# Estimation of Chaotic Probabilities

**Leandro Chaves Rêgo**
School of Electrical and Computer Engin.
Cornell University, Ithaca, NY, U.S.A.
lcr26@cornell.edu

**Terrence L. Fine**
School of Electrical and Computer Engin.
Cornell University, Ithaca, NY, U.S.A.
tlfine@ece.cornell.edu

## Abstract

A Chaotic Probability model is a usual set of probability measures, $\mathcal{M}$, the totality of which is endowed with an *objective, frequentist interpretation* as opposed to being viewed as a statistical compound hypothesis or an imprecise behavioral subjective one. In the prior work of Fierens and Fine, given finite time series data, the estimation of the Chaotic Probability model is based on the analysis of a set of relative frequencies of events taken along a set of subsequences selected by a set of rules. Fierens and Fine proved the existence of families of causal subsequence selection rules that can make $\mathcal{M}$ visible, but they did not provide a methodology for finding such family. This paper provides a universal methodology for finding a family of subsequences that can make $\mathcal{M}$ visible such that relative frequencies taken along such subsequences are provably close enough to a measure in $\mathcal{M}$ with high probability.

**Keywords.** Imprecise Probabilities, Foundations of Probability, Church Place Selection Rules, Probabilistic Reasoning, Complexity.

## 1 Introduction

A large portion of this and the following section is drawn in detail from Fierens 2003 [3] and Fierens and Fine 2003 [2]. They are included in this work to give the reader the necessary background to understand it.

### 1.1 Scope of Chaotic Probabilities

A Chaotic Probability model is a usual set of probability measures, $\mathcal{M}$, the totality of which is endowed with an *objective, frequentist interpretation* as opposed to being viewed as a statistical compound hypothesis or an imprecise behavioral subjective one. This model was proposed by Fierens and Fine 2003, [2] [3]. In this setting, $\mathcal{M}$ is intended to model stable (although not stationary in the traditional stochastic sense) physical sources of long finite time series data that have highly irregular behavior and not to model states of belief or knowledge that are assuredly imprecise. This work was in part inspired by the following quotation from Kolmogorov 1983 [6]:

*"In everyday language we call random those phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking, there is no ground to believe that random phenomena should possess any definite probability. Therefore, we should distinguish between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory)."*

The task of identifying real world data supporting the Chaotic Probability model is an important open research question. We believe the model may be useful for complex phenomena (e.g. weather forecast, financial data) where the probability of events, interpreted as propensity to occur, varies in a chaotic way with different initial conditions of the random experiment. Due to the complexity of such phenomena, inferring one probability for each possible initial condition is infeasible and unrealistic. The Chaotic Probability model gives a coarse-grained picture of the phenomena, keeping track of the range of the possible probabilities of the events.

### 1.2 Previous Work and Overview

There is some earlier literature that tries to develop a frequentist interpretation of imprecise probabilities (although most of the literature deals with a subjective interpretation, see for example Walley 1991 [11]). For work on asymptotics or laws of large numbers for interval-valued probability models, see Fine et al. [5] [7] [9] [10]. Cozman and Chrisman 1997 [1] studied the estimation of credal sets by analyzing limiting relative frequencies along a set of subsequences of a time series. However, as we said above, the focus of our Chaotic Probability model is the study of finite length time series.

In their work on Chaotic Probability models [2] [3], Fierens and Fine provided an instrumental interpretation of random processes measures consistent with $\mathcal{M}$ and the highly irregular physical phenomena they intended to model by $\mathcal{M}$. With that interpretation in mind they also provided a method for simulation of a data sequence given the Chaotic Probability model. This interpretation will be reviewed in Section 2.1.

Fierens and Fine also studied the estimation of $\mathcal{M}$ given a finite data sequence. They analyzed the sequence by looking at relative frequencies along subsequences selected by rules in some set $\Psi$. In particular, they introduced three properties of a set of subsequence selection rules: *Causal Faithfulness, Homogeneity and Visibility*. By Causally Faithful rules they meant rules that select subsequences such that the empirical and theoretical time averages along the selected subsequence are sufficiently close together. A set of rules renders $\mathcal{M}$ visible if all measures in $\mathcal{M}$ can be estimated by relative frequencies along the selected subsequences. Finally, a set of rules is homogeneous if it can not expose more than a small neighborhood of a single measure contained in the convex hull of $\mathcal{M}$, intuitively a set of rules is homogenous if the relative frequencies taken along the terms selected by the rules are all close to a single measure in the convex hull of $\mathcal{M}$.

Although, Fierens and Fine proved the existence of families of causal subsequence selection rules that can make $\mathcal{M}$ visible, they did not provide a methodology for finding such family. Section 3 describes a universal methodology for finding a family of causal subsequence selection rules that can make $\mathcal{M}$ visible. In Section 4, we strengthen this result by assuring that the relative frequency taken along every subsequence analyzed is close to some measure in $\mathcal{M}$ with high probability. This is a surprising result since Fierens and Fine believed the best one could hope to do was to estimate the convex hull of $\mathcal{M}$. We conclude in Section 5.

## 2 The Chaotic Probability Setup

### 2.1 Instrumental Interpretation

Let $\mathcal{X} = \{z_1, z_2, ..., z_\xi\}$ be a finite sample space. We denote by $\mathcal{X}^*$ the set of all finite sequences of elements taken from $\mathcal{X}$. A particular sequence of $n$ samples from $\mathcal{X}$ is denoted by $x^n = \{x_1, x_2, ..., x_n\}$. $\mathcal{P}$ denotes the set of all probability measures on the power set of $\mathcal{X}$. A chaotic probability model $\mathcal{M}$ is a subset of $\mathcal{P}$ and models the "marginals" of some process generating sequences in $\mathcal{X}^*$. This section provides an instrumental (that is, without commitment to reality) interpretation of such a process. [1]

Let $F$ be a chaotic[2] selection function, $F : \mathcal{X}^* \to \mathcal{M}$. At each instant $i$, a measure $\nu_i = F(x^{i-1})$ is chosen according to this selection function $F$. We require that the complexity of $F$ be neither too high, so that $\mathcal{M}$ can not be made visible on the basis of a finite time series, nor too low, so that as a working matter a standard stochastic process can be used to model the phenomena.

An actual data sequence $x^n$ is assessed by the "graded propensity" of the realization of a sequence of random variables $X^n$ described by:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_{k=1}^{n} \nu_k(x_k) \quad (1)$$

We denote by $\mathcal{M}^*$ the family of all such process measures $P$.

Although, it may be argued that $\mathcal{M}$ provides only an incomplete picture of the source, descriptions of $F$ being necessary to form a complete model, the chaotic probability is to be understood as providing a coarse-grained picture of the source with no empirical reality existing in the fine-grained description of the source given by $F$. An analogous situation occurs in quantum mechanics where there are fine-grained pictures that have no empirical reality, as we can see from the following quotes from Gell-Mann [4]:

*"Since quantum mechanics is probabilistic rather than deterministic, one might expect it to supply a probability for each fine-grained history. However, that is not the case. The interference terms between fine-grained histories do not usually vanish, and probabilities therefore cannot be assigned to such histories.... The answer is that in order to have actual probabilities, it is necessary to consider histories that are sufficiently coarse-grained." (p. 144)*

*"For histories of the universe in quantum mechanics, coarse-graining typically means following only certain things at certain times and only to a certain level of detail." (p.144)*

*"In practice, quantum mechanics is always applied to sets of decohering coarse-grained histories, and that is why it is able to predict probabilities." (p. 146)*

Note that no matter how complex the selection function is, the process measure $P$ is a standard stochastic

[1] As pointed out by an anonymous referee, the theory also applies if there is empirical reality in the description of $F$, but F is simply unknown.

[2] The word chaotic is not a technical term, it is ordinary language to describe such selection functions whose complexity is neither too small nor too large.

process. The issue is whether this process reflects the reality of the underlying phenomena, i.e., if the selection function is chaotic (in the sense of being neither too simple nor too complex) all we can hope to learn and therefore predict for future terms in the sequence is the coarse-grained description of the model given by $\mathcal{M}$.

The next subsection digresses on a new statistical model which can take advantage of the mathematical tools developed in this theory.

#### 2.1.1 Digression on a New Statistical Model

While our primary interest is not in a statistical compound hypothesis, the results of the Chaotic Probability model do bear on a new statistical estimation model. We can partially specify any stochastic process model $P \in \mathcal{P}$ by specifying the following set of conditional measures for the individual times:

$$\mathcal{M}^P = \{\nu : (\exists k)(\exists x^k),$$

$$\nu(X_{k+1} \in A) = P(X_{k+1} \in A | X^k = x^k), \ \forall A \subset \mathcal{X}\}$$

Note that we do not keep track of the conditioning event, only of the measure $\nu$. We then, wish to estimate $\mathcal{M}^P$ from data $x^n$. Also note that, in general, the process is not Markovian since the selection function $F$ depends on the whole history.

The model can be used in the following situation. Suppose we have an opponent in a game who can decide whether or not to act upon a trial $t$ after examining the history of outcomes $x^{t-1}$ prior to that trial. Certain distributions $P(X_t \in A | X^{t-1} = x^{t-1})$ for the trial at time $t$ are favorable to us and others to our opponent. An assessment of the range of consequences to us from choices made by an intelligent opponent can be calculated from $\mathcal{M}^P$.

### 2.2 Estimation

The estimation process in the chaotic probability framework uses a finite time series and analyzes it calculating a set of relative frequencies taken along subsequences selected by causal subsequence selection rules (also known as Church place selection rules). These rules are called causal because the next choice is a function of only past values in the sequence and not, say, of the whole sequence. These rules satisfy the following:

**Definition 1** *An effectively computable function $\varphi$ is a causal subsequence selection rule if:*

$$\varphi : \mathcal{X}^* \to \{0,1\}$$

*and, for any any $x^n \in \mathcal{X}^*$, $x_k$ is the $j$-th term in the generated subsequence $x^{\varphi,n}$, of length $\lambda_{\varphi,n}$, if:*

$$\varphi(x^{k-1}) = 1, \ \sum_{i=1}^{k} \varphi(x^{i-1}) = j, \ \lambda_{\varphi,n} = \sum_{k=1}^{n} \varphi(x^{k-1}).$$

Given a set of causal subsequence selection rules, $\Psi$, for each $\varphi \in \Psi$, define the empirical and theoretical time averages along a chosen subsequence as:

$$(\forall A \subset \mathcal{X}),$$

$$\overline{\mu}_{\varphi,n}(A) \doteq \frac{1}{\lambda_{\varphi,n}} \sum_{k=1}^{n} I_A(x_k)\varphi(x^{k-1})$$

$$\overline{\nu}_{\varphi,n}(A) \doteq \frac{1}{\lambda_{\varphi,n}} \sum_{k=1}^{n} E[I_A(X_k)|X^{k-1} = x^{k-1}]\varphi(x^{k-1})$$

where $I_A$ is the $\{0,1\}$-valued indicator function of the event $A$.

$\overline{\nu}_{\varphi,n}$ can be rewritten in terms of the instrumental understanding as:

$$(\forall A \subset \mathcal{X}) \ \overline{\nu}_{\varphi,n}(A) \doteq \frac{1}{\lambda_{\varphi,n}} \sum_{k=1}^{n} \nu_k(A)\varphi(x^{k-1})$$

As we want to make $\mathcal{M}$ visible by means of rules in $\Psi$, we define a metric, $d$, between measures to quantify how good an estimator $\overline{\mu}_{\varphi,n}$ is for $\overline{\nu}_{\varphi,n}$, as:

$$(\forall \mu, \mu' \in \mathcal{P}) \ d(\mu, \mu') \doteq \max_{z \in \mathcal{X}} |\mu(z) - \mu'(z)|$$

A rule $\varphi$ applied to $x^n$ is said to be *causally faithful* if $d(\overline{\nu}_{\varphi,n}, \overline{\mu}_{\varphi,n})$ is small. Fierens and Fine guaranteed the existence of such rules by proving the following theorem.

**Theorem 1 (Causal Faithfulness)** *Let $\xi$ be the cardinality of $\mathcal{X}$ and denote the cardinality of $\Psi$ by $||\Psi||$. Let $m \le n$. If $||\Psi|| \le t$, then for any process measure $P \in \mathcal{M}^*$:*

$$P(\max_{\varphi \in \Psi}\{d(\overline{\mu}_{\varphi,n}, \overline{\nu}_{\varphi,n}) : \lambda_{\varphi,n} \ge m\} \ge \epsilon) \le 2\xi t e^{\frac{-\epsilon^2 m^2}{2n}}$$

Note that, if we take $m = \alpha n$, for $\alpha \in (0,1)$, the size $t$ of the family of selection rules can be as large as $e^{\rho n}$, for $\rho < \frac{\alpha^2 \epsilon^2}{2}$; faithfulness of the rules is guaranteed with high probability for large $n$.

The property that a set of rules, $\Psi$, must satisfy in order to make $\mathcal{M}$ visible is given by the following definition:

**Definition 2 (Visibility)** *$\mathcal{M}$ is made visible $(\Psi, \theta, \delta, m, n)$ by $P \in \mathcal{M}^*$ if:*

$$P(\bigcap_{\nu\in\mathcal{M}} \bigcup_{\varphi\in\Psi} \{X^n : \lambda_{\varphi,n} \geq m, d(\nu, \overline{\mu}_{\varphi,n}) \leq \theta\}) \geq 1 - \delta$$

Let $B(\theta, \mu') \doteq \{\mu \in \mathcal{P} : d(\mu, \mu') < \theta\}$ and define an estimator, $\hat{\mathcal{M}}_{\theta,\Psi}$, of $\mathcal{M}$ by:

$$\forall x^n \in \mathbf{X}^*,\ \hat{\mathcal{M}}_{\theta,\Psi}(x^n) = \bigcup_{\varphi\in\Psi:\lambda_{\varphi,n}\geq m} B(\theta, \overline{\mu}_{\varphi,n})$$

Fierens and Fine proved both the existence of families of causal selection rules $\Psi$'s that can make $\mathcal{M}$ visible and that if $\mathcal{M}$ is made visible by $P$ through $\Psi$, then $\hat{\mathcal{M}}_{\theta,\Psi}$ is a good estimator of $\mathcal{M}$, in the sense that with high probability $P$, $\hat{\mathcal{M}}_{\theta,\Psi}$ contains $\mathcal{M}$ and is contained in a certain enlargement of the convex hull of $\mathcal{M}$.

A gap in the theory so far is identifying a procedure for finding a family of selection rules $\Psi$ that renders $\mathcal{M}$ visible. In the next section, we provide a methodology for finding such a family of rules $\Psi$ that works for any chaotic probability source, and we call it a *universal family of selection rules*. Unfortunately, such a family may "extract" more than $\mathcal{M}$. We return to this point in Section 4.

## 3 Universal Family of Selection Rules

In the previous section, we noted the importance of finding a family of rules, $\Psi$, that is capable of making our chaotic probability model $\mathcal{M}$ visible. This section provides a universal methodology for selecting such a family.

Define for each family of causal selection rules, $\Psi$, the estimator based on this family as:

$$\hat{\mathcal{M}}_{\Psi} \doteq \{\overline{\mu}_{\varphi,n} : \varphi \in \Psi, \lambda_{\varphi,n} \geq m\}$$

Let $ch(\mathcal{M})$ denote the convex hull of $\mathcal{M}$. And define the $\epsilon$-enlargement, $[\mathbf{A}]^\epsilon$, of a set $\mathbf{A}$, as:

$$[\mathbf{A}]^\epsilon \doteq \{\mu : \mu \in \mathcal{P} \text{ and}$$

$$\exists\mu' \in \mathbf{A} \text{ such that } \mathbf{d}(\mu, \mu') < \epsilon\}$$

Then, choose a small $\epsilon > 0$, and a minimal finite covering of $\mathcal{P}$ by $Q_\epsilon$ balls of radius $\epsilon$, $\{B(\epsilon, \mu_i)\}$, where $\mu_i$ are computable measures. Let $N_\epsilon$ be the size of the smallest subset of the above covering of the simplex that covers the actual chaotic probability model $\mathcal{M}$, and denote this subset by $\mathcal{M}_\epsilon$. Approximate $F(x^{k-1})$ by $F_\epsilon(x^{k-1}) = \mu_j$ if $\mu_j$ is the closest measure to $F(x^{k-1})$ among all $\mu_i$'s that belongs to $\mathcal{M}_\epsilon$. Let $F_{\epsilon,n}$ be the restriction of $F_\epsilon$ to $\mathcal{X}^{1:n}$ (all sequences of length not greater than $n$). The following theorem provides the desired method of finding a universal family of selection rules for chaotic probability sources.

Intuitively, Theorem 2 states that as long as the Kolmogorov Complexity [8] of the chaotic measure selection function is not too high, and we have a long enough data sequence, we are able to make visible with high probability all measures that were selected frequently enough in the sequence.

**Theorem 2** *Choose $f \geq 1$, $\alpha = (fN_\epsilon)^{-1}$ and let $m = \alpha n$. Define $\mathcal{M}_f \doteq \{\nu : \nu \in \mathcal{M}$ and $\exists\mu_i \in \mathcal{M}_\epsilon$ such that $d(\nu, \mu_i) < \epsilon$ and $\mu_i$ is selected at least $m$ times by $F_{\epsilon,n}$ $\}$. Given $\beta$ smaller than $\frac{\alpha^2\epsilon^2}{2}$, choose $\epsilon' \in (0, \beta\log_2 e)$ and assume the Kolmogorov complexity, $K(F_{\epsilon,n})$, of $F_{\epsilon,n}$ satisfies the following condition:*

$$\exists\kappa \geq 0,\ \exists L_{\epsilon',\kappa} \text{ such that } \forall n \geq L_{\epsilon',\kappa},$$
$$\frac{K(F_{\epsilon,n})}{n} < \beta\log_2 e + \frac{\kappa\log_2 n}{n} - \epsilon' \quad (2)$$

*Define $\mathcal{M}^*_R \doteq \{P : P \in \mathcal{M}^*$ and the corresponding $F$ satisfies condition 2$\}$. Then, for $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil\log_2 Q_\epsilon\rceil}{\epsilon'}\}$, there exists a family of causal subsequence selection rules $\Psi_U$, depending only on $\alpha$, $\kappa$ and $\epsilon$, such that $\forall\mathcal{M}$, $\forall P \in \mathcal{M}^*_R$:*

$$P(X^n : [ch(\mathcal{M})]^{4\epsilon} \supset [\hat{\mathcal{M}}_{\Psi_U}]^{3\epsilon} \supset \mathcal{M}_f) \geq 1 - \delta$$

*where $\gamma = \frac{\alpha^2\epsilon^2}{2} - \beta$ and $\delta = 2\xi n^\kappa e^{-\gamma n}$.*

**Remark 1** *Note that the fraction of times a measure in $\mathcal{M} \setminus \mathcal{M}_f$ is used to generate an outcome in a sequence $X^n$ is upper bounded by $(1/f)$. Therefore, for $f$ sufficiently large it is reasonable to expect that such measures may not be estimated.*

**Proof:** Define a family of selection functions, $\Psi_G$, that corresponds to $F_{\epsilon,n}$ as follows: $\Psi_G = \{\varphi_i^G$, for $1 \leq i \leq N_\epsilon\}$, where, for $0 \leq j \leq n-1$:

$$\varphi_i^G(x^j) \doteq \begin{cases} 1 & \text{if } F_{\epsilon,n}(x^j) = \mu_i \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

As each $\varphi_i^G$ is a function of $F_{\epsilon,n}$ and $\mu_i$, and $\lceil\log_2 Q_\epsilon\rceil$ is an upper bound on the number of bits necessary to specify the index $i$ of the particular measure $\mu_i$, the Kolmogorov complexity, $K(\varphi_i^G)$, of $\varphi_i^G$ satisfies:

$$\max_i K(\varphi_i^G) \leq K(F_{\epsilon,n}) + \lceil\log_2 Q_\epsilon\rceil \quad (4)$$

It then follows, from our hypothesis, that for $1 \leq i \leq N_\epsilon$ and $\forall n \geq L_{\epsilon',\kappa}$, $K(\varphi_i^G)$ satisfies the following condition:

$$\frac{K(\varphi_i^G)}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \epsilon' + \frac{\lceil \log_2 Q_\epsilon \rceil}{n}$$

Therefore, for $n > \max(L_{\epsilon',\kappa}, \frac{2\lceil \log_2 Q_\epsilon \rceil}{\epsilon'})$:

$$\frac{K(\varphi_i^G)}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \frac{\epsilon'}{2}$$

Let $\Psi_U$ consists of all rules of Kolmogorov complexity less than or equal to $\beta n \log_2 e + \kappa \log_2 n - 1$. Note that since for $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil \log_2 Q_\epsilon \rceil}{\epsilon'}\}$, $\frac{n\epsilon'}{2} > 1$, it follows that $\Psi_U$ includes $\Psi_G$ for $n$ large enough.

As $||\Psi_U|| \leq 2^{n\beta \log_2 e + \kappa \log_2 n} = n^\kappa e^{\beta n}$, $m = \alpha n$ and $\gamma = \frac{\alpha^2 \epsilon^2}{2} - \beta > 0$, by the causal faithfulness theorem, for any $P \in \mathcal{M}^*$:

$$P(\max_{\varphi \in \Psi_U} \{d(\overline{\mu}_{\varphi,n}, \overline{\nu}_{\varphi,n}) : \lambda_{\varphi,n} \geq m\} \geq \epsilon) \leq 2\xi n^\kappa e^{-\gamma n}$$

Note that, since for $\alpha = (f N_\epsilon^{-1})$, we know that for all $X^n$, there exists $i$ such that $\lambda_{\varphi_i^G,n} \geq m$, as $\Psi_G \subset \Psi_U$, we have that for all $X^n$ the maximum above is taken over a non-empty set.

To prove the theorem, let $\varphi_i^G$ be as defined in equation 3, then for a fixed $X^n$, by definition of $\mathcal{M}_f$, $\forall \nu \in \mathcal{M}_f$, $\exists \mu_i \in \mathcal{M}_\epsilon$ such that $d(\nu, \mu_i) < \epsilon$ and $\lambda_{\varphi_i^G,n} \geq m$ (Note the index $i$ depends on $X^n$). Then, using the triangle inequality property:

$$\sup_{\nu \in \mathcal{M}_f} d(\nu, \overline{\mu}_{\varphi_i^G,n}) \leq \sup_{\nu \in \mathcal{M}_f} d(\overline{\mu}_{\varphi_i^G,n}, \overline{\nu}_{\varphi_i^G,n}) + $$
$$+ \sup_{\nu \in \mathcal{M}_f} d(\overline{\nu}_{\varphi_i^G,n}, \mu_i) + \sup_{\nu \in \mathcal{M}_f} d(\mu_i, \nu).$$

Since $\overline{\nu}_{\varphi_i^G,n}$ is the time average of the actual measures selected by $F$ in the ball $B(\epsilon, \mu_i)$, $d(\overline{\nu}_{\varphi_i^G,n}, \mu_i) < \epsilon$, and as $\Psi_G \subset \Psi_U$, the following holds,

$$\{X^n : \max_{\varphi \in \Psi_U} \{d(\overline{\mu}_{\varphi,n}, \overline{\nu}_{\varphi,n}) : \lambda_{\varphi,n} \geq m\} < \epsilon\} \subset$$
$$\subset \{X^n : \sup_{\nu \in \mathcal{M}_f} \min_{\varphi \in \Psi_U} \{d(\nu, \overline{\mu}_{\varphi,n}) : \lambda_{\varphi,n} \geq m\} < 3\epsilon\}. \quad (5)$$

Equation 5 implies,

$$\{X^n : \max_{\varphi \in \Psi_U} \{d(\overline{\mu}_{\varphi,n}, \overline{\nu}_{\varphi,n}) : \lambda_{\varphi,n} \geq m\} < \epsilon\} \subset$$
$$\subset \{X^n : [ch(\mathcal{M})]^{4\epsilon} \supset [\hat{\mathcal{M}}_{\Psi_U}]^{3\epsilon} \supset \mathcal{M}_f\} \quad (6)$$

Theorem 2 follows from the causal faithfulness Theorem 1. □

The problem with the sort of estimator provided by the above theorem is that although it is able to make visible all measures in $\mathcal{M}$ that appeared frequently enough in the process, it is only guaranteed to be included in an enlarged neighborhood of $\mathcal{M}$'s convex hull and in some cases this can be rather larger than $\mathcal{M}$.

The following section proves a theorem that given $x^n$ provides a methodology for finding a universal family of subsequences, $\Psi(x^n)$, that is both able to make visible all measures in $\mathcal{M}$ that appears frequently enough in the process and contain only these subsequences whose empirical time averages are close enough to measures in $\mathcal{M}$ with high probability. We call this family *strictly faithful*.

# 4 Strictly Faithful Family of Subsequences

In this section, we propose a methodology for finding a set of strictly faithful family of subsequences that can both make visible all measures that are used frequently enough in any chaotic probability source $\mathcal{M}$ and contain only these subsequences whose empirical time averages are close enough to $\mathcal{M}$ with high probability.

The problem with the set of rules $\Psi_U$ is that it may contain rules that are not homogeneous, i.e., rules that select a subsequence generated by mixtures of measures $\mu_i$'s. In our proposed methodology in this section, we will analyze each rule $\varphi \in \Psi_U$ with a universal family $\Psi_U^\varphi$ (see definition below) and include $\varphi$ in $\Psi(x^n)$ only if it is homogeneous. As $\Psi_U^\varphi$ is universal for the subsequence selected by $\varphi$, it will be able to identify if it is or not homogeneous with high probability. Thus, our family of sequences $\Psi(x^n)$ is constructed in a two-stage process: first we consider the family of selection rules $\Psi_U$ which consists of all rules of at most a certain complexity value which is able to make $\mathcal{M}$ visible; then we filter the rules contained in $\Psi_U$ so that it contains only homogenous subsequences whose relative frequencies are close enough to a measure in $\mathcal{M}$.

The following theorem proves the desired result, i.e., if the Kolmogorov complexity of the chaotic measure selection function is not too high and we have a long enough data sequence, with high probability we can make visible all and only measures that were used frequently enough in the sequence.

**Theorem 3** *Choose* $f \geq 1$, $\alpha_1 = (f N_\epsilon)^{-1}$, $\alpha_2 = N_\epsilon^{-1}$ *and let* $m = \alpha_1 n$. *Define* $\mathcal{M}_f \doteq \{\nu : \nu \in \mathcal{M}$ *and* $\exists \mu_i \in \mathcal{M}_\epsilon$ *such that* $d(\nu, \mu_i) < \epsilon$ *and* $\mu_i$ *is selected at least* $m$ *times by* $F_{\epsilon,n}$ $\}$ *and* $\mathcal{M}_{\epsilon,f} \doteq \{\mu_i : \mu_i \in \mathcal{M}_\epsilon$ *and* $\mu_i$ *is selected at least* $\alpha_2 m$ *times by* $F_{\epsilon,n}$ $\}$. *Given* $\beta$ *smaller than* $\frac{\alpha_1^2 \alpha_2^2 \epsilon^2}{2}$, *choose* $\epsilon' \in (0, \beta \log_2 e)$ *and* *assume the Kolmogorov complexity, $K(F_{\epsilon,n})$, of $F_{\epsilon,n}$ satisfies the following condition:*

$$\exists\kappa \geq 0,\ \exists L_{\epsilon',\kappa} \text{ such that } \forall n \geq L_{\epsilon',\kappa}, \quad \frac{K(F_{\epsilon,n})}{n} < \beta \log_2 e + \frac{\kappa \log_2 n}{n} - \epsilon' \quad (7)$$

*Define $\mathcal{M}_R^* \doteq \{P : P \in \mathcal{M}^*$ and the corresponding $F$ satisfies condition 7$\}$. Then, for $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil\log_2 Q_\epsilon\rceil}{\epsilon'}\}$, for each $x^n$, there exists a family of subsequences $\Psi(x^n)$, depending only on $\alpha_1$, $\alpha_2$, $\kappa$ and $\epsilon$, such that $\forall\mathcal{M}$ and $\forall P \in \mathcal{M}_R^*$:*

$$P(\{X^n : \sup_{\mu\in\mathcal{M}_f} \min_{\nu\in\hat{\mathcal{M}}_{\Psi(X^n)}} d(\mu,\nu) < 3\epsilon\}\cap$$

$$\cap\{X^n : \max_{\nu\in\hat{\mathcal{M}}_{\Psi(X^n)}} \min_{\mu\in\mathcal{M}_{\epsilon,f}} d(\mu,\nu) < 6\epsilon\}) \geq 1 - \delta_1{}^{3}$$

*where $\gamma_1 = \frac{\alpha_1^2\alpha_2^2\epsilon^2}{2} - \beta$, $S_\epsilon \doteq \min\{Q_\epsilon, n^\kappa e^{\beta n}\}$ and $\delta_1 = 4\xi S_\epsilon n^\kappa e^{-\gamma_1 n}$.*

## 5 Conclusions and Future Work

We proposed a methodology for finding a universal family of subsequences that can strictly make visible all, and only these, measures in any chaotic probability source $\mathcal{M}$.

This is an important conceptual result that can provide alternative mathematical methods for analyzing time series data. We believe our capacity for recognizing new phenomena is conditioned by our existing mathematical constructs. Although, we do not have any real world example that supports our model, as development of the theory evolves, we expect to be able to identify phenomena with this new set of mathematical tools.

Although theoretically this is an important result, the procedure is not yet useful to apply in practical problems. First there is the issue of well-known non-computability of Kolmogorov complexity, but as pointed out by an anonymous referee, this could in principle be resolved since we only need an upper bound on its value to apply our result. The real problem is that $F_{\epsilon,n}$ is unknown. Our results guarantees that if the phenomena is not too complex, then we will learn it. But, we do not know usually how complex it actually is. Fierens [3] did some numerical examples of chaotic probability estimation from simulated sequences, but he focused on different family of rules that are not universal as the ones we propose.

On another front, much of the work done on chaotic probabilities focused on the univariate case. Although Fierens [3] also proposes a model for conditional chaotic probabilities, it is not clear that his approach to deal with the multivariate case is the most appropriate one. Finally, in order to build a complete theory of chance phenomena, the concepts of expectation and independence must be studied in this framework.

## Appendix: Proof of Theorem 3

**Proof:** Let $\Psi_G$ be as in the proof of theorem 2. Let $\Psi_U$ consists of all rules of Kolmogorov complexity not greater $\beta n \log_2 e + \kappa \log_2 n - 1$. From the hypothesis of the theorem 3 and from the proof of theorem 2, we know if $n > \max\{L_{\epsilon',\kappa}, \frac{2\lceil\log_2 Q_\epsilon\rceil}{\epsilon'}\}$, $\Psi_G \subset \Psi_U$.

Define $\phi_i \cdot \phi_j$ to be the selection rule that selects a term $x_k$ if and only if both $\phi_i$ and $\phi_j$ select it. For each $\varphi \in \Psi_U$, define:

$$\Psi_G^\varphi = \{\varphi_i^G \cdot \varphi, 1 \leq i \leq N_\epsilon\}$$

$$\Psi_U^\varphi = \{\phi \cdot \varphi : \phi \in \Psi_U\}$$

Again, from the hypothesis of the theorem 3 and from the proof of theorem 2, $\forall\varphi \in \Psi_U$, $\Psi_G^\varphi \subset \Psi_U^\varphi$.

The basic technical idea in the proof is to bound the distance we want to calculate using the triangle inequality property, so that we are left with distances that are of one of the following forms: distances between empirical and theoretical time averages taken along the same subsequence ($d(\overline{\mu},\overline{\nu})$), for which we can apply the causal faithfulness result; distances between a subsequence selected by $\varphi_i^G$, or any further subsequence of it, and the measure $\mu_i$, which, by definition of $\varphi_i^G$, is less than $\epsilon$; or distances that can be calculated from properties of $\mathcal{M}_f$, $\mathcal{M}_{\epsilon,f}$ or $\hat{\mathcal{M}}_{\Psi_{x^n}}$.

For each $x^n$, consider the following construction of $\Psi(x^n)$:

- $\Psi(x^n) = \{\}$
- For each $\varphi \in \Psi_U$ such that $\lambda_{\varphi,n}(x^n) \geq m$
  - $m_1^\varphi = \alpha_2 \lambda_{\varphi,n}$
  - If $\max_{\phi\in\Psi_U^\varphi}\{d(\overline{\mu}_{\phi,n},\overline{\mu}_{\varphi,n}) : \lambda_{\phi,n} \geq m_1^\varphi\} < 4\epsilon$
    - $\Psi(x^n) = \Psi(x^n) \cup \{\varphi\}$

[3] If $\Psi(X^n) = \emptyset$, we adopt the following convention: $\sup_{\mu\in\mathcal{M}_f} \min_{\nu\in\hat{\mathcal{M}}_\emptyset} d(\mu,\nu) = \infty$ and $\max_{\nu\in\hat{\mathcal{M}}_\emptyset} \min_{\mu\in\mathcal{M}_{\epsilon,f}} d(\mu,\nu) = 0$.

Note that, for $\alpha_1 = (fN_\epsilon^{-1})$, we know that for all $X^n$, there exists $i$ such that $\lambda_{\varphi_i^G,n} \geq m$, then we have that, for all $X^n$, $\mathcal{M}_f \neq \emptyset$ and $\mathcal{M}_{\epsilon,f} \neq \emptyset$. And for $\alpha_2 = N_\epsilon^{-1}$, we know that for all $X^n$ and for all $\varphi$, there exists $i$ such that $\lambda_{\varphi_i^G \cdot \varphi,n} \geq m_1^\varphi$, and also since $\Psi_G \subset \Psi_U$, $\forall i$, $\exists \phi \in \Psi_U^{\varphi_i^G}$ such that $\phi = \varphi_i^G$, so for $\phi = \varphi_i^G$, $\lambda_{\phi,n} \geq m_1^{\varphi_i^G} = \alpha_2 \lambda_{\varphi_i^G,n}$. This implies most of the maximums in the following proof are taken over a non-empty set.

But, by the above construction of $\Psi(x^n)$, it is possible that there exists some $x^n$ such that $\Psi(x^n)$ is empty in such case remember our convention:

$$\sup_{\mu \in \mathcal{M}_f} \min_{\nu \in \hat{\mathcal{M}}_\emptyset} d(\mu,\nu) = \infty \text{ and } \max_{\nu \in \hat{\mathcal{M}}_\emptyset} \min_{\mu \in \mathcal{M}_{\epsilon,f}} d(\mu,\nu) = 0$$

First, we are going to prove the following Lemma 1. In the first part of the proof, we want to verify that all measures in $\Psi_G$ that can make visible a part of $\mathcal{M}_f$ are included in $\Psi(x^n)$. The second part verifies that in all subsequences in $\Psi(x^n)$, the underlying measures that generated each term of the subsequence are contained in a small neighborhood of each other, i.e. the subsequence is homogeneous.

**Lemma 1**

$$\{\{X^n : \max_i \max_{\phi \in \Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) :$$

$$\lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < \epsilon\} \cap$$

$$\cap \{X^n : \max_i \max_{\varphi \in \Psi_U} \{d(\overline{\mu}_{\varphi_i^G \cdot \varphi,n}, \overline{\nu}_{\varphi_i^G \cdot \varphi,n}) :$$

$$\lambda_{\varphi,n} \geq m, \lambda_{\varphi_i^G \cdot \varphi,n} \geq m_1^\varphi\} < \epsilon\}\} \subset$$

$$\subset \{\{X^n : \sup_{\mu \in \mathcal{M}_f} \min_{\nu \in \hat{\mathcal{M}}_{\Psi(X^n)}} d(\mu,\nu) < 3\epsilon\} \cap$$

$$\cap \{X^n : \max_{\nu \in \hat{\mathcal{M}}_{\Psi(X^n)}} \min_{\mu \in \mathcal{M}_{\epsilon,f}} d(\mu,\nu) < 6\epsilon\}\}$$

**Proof of Lemma:** Let's first consider the event $\{X^n : \sup_{\mu \in \mathcal{M}_f} \min_{\nu \in \hat{\mathcal{M}}_{\Psi(X^n)}} d(\mu,\nu) < 3\epsilon\}$. From theorem 2, we know we will be able to make visible all measures in $\mathcal{M}_f$ if $\forall i$ such that $\lambda_{\varphi_i^G,n} \geq m$, the empirical and theoretical time averages on the subsequence selected by $\varphi_i^G$ are close enough; and if $\varphi_i^G$ is included in $\Psi(x^n)$.

Then note that $\forall \mathcal{A} \subset \mathcal{P}$ and $\forall \theta > 0$, if $\forall \mu \in \mathcal{A}$, $\exists i$ such that $\lambda_{\varphi_i^G,n} \geq m$ and $d(\mu, \overline{\mu}_{\varphi_i^G,n}) < \theta$; and if $\forall i$ such that $\lambda_{\varphi_i^G,n} \geq m$ implies $\varphi_i^G \in \Psi(X^n)$, then $\forall \mu \in \mathcal{A}, \exists \varphi \in \Psi(X^n)$ such that $d(\mu, \overline{\mu}_{\varphi,n}) < \theta$. This, by definition of $\Psi(X^n)$ and $\hat{\mathcal{M}}_{\Psi(X^n)}$, is equivalent to the following expression:

$$\{\{X^n : \sup_{\mu \in \mathcal{A}} \min_i \{d(\mu, \overline{\mu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < \theta\} \cap$$
$$\cap \{X^n : \max_i \max_{\phi \in \Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\mu}_{\varphi_i^G,n}) :$$
$$\lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < 4\epsilon\}\} \subset$$
$$\subset \{X^n : \sup_{\mu \in \mathcal{A}} \min_{\nu \in \hat{\mathcal{M}}_{\Psi(X^n)}} d(\mu,\nu) < \theta\} \quad (8)$$

In order to prove Lemma 1, let $\mathcal{A} = \mathcal{M}_f$ and $\theta = 3\epsilon$. We will now prove that the event $\{X^n : \sup_{\mu \in \mathcal{M}_f} \min_i \{d(\mu, \overline{\mu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < 3\epsilon\}$ occurs if $\forall i$ such that $\lambda_{\varphi_i^G,n} \geq m$, the empirical and theoretical time averages of the subsequence selected by $\varphi_i^G$ are close together. By the triangle inequality, we know:

$$d(\mu, \overline{\mu}_{\varphi_i^G,n}) \leq d(\overline{\nu}_{\varphi_i^G,n}, \overline{\mu}_{\varphi_i^G,n}) + d(\overline{\nu}_{\varphi_i^G,n}, \mu_i) + d(\mu_i, \mu)$$

By definition of $\varphi_i^G$, $d(\overline{\nu}_{\varphi_i^G,n}, \mu_i) < \epsilon$. For all $X^n$, by definition of $\mathcal{M}_f$, $\forall \mu \in \mathcal{M}_f$, $\exists \mu_i \in \mathcal{M}_\epsilon$ such that $d(\mu, \mu_i) < \epsilon$ and $\lambda_{\varphi_i^G,n} \geq m$, the previous triangle inequality implies that:

$$\{X^n : \max_i \{d(\overline{\mu}_{\varphi_i^G,n}, \overline{\nu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < \epsilon\} \subset$$
$$\subset \{X^n : \sup_{\mu \in \mathcal{M}_f} \min_i \{d(\mu, \overline{\mu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < 3\epsilon\}. \quad (9)$$

We will now prove that the condition for $\varphi_i^G$ be included in $\Psi(x^n)$ is implied by causal faithfulness of $\Psi_U^{\varphi_i^G}$, the universal family with which we analyze the subsequence selected by $\varphi_i^G$. In fact, expression 12 below will guarantee that if for all $i$ such that $\lambda_{\varphi_i^G,n} \geq m$, the family $\Psi_U^{\varphi_i^G}$ is faithful then $\mathcal{M}_f$ can be made visible by $\Psi(x^n)$. Again, use triangle inequality property to obtain:

$$d(\overline{\mu}_{\phi,n}, \overline{\mu}_{\varphi_i^G,n}) \leq d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) + d(\overline{\nu}_{\phi,n}, \mu_i) +$$

$$+ d(\mu_i, \overline{\nu}_{\varphi_i^G,n}) + d(\overline{\nu}_{\varphi_i^G,n}, \overline{\mu}_{\varphi_i^G,n}).$$

Since by definition of $\varphi_i^G$, $d(\mu_i, \overline{\nu}_{\varphi_i^G,n}) < \epsilon$, and by definition of $\Psi_U^{\varphi_i^G}$, for $\phi \in \Psi_U^{\varphi_i^G}$, $d(\overline{\nu}_{\phi,n}, \mu_i) < \epsilon$, if we assume the two remaining terms in the previous triangle inequality are less than $\epsilon$, we have:

$$\{\{X^n : \max_i \max_{\phi\in\Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) : \lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < \epsilon\} \cap \cap \{X^n : \max_i \{d(\overline{\mu}_{\varphi_i^G,n}, \overline{\nu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < \epsilon\}\} \subset \subset \{X^n : \max_i \max_{\phi\in\Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\mu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < 4\epsilon\}. \quad (10)$$

But the first event in the previous expression is included in the second because as $\Psi_G \subset \Psi_U$, $\forall i$, $\exists \phi \in \Psi_U^{\varphi_i^G}$ such that $\phi = \varphi_i^G$, so, as $\lambda_{\varphi_i^G,n} \geq m_1^{\varphi_i^G} = \alpha_2 \lambda_{\varphi_i^G,n}$:

$$\{X^n : \max_i \max_{\phi\in\Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) : \lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < \epsilon\} \subset \subset \{X^n : \max_i \{d(\overline{\mu}_{\varphi_i^G,n}, \overline{\nu}_{\varphi_i^G,n}) : \lambda_{\varphi_i^G,n} \geq m\} < \epsilon\}. \quad (11)$$

From expression 8(with $\mathcal{A} = \mathcal{M}_f$ and $\theta = 3\epsilon$), 9, 10 and 11, we get:

$$\{X^n : \max_i \max_{\phi\in\Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) : \lambda_{\varphi_i^G,n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < \epsilon\} \subset \subset \{X^n : \sup_{\mu\in\mathcal{M}_f} \min_{\nu\in\hat{\mathcal{M}}_{\Psi(X^n)}} d(\mu, \nu) < 3\epsilon\}. \quad (12)$$

Observe that:

$$\max_i \max_{\varphi\in\emptyset} \{d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}\} = 0$$

$$\text{and} \quad \max_i \max_{\varphi\in\emptyset} \{d(\overline{\mu}_{\varphi,n}, \mu_i) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}\} = 0.$$

Let's find a subset of the last event on Lemma 1. Consider $x^n$ such that $\Psi(x^n) \neq \emptyset$, otherwise note that the cases where $\Psi(x^n) = \emptyset$, by our convention, do not contradict the expressions 13 or 14, given below.

We will now prove that if all rules of the form $\varphi_i^G \cdot \varphi$, where $\varphi \in \Psi(X^n)$, are faithful, then $\forall \varphi \in \Psi(X^n)$, $\overline{\mu}_{\varphi,n}$ is close enough to a measure in $\mathcal{M}_{\epsilon,f}$. The reasoning behind the proof is as follows: as we know $\forall \varphi \in \Psi(X^n)$, $\exists i$ such that $\lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}$, then there are two possible cases, either only one measure $\mu_i$ satisfies this condition or more than one satisfies.

If only one $\mu_i$ satisfies $\lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}$, then by definition of $\Psi(x^n)$, $d(\overline{\mu}_{\varphi,n}, \overline{\mu}_{\varphi_i^G\cdot\varphi,n}) < 4\epsilon$. So, faithfulness of rules of the form $\varphi_i^G \cdot \varphi$ implies $d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) < \epsilon$ and by definition of $\varphi_i^G$, $d(\overline{\nu}_{\varphi_i^G\cdot\varphi,n}, \mu_i) < \epsilon$. And as $\lambda_{\varphi_i^G,n} \geq \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi} = \alpha_2\lambda_{\varphi,n} \geq \alpha_2 m$, $\mu_i \in \mathcal{M}_{\epsilon,f}$, so $\overline{\mu}_{\varphi,n}$ is close enough to $\mu_i \in \mathcal{M}_{\epsilon,f}$, as desired.

If $\exists \varphi \in \Psi(x^n)$ such that more than one measure $\mu_i$ satisfies $\lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}$ (call two of these indexes $i$ and $j$), then if all rules of the form $\varphi_i^G \cdot \varphi$ are faithful, triangle inequality and the definition of $\Psi(x^n)$ will imply $d(\mu_i, \mu_j) < 12\epsilon$ and this implies $\overline{\mu}_{\varphi,n}$ will be within $6\epsilon$ of a measure in $\mathcal{M}_{\epsilon,f}$. Let's see a formal proof. By triangle inequality, we have:

$$\max_i \max_{\varphi\in\Psi(X^n)} d(\overline{\mu}_{\varphi,n}, \mu_i) \leq \leq \max_i \max_{\varphi\in\Psi(X^n)} d(\overline{\mu}_{\varphi,n}, \overline{\mu}_{\varphi_i^G\cdot\varphi,n}) + + \max_i \max_{\varphi\in\Psi(X^n)} d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) + + \max_i \max_{\varphi\in\Psi(X^n)} d(\overline{\nu}_{\varphi_i^G\cdot\varphi,n}, \mu_i)$$

Then note that, by definition of $\hat{\mathcal{M}}_{\Psi(X^n)}$, $\forall \nu \in \hat{\mathcal{M}}_{\Psi(X^n)}$, $\exists \varphi \in \Psi(X^n)$ such that $\nu = \overline{\mu}_{\varphi,n}(\lambda_{\varphi,n} \geq m$ since $\varphi \in \Psi(X^n))$.

Then notice that by definition of $\varphi_i^G$, $d(\overline{\nu}_{\varphi_i^G\cdot\varphi,n}, \mu_i) < \epsilon$ and by definition of $\Psi(X^n)$, $\forall \varphi \in \Psi(X^n)$, if $\lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}$ then $d(\overline{\mu}_{\varphi,n}, \overline{\mu}_{\varphi_i^G\cdot\varphi,n}) < 4\epsilon$ and as we know that $\forall \varphi \in \Psi(X^n)$, $\forall x^n$, $\exists i(\varphi)$ such that $\lambda_{\varphi_{i(\varphi)}^G\cdot\varphi,n} \geq m_1^{\varphi} = \alpha_2\lambda_{\varphi,n} \geq \alpha_2 m$, then by definition of $\mathcal{M}_{\epsilon,f}$, $\mu_{i(\varphi)} \in \mathcal{M}_{\epsilon,f}$. So, the previous triangle inequality implies that:

$$\{X^n : \max_i \max_{\varphi\in\Psi(X^n)} \{d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}\} < \epsilon\} \subset \subset \{X^n : \max_i \max_{\varphi\in\Psi(X^n)} \{d(\overline{\mu}_{\varphi,n}, \mu_i) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}\} < 6\epsilon\} \subset \subset \{X^n : \max_{\nu\in\hat{\mathcal{M}}_{\Psi(X^n)}} \min_{\mu\in\mathcal{M}_{\epsilon,f}} \{d(\nu, \mu)\} < 6\epsilon\} \quad (13)$$

Finally note that, by definition of $\Psi(X^n)$, since $\Psi(X^n) \subset \Psi_U$ and $\forall \varphi \in \Psi(X^n)$, $\lambda_{\varphi,n} \geq m$:

$$\{X^n : \max_i \max_{\varphi\in\Psi_U} \{d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}, \lambda_{\varphi,n} \geq m\} < \epsilon\} \subset \subset \{X^n : \max_i \max_{\varphi\in\Psi(X^n)} \{d(\overline{\mu}_{\varphi_i^G\cdot\varphi,n}, \overline{\nu}_{\varphi_i^G\cdot\varphi,n}) : \lambda_{\varphi_i^G\cdot\varphi,n} \geq m_1^{\varphi}\} < \epsilon\} \quad (14)$$

Expressions 12, 13 and 14 prove the lemma. □

Applying the causal faithfulness theorem, since $||\Psi_G|| = N_\epsilon \leq S_\epsilon$ and $||\Psi_U^{\varphi_i^G}|| = ||\Psi_U|| \leq n^\kappa e^{\beta n}$, we have:

$$P(\{X^n : \max_i \max_{\varphi \in \Psi_U} \{d(\overline{\mu}_{\varphi_i^G \cdot \varphi, n}, \overline{\nu}_{\varphi_i^G \cdot \varphi, n}) : \lambda_{\varphi,n} \geq m, \lambda_{\varphi_i^G \cdot \varphi, n} \geq m_1^\varphi\} < \epsilon\}) \geq \geq 1 - 2\xi n^\kappa S_\epsilon e^{\beta n} \exp(\frac{-\epsilon^2 \alpha_1^2 \alpha_2^2 n}{2}) = 1 - 2\xi n^\kappa S_\epsilon e^{-\gamma_1 n} \quad (15)$$

$$P(\{X^n : \max_i \max_{\phi \in \Psi_U^{\varphi_i^G}} \{d(\overline{\mu}_{\phi,n}, \overline{\nu}_{\phi,n}) : \lambda_{\varphi_i^G, n} \geq m, \lambda_{\phi,n} \geq m_1^{\varphi_i^G}\} < \epsilon\}) \geq \geq 1 - 2\xi n^\kappa S_\epsilon e^{\beta n} \exp(\frac{-\epsilon^2 \alpha_1^2 \alpha_2^2 n}{2}) = 1 - 2\xi n^\kappa S_\epsilon e^{-\gamma_1 n} \quad (16)$$

Equations 15, 16 and Lemma 1 prove the theorem. □


### Acknowledgements

The first author received a scholarship from the Brazilian Government through the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) while this research was being done. We also would like to thank Pablo Fierens for the first two sections and for useful talks we had in the early stages of this work.


## References


[1] Fabio Cozman and L. Chrisman. Learning convex sets of probability from data. Technical Report CMU-RI-TR-97-25, Robotics Institute, Carnegie Mellon University, 1997.

[2] Pablo I. Fierens and Terrence L. Fine. Toward a Chaotic Probability Model for Frequentist Probability: The Univariate Case. *ISIPTA'03 Proceedings*, pages 245–259, July 2003.

[3] Pablo Ignácio Fierens. *Towards a Chaotic Probability Model for Frequentist Probability.* PhD thesis, Cornell University, 2003.

[4] M. Gell-Mann. *The Quark and The Jaguar.* W. H. Freeman and Company, 1994.

[5] Y.-L. Grize and Terrence L. Fine. Continuous lower probability-based models for stationary processes with bounded and divergent time averages. *Annals of Probability*, 15:783–803, 1987.

[6] A. N. Kolmogorov. *On logical foundations of probability theory*, volume 1021 of *Lecture Notes in Mathematics.* Springer-Verlag, 1983.

[7] A. Kumar and Terrence L. Fine. Stationary lower probabilities and unstable averages. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 69:1–17, 1985.

[8] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications.* Graduate Texts in Computer Science. Springer, second edition, 1997.

[9] A. Papamarcou and Terrence L. Fine. A note on undominated lower probabilities. *Annals of Probability*, 14:710–723, 1986.

[10] A. Sadrolhefazi and Terrence L. Fine. Finite-dimensional distribution and tail behavior in stationary interval-valued probability models. *Annals of Statistics*, 22:1840–1870, 1994.

[11] Peter Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall Pubs., 1991.