmic composition is that it only captures short-term dependencies. For example, assume that the states of the chain are pitches, and that the probability transitions were estimated from a sufficiently large *corpus*, in order to overcome the difficulty of estimating transition probabilities in high-order chains. If the order being used is low, the learned model is quite simple and the new generated melodies usually does not resemble the *corpus* that one desires to mimic. However, increasing the order does not solve the problem, since only still short-term structures are being captured, only the definition of "short" being just a little enlarged. Therefore, it is possible that with a higher-order model (around 10, for example), the generated melodies will be only a plagiarism of some content already present in the *corpus*, a behavior one does not wish to observe, if the creation of new musical material is the researcher's goal.

However, the former paragraph does not mean that the Markov model needs to be abandoned at all! Instead of applying it to low-level musical aspects such as pitch and duration of notes, some researchers obtained good results in another structures, such as chord classes [1], harmony [20], interval between pair of notes [12] among others ([16, p. 24-43]). Indeed, forwarding somewhat in time, in 2002 François Pachet proposed the *Continuator* [19], a system that is not intended to create new music material, but to continue musical phrases played by a musician in a MIDI controller, responding to them in the same style of the excerpt being played, by learning transitions between pitches, velocity, beginning and length of the notes, among others. His work does not employed the usual transition matrices, but another structure called *prefix tree*, which can be proven to be equivalent of a varying-order Markov model. By only "filling the gaps" and composing short excerpts of music, his approach masks the drawbacks previously stated.

[10] One also could be interested in analyzing the style of a composer, for example, by means of its transition matrix.

[11] A Markov model of order 0 consider only the relative frequency of states within a *corpus*, as the black and white keys in Example 1.

ii. Iannis Xenakis' *Analogique A*

In 1958 and 1959, previously to the publication of the work of Olson and Belar [18], Iannis Xenakis, a Greek-French composer, music theorist, architect, performance director and engineer has pioneered the use of Markov chains in music composition. Moreover, instead of estimating the transition probabilities from a *corpus*, he fixed beforehand the transition probabilities and with calculations performed by hand[12] he created entirely new musical material in *Analogique A* and *Analogique B*, for string orchestra and sinusoidal sounds, respectively. This process is fully described in Chapters 2 and 3 of [28], with a quite cumbersome and intricate language, from both mathematical and musical viewpoints. In this subsection we briefly describe some remarkable aspects of this material, with a more accessible and updated terminology, mainly focused on *Analogique A* since the mathematical formulation of both compositions are similar. An expansion of this subsection to fully update some excerpts of [28] is addressed as a future work.

The first question posed by Xenakis is essentially about the adequate musical object over which impose the Markov structure. This leads him to a deep analysis of the nature of sounds, from which this quote is particularly interesting:

> All sound is an integration of grains, of elementary sonic particles, of sonic quanta. Each of these elementary grains has a threefold nature: duration, frequency, and intensity. All sound, even all continuous sonic variation, is conceived as an assemblage of a large number of elementary grains adequately disposed in time. [...] In the attack, body, and decline of a complex sound, thousands of pure sounds appear in a more or less short interval of time, Δt. Hecatombs of pure sounds are necessary for the creation of a complex sound. ([28, p. 43-44]).

It becomes clear that Xenakis was aware of the work of Dennis Gabor, a Hungarian-British electrical engineer and physicist who received the Nobel Prize in 1971, about time-frequency analysis and the nature of sound [11], since Xenakis' claim is essentially the same as saying that the sonic content in a given short interval of time can be decomposed as a superposition of more simple sounds. This is the principle behind the *spectrogram*, a visual representation of the frequency spectrum of a sound as it varies with time. Figure 2 illustrates a spectrogram of the subject of *Ricercar a 6*, the six-voice fugue from *The Musical Offering* from J. S. Bach.

Xenakis assumes that the length of the time frame being analyzed, Δt, is small but invariable, in order to ignore it and consider only the components of frequency and intensity, which he denotes as F and G, respectively. It is important to note that not every possible combination of F and G is audible to the human ear[13], a point that he is still aware, and without loss of generality, the audible region of the FG plane can be put in a one-to-one correspondence with a rectangle, as illustrated in Figure II-6 in [28, p. 49]. This leads us to his definition of *screen*:

> The screen is the audible area (FG) fixed by a sufficiently close and homogeneous grid [...], the cells of which may or may not be occupied by grains. In this way, any sound and its history may be described by means of a sufficiently large number of sheets of paper carrying a given screen S. These sheets are placed in a fixed lexicographical order. ([28, p. 51])

[12] The invention of the digital computer dates from the decade of 1950, but their widespread use was only possible several decades later.

[13] The way we process and perceive sounds is the main object of study of *psychoacoustics* [30]. The curves in Figure II-6 in [28, p. 49] are called the *equal loudness curves*. Psychoacoustics is quite important nowadays, since several lossy audio codecs such as MP3 depends on this theory in order to properly "throw away" information we do not perceive.

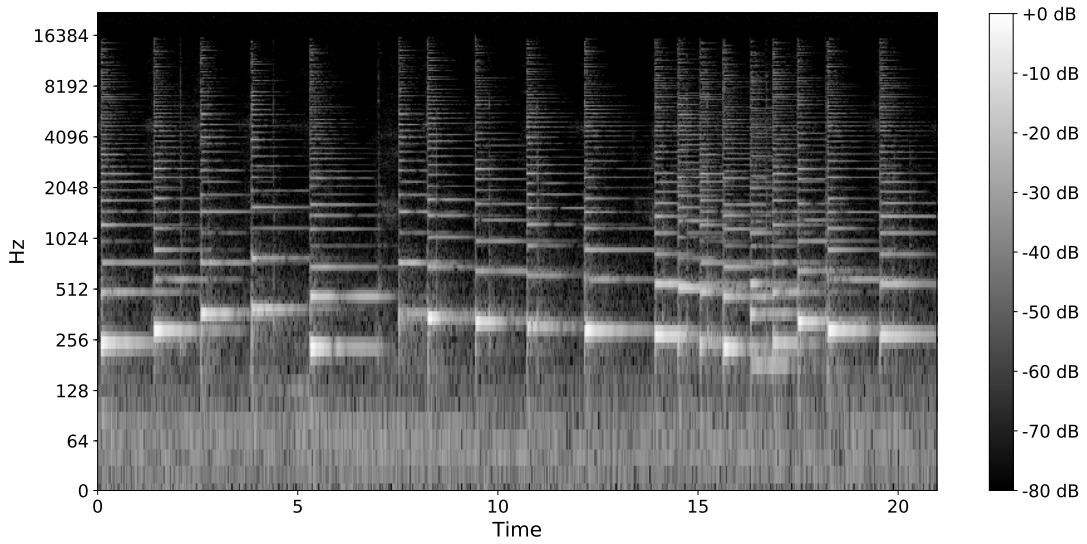

Figure 2: *Spectrogram of the subject of* Ricercar a 6. *Each point in the figure represents the intensity (in dB) of the respective frequency at time t (y and x coordinates of the point, respectively).*

Recalling again Figure 2, each vertical line can be understood as a screen, and the ordered set of screens forms the whole sound. In Figure 3 we can see one particular screen associated with this spectrogram.

One last variable that Xenakis considers important is the density of grains per unit of volume $\Delta F \Delta G \Delta t$, denoted by D. He suggests that it should also be measured in logarithmic scale, with base between 2 and 3. This particular choice is due to the fact that later in Chapter 2 Xenakis compares the density of grains, the complexity of a music and the *entropy* of a screen[14]. Along much of this Chapter he discuss several aspects of the screens and draws many parallels with Physics. In particular, he models the density of grains per screen *via* the Poisson distribution, and imagining that the continuous evolution of grains along time is similar to the movement of the molecules on a gas, its interaction can also be described by the Maxwell-Boltzmann distribution. He also operate screens with the usual operators of set theory like union, intersection and complements. Although very interesting, these discussion are outside the scope of this work, since they are not critical to have a good understanding of the stochastic structure of *Anaogique A*.

When analyzing the way screens are linked, he propose that transitions can be stochastic, in particular, following a Markov chain. Therefore, from Xenakis' viewpoint, screens are the most general object one can impose a Markov structure in order to create new music. However, as noted earlier in Chapter 2 *via* another parallel with Physics, the manipulation of individuals grains in order to achieve this goal is infeasible:

> Theoretically, a complex sound can only be exhaustively represented on a three-dimensional diagram F, G, t, giving the instantaneous frequency and intensity as a function of time. But in practice this boils down to saying that in order to represent a momentary sound, such as a simple noise made by a car, months of calculations and graphs are necessary. This impasse is strikingly reminiscent of classical mechanics, which claimed that, given sufficient time, it could account for all physical and even

[14]The entropy of a random variable will briefly appear in Section VII, and it is also measured in logarithmic scale, usually in base 2.

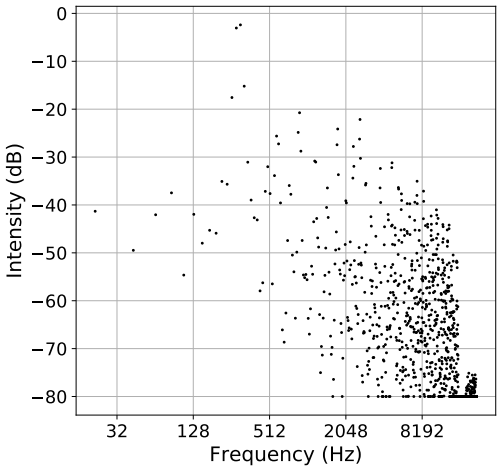

Figure 3: *Vertical line of the spectrogram in Figure 2, representing a screen.*

> biological phenomena using only a few formulae. But just to describe the state of a gaseous mass of greatly reduced volume at one instant t, even if simplifications are allowed at the beginning of the calculation, would require several centuries of human work! [...]
>
> The same thing holds true for complex as well as quite simple sounds. It would be a waste of effort to attempt to account analytically or graphically for the characteristics of complex sounds when they are to be used in an electromagnetic composition. For the manipulation of these sounds macroscopic methods are necessary. [...]
>
> Microsounds and elementary grains have no importance on the scale which we have chosen. Only groups of grains and the characteristics of these groups have any meaning. ([28, p. 49-50])

In order to compose *Analogique A*, Xenakis firstly restricts himself to impose a Markov structure over parameters F, G and D of the screens, each taking only two possible values. This will lead to a Markov structure on a set of screens as will become clear. He claims that more complex structures could lead to more interesting musical result, but the necessary volume of calculation is unfeasible to perform by hand, necessitating a computer to perform them.

Consider f_0 and f_1, g_0 and g_1, and d_0 and d_1 two distinct audible frequency, intensity and density regions[15], respectively. Each of these pair of variables is governed by one of the two transition matrices below:

$$\mathbf{P}_1 = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix} \qquad \mathbf{P}_2 = \begin{bmatrix} 0.85 & 0.15 \\ 0.4 & 0.6 \end{bmatrix}. \tag{33}$$

The Markov structure that Xenakis imposes on F, G, and D can be described in two steps: the current values of F, G and D determines if the next value of each one will be chosen accordingly to $\mathbf{P}_1$ or $\mathbf{P}_2$, and then, its respective values are chosen accordingly to the corresponding transition matrix. He calls this process as *coupling*, and it is completely described by the correspondence in Table 1, where the first line contains the current value of the variables, whereas the second contains the respective transition matrix for the variable in the third line.

[15] Here Xenakis measures density in *terts*, logarithm in base 3.

Table 1: *Coupling of transition matrices* $\mathbf{P}_1$ *and* $\mathbf{P}_2$.

f_0	f_1	d_0	d_1	g_0	g_1	g_0	g_1	f_0	f_1	d_0	d_1
$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_1$	$\mathbf{P}_2$
D	D	F	F	D	D	F	F	G	G	G	G

Table 2: *Transition matrix between the screens formed by the possible values of F, G and D. Letters from A to H are used to abbreviate the corresponding screens.*

	A $(f_0g_0d_0)$	B $(f_0g_0d_1)$	C $(f_0g_1d_0)$	D $(f_0g_1d_1)$	E $(f_1g_0d_0)$	F $(f_1g_0d_1)$	G $(f_1g_1d_0)$	H $(f_1g_1d_1)$
A	0.021	0.084	0.084	0.336	0.019	0.076	0.076	0.304
B	0.357	0.089	0.323	0.081	0.063	0.016	0.057	0.014
C	0.084	0.076	0.021	0.019	0.336	0.304	0.084	0.076
D	0.189	0.126	0.126	0.084	0.171	0.114	0.114	0.076
E	0.165	0.150	0.150	0.135	0.110	0.100	0.100	0.090
F	0.204	0.136	0.036	0.024	0.306	0.204	0.054	0.036
G	0.408	0.072	0.272	0.048	0.102	0.018	0.068	0.012
H	0.096	0.144	0.144	0.216	0.064	0.096	0.096	0.144

As an example, if the current state is $(f_0\ g_0\ d_1)$, the next value of F is conditioned to g_0 and d_1, and the table informs us that its next value should be chosen according to transition matrix $\mathbf{P}_2$. However, f_0 indicates that the next value of G should be chosen according to $\mathbf{P}_1$, but d_1 says that it should be chosen according to $\mathbf{P}_2$. Since in this case there is not agreement, one of the transition matrices are chosen at random with equal probability, and the next value of G will be chosen from this matrix.

Since each variable has two possible values, when combined we have 8 possibilities, that forms exactly 8 screens. From the methodology described above it is possible to compute the transition matrix between these screens, described in Table 2.

Therefore, this entire procedure is the *macroscopic method* that Xenakis mentioned before: there is a mechanism of transformation for frequency, intensity and density (transition matrices $\mathbf{P}_1$ and $\mathbf{P}_2$) and a interaction protocol between them (Table 1), that when combined implies in a mechanism of transformation between screens (transition matrix in Table 2).

Note that the transition matrix in Table 2 is ergodic, since each entry is strictly positive. Therefore, the respective Markov chain possess a stationary distribution, shown in Table 3. However, Xenakis notes that the chain converges quite quickly to the stationary distribution, and

Table 3: *Stationary distribution of transition matrix in Table 2.*

A $(f_0g_0d_0)$	B $(f_0g_0d_1)$	C $(f_0g_1d_0)$	D $(f_0g_1d_1)$	E $(f_1g_0d_0)$	F $(f_1g_0d_1)$	G $(f_1g_1d_0)$	H $(f_1g_1d_1)$
0.17	0.13	0.13	0.11	0.14	0.12	0.10	0.10

although not explicitly mentioned, he does not want that a initial screen freely evolve according to this transition matrix, maybe because of the "lack of innovation" of the stationary distribution. Therefore, an idea to overcome this difficulty can be loosely described as evolving a particular screen until it is close to the stationary distribution, applying a "perturbation", evolve the obtained screen again, and so on. Xenakis claims that this procedure does not diminish the importance of the Markov structure imposed by transition matrix on Table 2 but *confirms* its importance, since we are successively confirming and negating its structure, which he calls *mechanism Z* in the following quote:

> In effect the intrinsic value of the organism thus created lies in the fact that it must manifest itself, be. The perturbations which apparently change its structure represent so many negations of this existence. And if we create a succession of perturbations or negations, on the one hand, and stationary states or existences on the other, we are only *affirming* mechanism Z. In other words, at first we argue positively by proposing and offering as evidence the existence itself; and then we confirm it negatively by opposing it with perturbatory states. ([28, p. 94])

Therefore, the stochastic process underlying *Analogique A* can be described by the following *kinetic diagram*, as named by Xenakis:

$$E \to P_A^0 \to P_A' \to E \to P_C' \to P_C^0 \to P_B^0 \to P_B' \to E \to P_A', \tag{34}$$

where each of these symbols is a *protocol*, which means that screens are sampled according to a particular probability distribution:

- E is the stationary distribution on Table 3;
- P_A^0 is the distribution which assigns probability 1 to screen A and zero to the others (a *perturbation* towards screen A);
- P_A' is the result of applying the transition matrix in Table 2 to P_A^0;
- P_B^0, P_B', P_C^0, P_C' are similarly described.

Therefore, *Analogique A* is the result of some realization of this stochastic process, and the particular choices of f_0, f_1, g_0, g_1, d_0, and d_1 are shown in Figures III-8, III-9, and III-10 of [28, p. 98-99], respectively. The combinations between these features correspond to screens A from H in Figure III-13 of [28, p. 101], where the Roman numerals indicates the location of specific clouds of grains and Arabic numerals are the mean densities in grains per second. Finally, in order to properly transform this realization in a music, Xenakis sets the duration of each screen Δt as 1.11 s, the duration of one half note, and within this duration the densities of the occupied cells must be realized. Each of the protocol in the kinetic diagram in Equation 34 is explored with 30 screens, sampled from the corresponding probability distribution. This implies in 15 measures for each protocol, and since there are 10 protocols to be explored, *Analogique A* consists of 150 measures. Its underlying stochastic process can then be denoted as $S_1, \ldots, S_{300}$ where the individual probability distributions are described *via* the kinetic diagram in Equation 34, that is, $S_1, \ldots, S_{30}$ are independently sampled from distribution E, $S_{31} \ldots, S_{60}$ from P_A^0, and so on. Note that this process is not homogeneous, since the probability distribution being sampled changes at every 30 screens, despite its major inspiration, the Markov chain whose transition matrix is Table 2, being a homogeneous process.

It is important to note that since this composition is to be performed by string instruments, its execution does not correspond to the screens in Figure III-13 of [28, p. 101], because of the timbre of the particular instruments being played. Therefore, this stochastic structure corresponds only to the *fundamental frequencies*, and a particular execution possess much more complex screens.

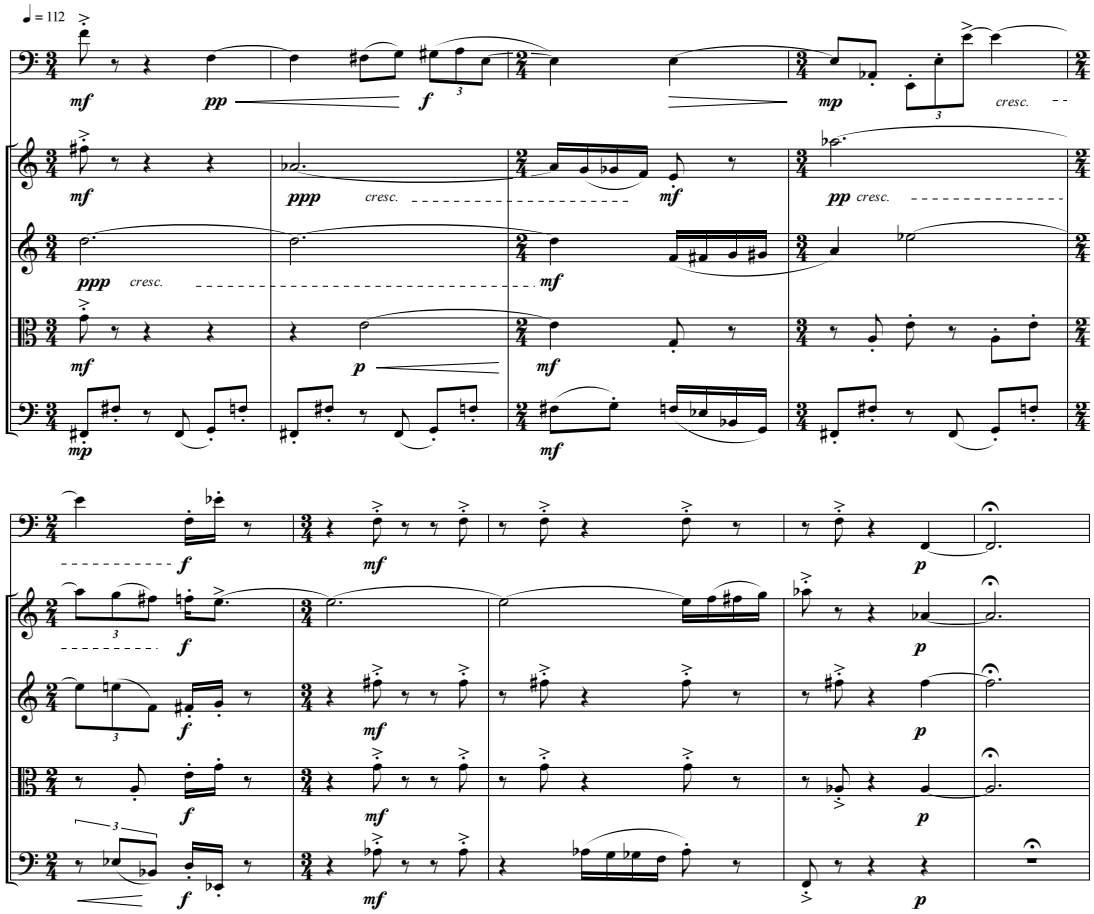

Figure 4: *Excerpt from Liduino Pitombeira's* Brazilian Landscapes No. 20 *for bassoon and string quartet.*

VI. Three more examples

The two musically-inspired examples previously developed, Examples 1 and 2, guided much of our intuition on mathematical concepts up to this point. However, one must agree that they are of low musical interest. In this section we present two applications of probabilistic tools in composition that, being less naive, leads to more interesting results.

i. Liduino Pitombeira's *Brazilian Landscapes No. 20* for bassoon and string quartet

The score on Figure 4 is an excerpt from a piece for bassoon and string quartet entitled *Brazilian Landscapes No.20* by Liduino Pitombeira. Its pitch classes comes from the Binomial distribution; the rhythm and specific pitch of the notes were not inspired by randomness, and we will explain only the first aforementioned aspect.

Firstly, consider $X \sim \text{Bin}(12, 1/2)$. With the intuition that X counts the number of heads obtained when tossing a fair coin 12 times, it becomes clear that its possible outcomes are $x = 0, \ldots, 12$. If the 12 pitch classes from C to B were put in a one-to-one correspondence with $\{0, \ldots, 11\}$ the possible outcomes of X except the last, we can loosely say that the notes in Figure

Table 4: *Probabilities associated with the notes in Liduino Pitombeira's* Brazilian Landscapes No. 20 *for bassoon and string quartet.*

x	$\mathbb{P}(X = x)$ (in %)	Pitch class	Quantity
0	0.02	C	0
1	0.29	C♯	0
2	1.61	D	2
3	5.37	D♯	5
4	12.08	E	12
5	19.34	F	19
6	22.56	F♯	23
7	19.34	G	19
8	12.08	G♯	12
9	5.37	A	5
10	1.61	A♯	2
11	0.29	B	0

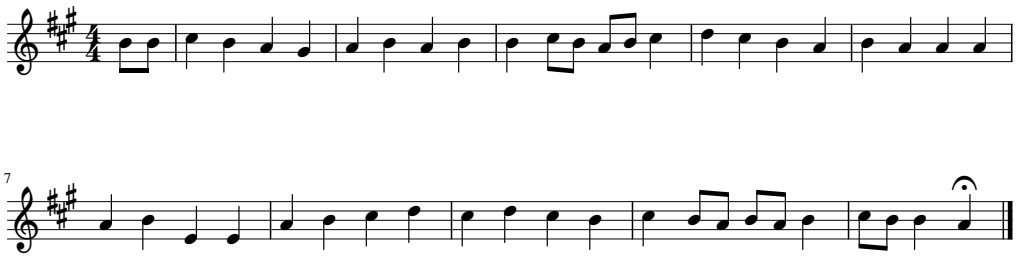

Figure 5: *A short composition based on the first order transitions from the chorales of J. S. Bach in the key of A major, with the rhythm of the chorale from BWV 104.*

4 were chosen accordingly to X', the random variable X truncated to only assume values in the set $\{0, \ldots, 11\}$. However, the procedure employed by Pitombeira was to list the probabilities of outcomes of X and then round its values to integer numbers, as in the second and fourth columns of Table 4, where the probability is measured in %. Therefore, the integer number in fourth column represents how many times the respective pitch class in third column will be present within the excerpt.

ii. Johann "Markov" Bach

On the other hand, the short pieces in Figures 5 and 6 were created with pure randomness, from a Markov structure of order 1 and 4, respectively, estimated from some of the chorales from Johann Sebastian Bach and using some tools from the Python package `music21` [8], a software tailored to perform computer-aided musicology. We now describe these excerpts in more details.

In order to estimate the transition probabilities, I considered only the soprano voice from Bach's chorales on the database of `music21` in the key of A major. This quite small database consisted of 1,637 notes, being only 13 of them distinct, whose MIDI numbers are 64, 66, 67, 68, 69,

Table 5: *Transition probabilities of order 1 estimated from the chorales of J. S. Bach in the key of A major. The states of the chain are the MIDI numbers of the notes, and probabilities are measured in %.*

	64	66	67	68	69	70	71	73	74	75	76	78	79
64	17.39	19.57	0	0	41.3	0	15.22	0	0	0	6.52	0	0
66	48.0	0	0	42.0	0	0	10	0	0	0	0	0	0
67	0	100	0	0	0	0	0	0	0	0	0	0	0
68	11.69	37.66	0	3.9	45.45	0	0	1.3	0	0	0	0	0
69	0.7	2.46	0	17.89	25.26	0	38.25	11.23	2.46	0	1.75	0	0
70	0	0	0	0	0	0	100	0	0	0	0	0	0
71	0.84	0.28	0.56	0.56	34.83	0.84	14.89	44.1	1.69	0	1.4	0	0
73	0	0.52	0	0	8.31	0	41.56	14.29	30.91	1.56	2.08	0.78	0
74	0	0	0	0	0	0	3.56	56.89	4.44	0	35.11	0	0
75	0	0	0	0	0	0	0	0	0	0	100	0	0
76	0	0	0	0	1.74	0	6.4	6.98	48.26	2.33	24.42	9.88	0
78	0	0	0	0	0	0	0	0	0	4.17	79.17	12.5	4.17
79	0	0	0	0	0	0	0	0	0	0	0	100	0

Table 6: *Stationary distribution of the transition matrix displayed in Table 5. Probabilities are measured in %.*

64	66	67	68	69	70	71	73	74	75	76	78	79
E4	F♯4	G4	G♯4	A4	A♯4	B4	C♯5	D5	D♯5	E5	F♯5	G5
2.81	3.05	0.12	4.7	17.41	0.18	21.75	23.52	13.74	0.67	10.51	1.47	0.06

70, 71, 73, 74, 75, 76, 78, and 79. The transition matrix of the chain of order 1 was estimated from this data by using the simple counting procedure described in Section IV and is displayed in Table 5, where the probabilities are measured in %. The first note on Figure 5 was chosen at random, uniformly from all the 13 possible notes, and 45 more notes were generated accordingly to the transition matrix in Table 5. The rhythm displayed in Figure 5 is from the chorale of BWV 104, one of the excerpts in the database analyzed. It is important to observe that the last note being an A was purely by chance, since I have not tried to generate the music several times until something "good" was obtained, but simply picked up the first sequence generated by the algorithm.

It can be shown that the transition matrix in Table 5 is ergodic, since its fourth power only has strictly positive entries. Therefore, it has a stationary distribution, displayed in Table 6 in %, together with the corresponding note names. Recall that this distribution can be interpreted as the proportion of appearance of each of these notes in a sufficiently large realization generated according to transition matrix in Table 5.

In order to illustrate the aforementioned drawback when considering higher order Markov chains, the short excerpt in Figure 6 was generated from a chain of order 4, whose transition probabilities were also inferred from the same *corpus*. Recall that now the estimation *via* the counting procedure is not quite reliable, since there are $13^5 = 371,293$ transition probabilities to estimate from the 1,637 observed notes, and several of these transitions does not even appear in the *corpus*. Indeed, from all of the $13^4 = 28,561$ distinct groups of four notes that can be formed with the 13 available notes, only 375 appears in the database, and for several of them there is

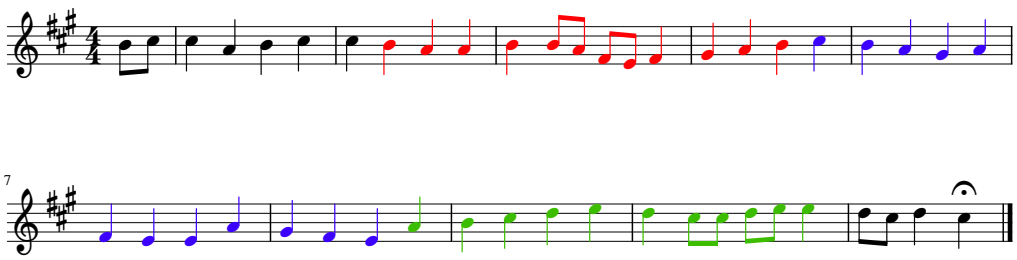

Figure 6: *A short composition based on the fourth order transitions from the chorales of J. S. Bach in the key of A major, with the rhythm of the chorale from BWV 104. The blocks of colored noted are already present in the database being analyzed.*

only one possibility of note to transition to, meaning that is very likely that some excerpt of some chorale is being exactly replicated. Indeed, the excerpts of 12 notes in red and blue in Figure 6 are already present within the *corpus* once, and the excerpt in green, consisting of 11 notes, appears eight times! This example illustrates that even with a low order chain, longer structures from the *corpus* can be replicated, without the creation of new musical material.

VII. Interpretability and flexibility

When using some statistical method to extract information from a dataset, one must have in mind the duality between intepretability and flexibility. There are some formal definitions of the flexibility of some model in the Statistics literature, but here let us understand the main idea, by comparing models in both extremes of this spectrum: Markov chains and neural networks.

Despite being outside the scope of this work, neural networks represent the state of the art in several Machine Learning techniques, in particular, algorithmic composition [10, 9]. However, it is extremely difficult to have an intuitive understanding of what the hidden layers are *exactly* doing, except in some special architectures. Usually, the only interpretable information in neural networks are the input and output layers, and if someone is interested in inferring information about the style of a composer, for example, is it very unlikely that this approach will lead him there, but it may have the capability of creating new musical pieces which passes in Turing tests with trained musicians [3].

On the other hand, we are convinced that Markov chains are highly interpretable, and some examples were presented illustrating that in order not to simply mimic the content already present in some *corpus* it must be wisely employed, in particular when dealing with its order and the definition of state, placing it then further to the flexible side of the spectrum and closer to the interpretable one.

A promising approach that claims to be in between the two extremes of this duality is [23]. Essentially, François Pachet and his collaborators propose that in order to properly capture some long-term information it is necessary only to model order two interactions but not only between adjacent notes, as in the Markov chain scenario. More specifically, they propose that a probability distribution that properly describes the information contained within a *corpus* is the *distribution of maximum entropy that honors the proportion of single notes and pair of notes up to some distance*.

This claim is one particular instance of a more general framework proposed in 1957 by the American physicist Edwin Jaynes. In this year he published two papers where some gaps between the notion of entropy in Statistical Mechanics and Probability were bridged, and also proposed

an interpretation Thermodynamics in probabilistic terms [14, 15]. Essentially, he claimed that the probability distribution which best represents our current state of knowledge in some scenario is the one with maximum entropy, conditioned to the constraints obtained from the observed data.

The entropy of a random variable X can be understood as the *average surprise* contained therein, and is defined as shown in Equation 35:

$$\mathcal{H}(X) = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i) \log\left(\frac{1}{\mathbb{P}(X = x_i)}\right) = -\sum_{i=1}^{\infty} \mathbb{P}(X = x_i) \log(\mathbb{P}(X = x_i)), \tag{35}$$

in the particular case of a discrete random variable X.

A deeper discussion of entropy of random variables and the applications of maximum entropy methods to algorithmic composition is outside the scope of this work, but it is important to remark the importance of this concept nowadays, not only in Science but directly in our lives. It was firstly introduced in 1948 by the American mathematician and electrical engineer Claude Shannon, in a landmark paper titled "A Mathematical Theory of Communication", and the concepts and theory introduced in this work are present everywhere, since it allows to efficiently compress and transmit information in a secure manner, and were crucial to the success of the space missions Voyager and Apollo, the invention of the compact disc and other medias, development of Internet, among several others. Moreover, the theory introduced by Shannon, nowadays known as *Information Theory*, is a field of knowledge with intersection with many others such as Statistics, Computer Science, Physics, Linguistics, Cryptography, and now, also Music!

VIII. Conclusion

In this work some basic aspects of Probability theory and Markov chains were introduced, mainly in a intuitive manner, in order to motivate researchers to employ these and more recent tools to perform music composition and analysis. Several examples were presented, with special attention to the brief analysis of Xenakis' *Analogique A* in Section V, and the excerpt of *Brazilian Landscapes No. 20* for bassoon and string quartet and both simulations from Markov chains in Section VI. My main goal with this work is to raise more questions and curiosity than answering them, in order to motivate the reader to go deeper in the literature on this subject. I hope that this goal was achieved.

Quoting again Isaac Newton, this text was only possible because I was standing on shoulders of giants. Without the references [16, 17, 28], some deep conversations with my doctorate advisor Luiz W. P. Biscainho about Music, Mathematics, life, the Universe and everything, and the meetings with Carlos Almada, Liduino Pitombeira, Pauxy Gentil-Nunes, Stefanella Boatto and Petrucio Viana to discuss Music and Mathematics, this work would not be possible. I am also deeply grateful to my student Nathalie Deziderio, for reviewing and making suggestions to improve this text.

References

[1] Almada, C. de L., *et. al.* (2019). Composição algorítmica de progressões harmônicas ao estilo de Antônio Carlos Jobim através de cadeias de Markov. In: *Proceedings of the 4th. International Congress of Music and Mathematics*, (in press). Rio de Janeiro, Brasil, 2019. Federal University of Rio de Janeiro.

[2] Ames, C. (1989). The Markov Process as a Compositional Model: A Survey and Tutorial. *Leonardo*, vol. 22, no. 2, pp. 175–187.

[3] BachBot. *The BachBot Challenge*. Available in https://bachbot.com/, accesed on 24/11/2019.

[4] Balakrishnan, V. (2009). *Lecture 1 - Introduction to Quantum Physics: Heisenberg's uncertainty principle*. Available in: https://www.youtube.com/watch?v=TcmGYe39XG0, accesed on 23/11/2019.

[5] Briot, J.-P., Hadjeres, G., and Pachet, F.-D. (2019) Deep Learning Techniques for Music Generation – A Survey. Available in: https://arxiv.org/abs/1709.01620.

[6] Broemeling, L. D. (2011). An Account of Early Statistical Inference in Arab Cryptology. *The American Statistician*, vol. 65, no. 4, pp. 255–257.

[7] Casella, G., and Berger, R. L. (2001). *Statistical Inference*. Boston: Cengage Learning.

[8] Cuthbert, M., and Ariza, C. (2010). music21: A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In: *Proceedings of the 11th. International Society for Music Information Retrieval Conference*, pp. 637–642. Utrecht, Netherlands, 2010. International Society for Music Information Retrieval.

[9] Eck, D., and Schmidhuber, J. (2002). A First Look at Music Composition using LSTM Recurrent Neural Networks. *Technical Report No. IDSIA-07-02 of the Instituto Dalle Molle di studi sull' intelligenza artificiale*.

[10] Hadjeres, G., Pachet, F., and Nielsen, F. (2017). DeepBach: a Steerable Model for Bach Chorales Generation. In: *Proceedings of the 34th. International Conference on Machine Learning*, pp. 1362–1371. Sydney, Australia, 2017. International Conference on Machine Learning.

[11] Gabor, D. (1947). Acoustical Quanta and the Theory of Hearing. *Nature*, vol. 159, pp. 591–594.

[12] Hiller, Jr., L. A., and Isaacson, L. M. (1957). Musical Composition with a High Speed Digital Computer. In: *Audio Engineering Society Convention 9*. New York, USA, 1957. Audio Engineering Society.

[13] Ibrahim, A. A.-K. (1992). The origins of cryptology: The Arab contributions. *Cryptologia*, vol. 16, no. 2, pp. 97–126.

[14] Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review: Series II*, vol. 106, no. 4, pp. 620–630.

[15] Jaynes, E. T. (1957). Information Theory and Statistical Mechanics II. *Physical Review: Series II*, vol. 108, no. 2, pp. 171–190.

[16] Maia, L. S. (2016). *Formalismos da Composição Algorítmica – Um Experimento com Canções Folclóricas Brasileiras* Dissertation (M.Sc. in Electrical Engineering). Graduate Program in Electrical Engineering, COPPE, Federal University of Rio de Janeiro, Rio de Janeiro.

[17] Mlodinow, L. (2009). *The Drunkard's Walk: How Randomness Rules Our Lives*. New York: Vintage.

[18] Olson, H. F., and Belar, H. (1961). Aid to Music Composition Employing a Random Probability System. *The Journal of the Acoustical Society of America*, vol. 33, no. 9, pp. 1163–1170.

[19] Pachet, F. (2002) The Continuator: Musical Interaction With Style. In: *Proceedings of the 2002 International Computer Music Conference*, pp. 211–218. Göteborg, Sweden, 2002. International Computer Music Assotiation.

[20] Ponsford, D., Wiggins, G. and Mellish, C. (1999) Statistical learning of harmonic movement. *Journal of New Music Research*, vol. 28, no. 2, pp. 150-177.

[21] Ross, S. (1994). *A First Course in Probability*. Upper Saddle River: Prentice Hall.

[22] Ross, S. (2019). *Introduction to Probability Models*. Cambridge: Academic Press.

[23] Sakellariou, J., Tria, F., Loreto, V., and Pachet, F. (2017). Maximum entropy models capture melodic styles. *Nature Scientific Reports*, v. 7, no. 9172.

[24] Seneta, E. (1996). Markov and the Birth of Chain Dependence Theory. *International Statistical Review*, v. 64, no. 3, pp. 255-263.

[25] Singh, S. (2000). *The Code Book: The Science of Secrecy From Ancient Egypt to Quantum Cryptography*. New York: Anchor Books.

[26] Stigler, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge: Belknap Press.

[27] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.

[28] Xenakis, I. (1992). *Formalized Music: Thought and Mathematics in Composition*. Hillsdale: Pendragon Press.

[29] Yanchenko, A. (2017). *Classical Music Composition Using Hidden Markov Models*. Dissertation (Ph.D in Statistics). Department of Statistical Science, Graduate School of Duke University, Duke University.

[30] Zwicker, E., and Fastl, H. (1999). *Psychoacoustics: Facts and Models*. New York: Springer.