Towards a Strictly Frequentist Theory of Imprecise Probability

Christian Fröhlich CHRISTIAN.FROEHLICH@UNI-TUEBINGEN.DE
Rabanus Derr RABANUS.DERR@UNI-TUEBINGEN.DE
Department of Computer Science, University of Tübingen, Germany

Robert C. Williamson BOB.WILLIAMSON@UNI-TUEBINGEN.DE
Department of Computer Science, University of Tübingen, Germany
and Tübingen AI Center, Germany

Abstract

Strict frequentism defines probability as the limiting relative frequency in an infinite sequence. What if the limit does not exist? We present a broader theory, which is applicable also to statistical phenomena that exhibit diverging relative frequencies. In doing so, we develop a close connection with imprecise probability: the cluster points of relative frequencies yield a coherent upper prevision. We show that a natural frequentist definition of conditional probability recovers the generalized Bayes rule. We prove constructively that, for a finite set of elementary events, there exists a sequence for which the cluster points of relative frequencies coincide with a prespecified set, thereby providing strictly frequentist semantics for coherent upper previsions.

Keywords: imprecise probability, strict frequentism, divergent relative frequencies, von Mises, Ivanenko

1. Introduction

It is now almost universally acknowledged that probability theory ought to be based on Kolmogorov's [31] mathematical axiomatization (translated in [32]).[1] However, if probability is defined in this purely measure-theoretic fashion, what warrants its application to real-world problems of decision making under uncertainty? To those in the so-called *frequentist* camp, the justification is essentially due to the *law of large numbers*, which comes in both an empirical and a theoretical flavour. In this paper we question both of these presumptions.

By the empirical version of the law of large numbers (LLN), we mean not a "law" which can be proven to hold, but the following hypothesis, which seems to guide many scientific endeavours. Assume we have obtained data $x_1, .., x_n$ as the outcomes of some experiment, which has been performed n times under "statistically identical" conditions. Of course, conditions in the real-world can never truly be identical — otherwise the outcomes would be constant, at least under the assumption of a deterministic universe. Thus, "identical" in this context must be a weaker notion, that all factors which we have judged as relevant to the problem at hand have been kept constant over the repetitions.[2] The empirical "law" of large numbers, which Gorban [18] calls the *hypothesis of (perfect) statistical stability* then asserts that in the long-run, relative frequencies of events and sample averages converge. These limits are then conceived of as the *probability* of an event and the *expectation*, respectively. Thus, even if relative frequencies can fluctuate in the finite data setting, we presume that they stabilize as more and more data is acquired. Crucially, this hypothesis of perfect statistical stability is not amenable to falsification, since we can never refute it in the finite data setting. It is a matter of faith to assume convergence of relative frequencies. On the other hand, there is now ample experimental evidence that relative frequencies can fail to stabilize even under very long observation intervals [18, Part II]. We say that such phenomena display *unstable (diverging)* relative frequencies. Rather than refuting the stability hypothesis, which is impossible, we question its adequateness as an idealized modeling assumption: we view convergence as the idealization of approximate stability in the finite case, whereas divergence idealizes instability. Thus, if probability is understood as limiting relative frequency, then the applicability of Kolmogorov's theory to empirical phenomena is limited to those which are statistically stable; the founder himself remarked:

> Generally speaking there is no ground to believe that a random phenomenon should possess any definite probability [33].

Building on the works of von Mises and Geiringer [46], Walley and Fine [48] and Ivanenko [22], our goal is to establish a broader theory that is also applicable to statistical phenomena which are outside of the scope of Kolmogorov's theory by exhibiting unstable relative frequencies.

[1] An important exception is quantum probability [20, 30].

[2] In fact, we do not need that conditions stay exactly constant, but that they change merely in a way that is so benign that the relative frequencies converge. That is, in the limit we should obtain a stable statistical aggregate.

One attempt to "prove" (or justify) the empirical law of large numbers, which in our view is doomed to fail, is to invoke the theoretical law of large numbers, which is a purely formal, mathematical statement. The strong law of large numbers states that if $X_1, X_2, ..$ is a sequence of independent and identically distributed (i.i.d.) random variables with finite expectation $\mathbb{E}[X] := \mathbb{E}[X_1] = \mathbb{E}[X_2] = \cdots$, then the sample average $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to the expectation:

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mathbb{E}[X]\right) = 1,$$

where P is the underlying probability measure in the sense of Kolmogorov. To interpret this statement correctly, some care is needed. It asserts that P assigns measure 1 to the set of sequences for which the sample mean converges, but not that this happens *for all* sequences. Thus one would need justification for identifying "set of measure 0" with "is negligible" ("certainly does not happen"), which in particular requires a justification for P. With respect to a different measure, this set might not be negligible at all [41, p. 8]; see also [3, 42] for critical arguments. Moreover, the examples in [18, Part II] show that sequences with seemingly non-converging relative frequencies (fluctuating substantially even for long observation intervals) are not "rare" in practice.

Given these complications, we opt for a different approach, namely a *strictly frequentist* one. Reaching back to Richard von Mises' [45] foundational work, a strictly frequentist theory explicitly defines probability in terms of limiting relative frequencies in a sequence. Importantly, we here *do not assume* that the elements of the sequence are random variables with respect to an abstract, countably additive probability measure. Instead, like von Mises, we take the notion of a sequence as the primitive entity in the theory. As a consequence, countable additivity does not naturally arise in this setting, and hence we do not subscribe to the frequentist interpretation of the classical strong LLN.

The core motivation for our work is to drop the assumption of perfect statistical stability and instead to explicitly model the possibility of unstable (diverging) relative frequencies. Rather than merely conceding that the "probability" might vary over time [2, pp. 27ff.] (which begs the question what such "probabilities" mean) we follow the approach of Ivanenko [22], reformulate his construction of a *statistical regularity* of a sequence, and discover that it is closely connected to *imprecise probability*, more specifically, to the subjectivist theory of *lower and upper previsions* [47]. In essence, to each infinite sequence we can naturally associate a *set* of probability measures, which constitute the statistical regularity that describes the cluster points of relative frequencies and consequently also those of sample averages. Since this works for *any* sequence and *any* event, we have thus countered a typical argument against frequentism, namely that the limit may not exist and hence probability is undefined [21]. On an arbitrary (possibly infinite) set of outcomes, the limiting relative frequencies induce a coherent upper probability and the limiting sample averages induce a coherent upper prevision in the sense of Walley [47]. In the convergent case, this reduces to a precise, finitely additive probability and a linear prevision, respectively. We demonstrate that (for a finite set of outcomes) the converse direction works, too: any coherent upper prevision can be induced in a strictly frequentist way from a sequence, which we can explicitly construct. Furthermore, we derive in a natural way a conditional upper prevision; remarkably, this approach recovers the *generalized Bayes rule*, the arguably most important updating principle in imprecise probability.

This paper is accompanied by an extended preprint [17], which contains all of the proofs that were not included in this paper, as well as additional discussions.

1.1. Von Mises - The Frequentist Perspective

Our approach is inspired by, and generalizes, Richard von Mises [45] (refined and summarized in [46]) axiomatization of probability theory. In contrast to the subjectivist camp, von Mises' concern was to develop a theory for repetitive events, which gives rise to a theory of probability that is mathematical, but which can also be used to reason about the physical world. Hence, von Mises is not concerned with the probability of single events, which he deems meaningless, but instead always views an event as part of a larger *reference class*. Such a reference class is captured by what he terms a *collective*, a disorderly sequence which exhibits both global regularity and local irregularity. For the definition of a collective, we need a possibility set Ω of elementary outcomes $\omega \in \Omega$, together with a set system of events $\mathcal{A} \subseteq 2^{\Omega}$.

Definition 1 *Consider a tuple* $\left(\Omega, \vec{\Omega}, \mathcal{A}, \mathcal{S}\right)$ *with the following data:*

1. *a sequence* $\vec{\Omega}\colon \mathbb{N} \to \Omega$*;*
2. *a set of* selection rules $\mathcal{S} := \{\vec{S}_j : j \in \mathcal{J}\}$*, where for each* j *in a countable index set* $\mathcal{J}$*,* $\vec{S}_j\colon \mathbb{N} \to \{0,1\}$ *and* $\vec{S}_j(i) = 1$ *for infinitely many* $i \in \mathbb{N}$*;*
3. *a non-empty set system* $\mathcal{A} \subseteq 2^{\Omega}$*, where for simplicity we assume* $|\mathcal{A}| < \infty$*.* [3]

This tuple forms a collective if the following two axioms hold.

[3] In fact, $\mathcal{A}$ does not necessarily have to be finite. Since an infinite domain of probabilities does not contribute a lot to a better understanding of the frequentist definition at this point, we restrict ourselves to the finite case here. The reader can find details in [46].

vM1. *The limiting relative frequency for $A \in \mathcal{A} \subseteq 2^{\Omega}$ exists:*

$$P(A) := \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right).^{4}$$

We call this limit the probability of A.

vM2. *For each $j \in \mathcal{J}$, the selection rule $\vec{S}_j$ does not change limiting relative frequencies:*[5]

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right) \cdot \vec{S}_j(i)}{\sum_{i=1}^{n} \vec{S}_j(i)} = P(A) \quad \forall A \in \mathcal{A}.$$

Axiom vM1 explicitly defines the probability of an event in terms of the limit of its relative frequency. Demanding that this limit exists is non-trivial, since this need not be the case for an arbitrary sequence. Intuitively, vM1 expresses the hypothesis of statistical stability, which captures a global regularity of the sequence.

In contrast, vM2 captures a sense of *randomness* or local irregularity. It is best understood by viewing a selection rule $\vec{S}_j$ as selecting a subsequence of the original sequence $\vec{\Omega}$ and then demanding that the limiting relative frequencies thereof coincide with those of the original sequence. Why do we need axiom vM2? Von Mises calls this the "law of the excluded gambling system" and it is the key to capture the notion of randomness in his framework. Intuitively, if a selection rule were to change relative frequencies, an adversary could exploit this selection rule to strategically offer a bet on the next outcome and thereby make long-run profit, at the expense of a fictional decision maker. A *random* sequence, however, is one for which there does not exist such a betting strategy. It turns out that this statement cannot hold in its totality. A sequence cannot be random with respect to all selection rules except in trivial cases (cf. Kamke's critique of von Mises' notion of randomness, nicely summarized in [44]). Thus, von Mises explicitly relativizes randomness with respect to a problem-specific set of selection rules [46, p. 12].[6]

In our view, the role of the randomness axiom vM2 is similar to the role of more familiar randomness assumptions like the standard *i.i.d.* assumption: to empower inference from finite data. In this work, however, we will be exclusively concerned with the idealized case of infinite data, disregarding vM2, since our focus is the axiom (or hypothesis) of statistical stability (cf. the discussion in Section 5).

We are motivated by the following question. What is the suitable generalization of von Mises approach when axiom vM1 breaks down? That is, when relative frequencies of at least some events do not converge. Our answer leads to a confluence with a theory that is thoroughly grounded in the subjectivist camp: the theory of *lower and upper previsions*.

1.2. Imprecise Probability - The Subjectivist Perspective

We briefly introduce the *prima facie* unrelated, subjectivist theory of imprecise probability, or more specifically, the theory of *lower and upper previsions* as put forward by Walley [47]. Orthodox Bayesianism models belief via the assignment of precise probabilities to propositions, or equivalently, via a linear expectation functional. In contrast, in Walley's theory, belief is interval-valued and the linear expectation is replaced by a pair of lower and upper expectations. Hence, the theory is strictly more expressive than orthodox Bayesianism, which can be recovered as a special case.

We assume an underlying possibility set Ω, where $\omega \in \Omega$ is an elementary event, which includes all relevant information. We call a function $X\colon \Omega \to \mathbb{R}$, which is bounded, i.e. $\sup_{\omega\in\Omega} |X(\omega)| < \infty$, a *gamble* and collect all such functions in the set L^{∞}. The set of gambles L^{∞} carries a vector space structure with scalar multiplication $(\lambda X)(\omega) = \lambda X(\omega)$, $\lambda \in \mathbb{R}$, and addition $(X+Y)(\omega) = X(\omega) + Y(\omega)$. For a constant gamble $c(\omega) = c\ \forall\omega$ we write simply c. Note that Walley's theory in the general case does not require that a vector space of gambles is given, but definitions and results simplify significantly in this case.

We interpret a gamble as assigning an uncertain loss $X(\omega)$ to each elementary event, that is, in line with the convention in insurance and machine learning, we take positive values to represent loss and negative values to represent reward, with zero being neutral.[7] We imagine a decision maker who is faced with the question of how to value a gamble X; the orthodox answer would be the expectation $\mathbb{E}[X]$ with respect to a subjective probability measure.

Walley [47] proposed a betting interpretation of imprecise probability, which is inspired by de Finetti [8], who identifies probability with fair betting rates. The goal is to axiomatize a functional $\overline{R}\colon L^{\infty} \to \mathbb{R}$, which assigns to a gamble the infimum number $\overline{R}(X)$ so that $X - \overline{R}(X)$ is a desirable transaction to our decision maker, where she incurs the uncertain loss X but in exchange gets the reward $-\overline{R}(X)$.

[4] The function χ_A denotes the indicator gamble for a set $A \subseteq \Omega$, i.e. $\chi_A(\omega) := 1$ if $\omega \in A$ and $\chi_A(\omega) := 0$ otherwise.

[5] To be precise, a selection rule in the sense of von Mises is a map from the set of finite Ω-valued strings to $\{0, 1\}$, i.e. a selection rule is able to "see" all previous elements when deciding whether or not to select the next one. Our formulation is more restrictive to avoid notational overhead, but when a sequence is fixed, the two formulations are equivalent.

[6] This class of selection rules necessarily must be specified in advance; confer [43]. A prominent line of work aspires to fix the set of selection rules as all partially computable selection rules [5], but there is no compelling reason to elevate this to a universal choice; cf. [9] for an elaborated critique.

[7] Unfortunately, this introduces tedious sign flips when comparing results to Walley [47].

Formally:

$$\overline{R}(X) := \inf\{\alpha \in \mathbb{R} \colon X - \alpha \in \mathcal{D}\},$$

where $\mathcal{D}$ is a *set of desirable gambles*. Walley [47, Section 2.5] argued for a criterion of coherence, which any reasonable functional $\overline{R}$ should satisfy, and consequently obtained the following characterization [47, Theorem 2.5.5], which we shall take here as an axiomatic *definition* instead.[8]

Definition 2 *A functional* $\overline{R} \colon L^\infty \to \mathbb{R}$ *is a* coherent upper prevision *if* $\forall X, Y \in L^\infty$*:*

UP1. $\overline{R}(X) \leq \sup(X)$
UP2. $\overline{R}(\lambda X) = \lambda \overline{R}(X)$, $\forall \lambda \in \mathbb{R}^+$
UP3. $\overline{R}(X + Y) \leq \overline{R}(X) + \overline{R}(Y)$

These properties also imply $\forall X, Y \in L^\infty$ [47, p. 76]:

UP4. $\overline{R}(X + c) = \overline{R}(X) + c$, $\forall c \in \mathbb{R}$
UP5. $X(\omega) \leq Y(\omega)\ \forall \omega \in \Omega \Rightarrow \overline{R}(X) \leq \overline{R}(Y)$

To a coherent upper prevision, we can define its conjugate *lower prevision* by:

$$\begin{aligned}\underline{R}(X) &:= -\overline{R}(-X) = -\inf\{\alpha \in \mathbb{R} \colon -X - \alpha \in \mathcal{D}\}\\ &= \sup\{\alpha \in \mathbb{R} \colon \alpha - X \in \mathcal{D}\},\end{aligned}$$

which specifies the supremum certain loss α that the decision maker is willing to shoulder in exchange for giving away the uncertain loss X. Due to the conjugacy, it suffices to focus on the upper prevision throughout. In general, we have that $\underline{R}(X) \leq \overline{R}(X)$ for any $X \in L^\infty$. If $\underline{R}(X) = \overline{R}(X)$ $\forall X \in L^\infty$, we say that $R := \overline{R} = \underline{R}$ is a *linear prevision*, a definition which is then in line with de Finetti [8].

By applying an upper prevision to indicator gambles, we obtain an *upper probability* $\overline{P}(A) := \overline{R}(\chi_A)$, where $A \subseteq \Omega$. Correspondingly, the *lower probability* is $\underline{P}(A) := 1 - \overline{P}(A^{\mathrm{C}}) = \underline{R}(\chi_A)$. In the precise case, there is a unique relationship between (finitely) additive probabilities and linear previsions; however, upper previsions are more expressive than upper probabilities. Finally, we remark that via the so-called *natural extension*, a coherent upper probability which is defined on some subsets of events $\mathcal{A} \subseteq 2^\Omega$ can be extended to a coherent upper prevision $\mathrm{NatExt}(\overline{P})$ on L^∞, which is compatible with $\overline{P}$ in the sense that $\mathrm{NatExt}(\overline{P})(\chi_A) = \overline{P}(A)\ \forall A \in \mathcal{A}$ (cf. [47, Section 3.1]).

2. Unstable Relative Frequencies

Assume that we have some fixed sequence $\vec{\Omega} \colon \mathbb{N} \to \Omega$ on a possibility set Ω of elementary events, but that for some events $A \in \mathcal{A}$, where $\mathcal{A} \subseteq 2^\Omega$, the limiting relative frequencies do not exist. What can we do then?

In a series of papers [23, 24, 25, 26, 27, 28, 29] and a monograph [22], Ivanenko and collaborators have developed a strictly frequentist theory of "hyper-random phenomena" based on "statistical regularities". In essence, they tackle mass decision making in the context of sequences with possibly divergent relative frequencies. Like von Mises, they take the notion of a sequence as the primitive, that is, without assuming an a priori probability and then invoking the law of large numbers. The presentation of Ivanenko's theory is obscured somewhat by the great generality with which it is presented (they work with general nets, rather than just sequences). We build heavily upon their work but entirely restrict ourselves to working with sequences. While in some sense this is a weakening, our converse result (see Section 3) is actually stronger as we show that one can achieve any "statistical regularity" by taking relative frequencies of only sequences. For simplicity, we will dispense with integrals with respect to finitely additive measures in our presentation, so that there are less mathematical dependencies involved;[9] instead, we work with linear previsions. Moreover, we establish[10] the connection to imprecise probability. The contribution in this section may be viewed as unifying ideas from Ivanenko with Walley's framework.

2.1. Ivanenko's Argument — Informally

We begin by providing an informal summary of Ivanenko's construction of *statistical regularities* on sequences. Assume we are given a fixed sequence $\vec{\Omega}$ of elementary events $\vec{\Omega}(1), \vec{\Omega}(2), \ldots$, where we may intuitively think of $\mathbb{N}$ as representing time. In contrast to von Mises, who demands the existence of relative frequency limits to define probabilities, we ask for something like a probability for *all* events $A \subseteq \Omega$, even when the relative frequencies have no limit. To this end, we exploit that sequences of relative frequencies always have a non-empty set of cluster points, each of which is a finitely additive probability. Hence, a decision maker can use this set of probabilities to represent the global statistical properties of the sequence. Also, a decision maker may want to assess a value for each gamble $X \colon \Omega \to \mathbb{R}$, which is evaluated infinitely often over time. Here, the sequence of averages $n \mapsto \frac{1}{n}\sum_{i=1}^{n} X(\vec{\Omega}(i))$ is the object of interest. In the case of convergent relative frequencies, a decision maker would use the expectation to assess the risk in the limit, whereas in the general case of possible non-convergence, a different object is needed. This object turns out to be a coherent upper prevision.

[8] Here, we need the vector space assumption on the set of gambles. We also note that Walley [47, pp. 64–65] himself made a similar definition, but then proposed the more general coherence concept.

[9] Compared to integrals with finitely additive measures, working with linear previsions as in [47] appears to be an easier approach for our purposes and aids the unification with the imprecise probability literature.

[10] Ivanenko and Labkovskii [27] mention in passing that sets of probabilities also appear in [47].

2.2. Ivanenko's Argument — Formally

Let Ω be an arbitrary (finite, countably infinite or uncountably infinite) set of outcomes and fix $\vec{\Omega}\colon \mathbb{N} \to \Omega$, an Ω-valued sequence. We define a gamble $X\colon \Omega \to \mathbb{R}$ as a bounded function from Ω to $\mathbb{R}$, i.e. $\exists K \in \mathbb{R}\colon |X(\omega)| \leq K$ $\forall \omega \in \Omega$ and collect all such gambles in the set L^∞. We assume the vector space structure on L^∞ as in Section 1.2.

The set L^∞ becomes a Banach space, i.e. a complete normed vector space, under the supremum norm $\|X\|_{L^\infty} := \sup_{\omega\in\Omega} |X(\omega)|$. We denote the topological dual space of L^∞ by $(L^\infty)^*$. Recall that it consists exactly of the continuous linear functionals $\phi\colon L^\infty \to \mathbb{R}$. We endow $(L^\infty)^*$ with the *weak*-topology*, which is the weakest topology (i.e. with the fewest open sets) that makes all evaluation functionals of the form $X^*\colon (L^\infty)^* \to \mathbb{R}$, $X^*(E) := E(X)$ for any $X \in L^\infty$ and $E \in (L^\infty)^*$ continuous. Consider the following subset of $(L^\infty)^*$:

$$\mathrm{PF}(\Omega) := \{E \in (L^\infty)^*\colon E(X) \geq 0 \text{ if } X \geq 0, E(\chi_\Omega) = 1\}.$$

Due to the Alaoglu-Bourbaki theorem, this set is compact under the weak* topology.[11]

A finitely additive probability $P\colon \mathcal{A} \to [0,1]$ on some set system $\mathcal{A} \subseteq 2^\Omega$, where $\Omega \in \mathcal{A}$, is a function such that:

PF1. $P(\Omega) = 1$.

PF2. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$ and $A, B \in \mathcal{A}$.

From the sequence $\vec{\Omega}$ we induce a sequence of finitely additive probabilities $\vec{P}$, where for each $n \in \mathbb{N}$, $\vec{P}(n) := A \mapsto \frac{1}{n}\sum_{i=1}^n \chi_A(\vec{\Omega}(i))$. It is easy to check that indeed $\vec{P}(n)$ is a finitely additive probability on the whole powerset 2^Ω for any $n \in \mathbb{N}$. We shall call $\vec{P}$ the *sequence of empirical probabilities*. Recall that due to [47, Corollary 3.2.3], a finitely additive probability defined on 2^Ω can be uniquely extended (via natural extension) to a linear prevision $E_P\colon L^\infty \to \mathbb{R}$, so that $E_P(\chi_A) = P(A)\ \forall A \subseteq \Omega$. Furthermore, we know from [47, Corollary 2.8.5], that there is a one-to-one correspondence between elements of $\mathrm{PF}(\Omega)$ and linear previsions $E_P\colon L^\infty \to \mathbb{R}$. Hence, we associate to each empirical probability $\vec{P}(n)$ an *empirical linear prevision* $\vec{E}(n) := X \mapsto \mathrm{NatExt}(\vec{P}(n))(X)$, where $X \in L^\infty$ and we denote the natural extension by NatExt. We thus obtain a sequence $\vec{E}\colon \mathbb{N} \to \mathrm{PF}(\Omega)$.

On the other hand, each gamble $X \in L^\infty$ induces a sequence of evaluations as $\vec{X}\colon \mathbb{N} \to \mathbb{R}$, where $\vec{X}(n) := X\left(\vec{\Omega}(n)\right)$. For $X \in L^\infty$, we define the sequence of averages of the gamble over time as $\overrightarrow{\Sigma X}\colon \mathbb{N} \to \mathbb{R}$, where $\overrightarrow{\Sigma X}(n) := \frac{1}{n}\sum_{i=1}^n X\left(\vec{\Omega}(i)\right)$. For each fixed n, we can also view the average as a function in X, i.e. $X \mapsto \frac{1}{n}\sum_{i=1}^n X\left(\vec{\Omega}(i)\right)$. Observe that this is a coherent linear prevision and by applying it to indicator gambles χ_A, we obtain $\vec{P}(n)$. Hence, we know from [47, Corollary 3.2.3] that this linear prevision is in fact the natural extension of $\vec{P}(n)$, i.e. $\vec{E}(n) = X \mapsto \overrightarrow{\Sigma X}(n) \forall n \in \mathbb{N}$. This concludes the technical setup; we now begin reproducing Ivanenko's argument.

Since $\mathrm{PF}(\Omega)$ is a compact topological space under the subspace topology induced by the weak*-topology on $(L^\infty)^*$, we know that any sequence $\vec{E}\colon \mathbb{N} \to (L^\infty)^*$ has a non-empty closed set of cluster points. Recall that a point z is a *cluster point* of a sequence $\vec{S}\colon \mathbb{N} \to \mathcal{T}$, where $\mathcal{T}$ is any topological space, if for any neighbourhood N of z w.r.t. $\mathcal{T}$:

$$\forall n_0 \in \mathbb{N}\colon \exists n \geq n_0 : \vec{S}(n) \in N.$$

We remark that this *does not* imply that those cluster points are limits of convergent subsequences.[12] We denote the *set of cluster points* as $\mathrm{CP}(\vec{E})$. By applying these linear previsions to indicator gambles, we obtain the set of finitely additive probabilities $\mathcal{P} := \left\{A \mapsto E(\chi_A)\colon E \in \mathrm{CP}(\vec{E})\right\}$. Due to the one-to-one relationship, we might work with either $\mathrm{CP}(\vec{E})$ or $\mathcal{P}$. Following Ivanenko, we call $\mathcal{P}$ the *statistical regularity* of the sequence $\vec{\Omega}$; in the language of imprecise probability, we would call it a *credal set*. We further define

$$\begin{aligned}\overline{R}(X) &:= \sup\left\{E(X)\colon E \in \mathrm{CP}(\vec{E})\right\}, && \forall X \in L^\infty,\\ \overline{P}(A) &:= \sup\left\{P(A)\colon P \in \mathcal{P}\right\}, && \forall A \subseteq \Omega.\end{aligned}$$

Observe that $\overline{R}$ is defined on all $X \in L^\infty$ and $\overline{P}$ is defined on *all* subsets of Ω, even if Ω is uncountably infinite, since each $P \in \mathcal{P}$ is a finitely additive probability on 2^Ω. We further observe that $\overline{R}$ is a coherent upper prevision on L^∞. Correspondingly, $\overline{P}$ is a coherent upper probability on 2^Ω, which is obtained by applying $\overline{R}$ to indicator functions. This follows directly from the *envelope theorem* in [47, Theorem 3.3.3].

So far, the definition of $\overline{R}$ and $\overline{P}$ may seem unmotivated. Yet they play a special role, as we now show.

Proposition 3 *The sequence of averages $\overrightarrow{\Sigma X}$ has the set of cluster points*

$$\mathrm{CP}\left(\overrightarrow{\Sigma X}\right) = \left\{E(X)\colon E \in \mathrm{CP}(\vec{E})\right\},$$

and therefore

$$\overline{R}(X) = \sup \mathrm{CP}\left(\overrightarrow{\Sigma X}\right) = \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n).$$

Proof First observe that

$$\vec{E}(n)(X) = \overrightarrow{\Sigma X}(n).$$

We use the following result from [29, Lemma 3].[13]

[11] We refer the reader to the preprint [17] for an explication of this.

[12] This would hold under sequential compactness, which is not fulfilled here in general, but it is for finite Ω.

[13] A subtle point in the argument, which Ivanenko and Pasichnichenko [29] do not make visible, is the sequential compactness of $\mathbb{R}$, which means that for any cluster point of an $\mathbb{R}$-valued sequence we can find a subsequence converging to it.

Lemma 4 *Let $f: Y \to \mathbb{R}$ be a continuous function on a compact space Y and $\vec{y}$ a Y-valued sequence. Then* $\mathrm{CP}\left(n \mapsto f\left(\vec{y}(n)\right)\right) = f\left(\mathrm{CP}\left(\vec{y}\right)\right)$.

On the right side, the application of f is to be understood as applying f to each element in the set $\mathrm{CP}\left(\vec{y}\right)$. Consider now the evaluation functional $X^*: \mathrm{PF}(\Omega) \to \mathbb{R}$, which is continuous under the weak*-topology. The application of the lemma with $f = X^*$, $Y = \mathrm{PF}(\Omega), \vec{y} = \vec{E}$ gives:

$$\mathrm{CP}\left(n \mapsto X^*\left(\vec{E}(n)\right)\right) = X^*\left(\mathrm{CP}\left(\vec{E}\right)\right).$$

But since $X^*\left(\vec{E}(n)\right) = \overrightarrow{\Sigma X}(n)$, we obtain that $\mathrm{CP}\left(\overrightarrow{\Sigma X}\right) = X^*\left(\mathrm{CP}\left(\vec{E}\right)\right) = \left\{E(X): E \in \mathrm{CP}\left(\vec{E}\right)\right\}$. ■

A similar statement holds for the coherent upper probability.

Corollary 5 *For any $A \subseteq \Omega$, the upper probability is given by*

$$\overline{P}(A) = \limsup_{n\to\infty}\left(\vec{P}(n)(A)\right) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right).$$

Proof Just observe that $\vec{P}(n)(A) = \overrightarrow{\Sigma \chi_A}(n)$ and apply the previous result. ■

Thus the limes superior of the sequence of relative frequencies induces a coherent upper probability on 2^{Ω}; similarly, the limes superior of the sequence of a gamble's averages induces a coherent upper prevision on L^∞. By conjugacy, we have that the lower prevision and lower probability are ($\forall X \in L^\infty, \forall A \subseteq \Omega$):

$$\underline{R}(X) = \inf\left\{E(X): E \in \mathrm{CP}\left(\vec{E}\right)\right\} = \liminf_{n\to\infty} \overrightarrow{\Sigma X}(n).$$

$$\underline{P}(A) = \inf\left\{P(A): P \in \mathcal{P}\right\} = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right),$$

which are obtained in a similar way using the limes inferior. Finally, when an event is *precise* in the sense that $\overline{P}(A) = \underline{P}(A)$ (and thus the lim inf equals the lim sup and hence the limit exists), we denote the upper (lower) probability as $P(A)$ and say that the precise probability of A exists.

3. From Cluster Points to Sequence

In the previous section, we have shown how from a given sequence we can construct a coherent upper prevision from the set of cluster points $\mathrm{CP}(\vec{E})$. Here we show the converse, thus "closing the loop": given an arbitrary coherent upper prevision, we construct a sequence $\vec{\Omega}$ such that the induced upper prevision is just the specified one. We take this to be an argument for the well-groundedness of our approach. For simplicity, we assume a finite possibility set Ω.

Theorem 6 *Let $|\Omega| < \infty$. Let $\overline{R}$ be a coherent upper prevision on L^∞. There exists a sequence $\vec{\Omega}$ such that we can write $\overline{R}$ as ($\forall X \in L^\infty$):*

$$\overline{R}(X) = \sup\left\{E(X): E \in \mathcal{E}_{\vec{\Omega}}\right\}, \ \mathcal{E}_{\vec{\Omega}} := \mathrm{CP}\left(\vec{E}_{\vec{\Omega}}\right),$$

where $\vec{E}_{\vec{\Omega}}(n) = X \mapsto \frac{1}{n}\sum_{i=1}^{n} X\left(\vec{\Omega}(i)\right)$, $\forall X \in L^\infty$.

For the constructive proof and two examples, one of which corresponds to the vacuous upper prevision, see the extended preprint [17].

The significance of this result is that it establishes strictly frequentist *semantics* for coherent upper previsions. It shows that to any decision maker who, in the subjectivist fashion, uses a coherent upper prevision, we can associate a sequence, which would yield the same upper prevision in a strictly frequentist way. We interpret this result as evidence for the naturalness, and arguably completeness, of our theory.

Ivanenko [22] offers a somewhat similar result to Theorem 6 by generalizing from sequences to *sampling nets*. Ivanenko's [2010] main result states that "any sampling directedness has a regularity, and any regularity is the regularity of some sampling directedness." [22, Theorem 4.2]. We provide a brief introduction to these sampling nets in the extended preprint [17]. Our result is more parsimonious in the sense that it relies only on sequences, which are arguably more intuitive objects than such sampling nets.

Our result should also be compared to Theorem 4.2 in [48] and Theorem 2.2 in [38]. On the one hand, our result is stronger since it holds for upper previsions, whereas Theorem 4.2 in [48] and Theorem 2.2 in [38] hold for upper probabilities only; note that upper previsions are more expressive than upper probabilities.[14] On the other hand, Theorem 2.2 in [38] is stronger in the sense that it guarantees that the same upper probability is induced when applying selection rules.

4. Unstable Conditional Probability

Now consider again a fixed sequence $\vec{\Omega}$ on an arbitrary possibility set Ω. An interesting aspect of the strictly frequentist approach is that there is a natural way of introducing conditional probabilities for events $A, B \subseteq \Omega$, which is the same for the case of converging or diverging relative frequencies. Furthermore, this approach generalizes directly to gambles. We will observe that this, perhaps surprisingly, yields the *generalized Bayes rule*. In the precise case, the standard Bayes rule is recovered.

[14]Indeed, the proof of Theorem 4.2 in [48] exploits this simplification by assuming that the credal set has a finite number of extreme points.

Recall that for a countably or finitely additive probability Q on 2^{Ω}, we can define conditional probability as:

$$Q(A|B) \coloneqq \frac{Q(A \cap B)}{Q(B)}, \quad A, B \subseteq \Omega, \text{ if } Q(B) > 0. \quad (1)$$

Important here is the condition that $Q(B) > 0$. Conditioning on events of measure zero may create trouble. Kolmogorov then allows the conditional probability to be arbitrary. This is rather unfortunate, as there arguably are settings where one would like to condition on events of measure zero.

As a prerequisite, given a linear prevision $E \in \mathrm{PF}(\Omega)$, we define the conditional linear prevision as:

$$E(X|B) \coloneqq \frac{E(\chi_B X)}{E(\chi_B)}, \quad \text{if } E(\chi_B) > 0. \quad (2)$$

The application to indicator gambles then recovers conditional probability. As long as $E(\chi_B) > 0$, it is insignificant whether we condition the linear prevision, or instead condition on the level of its underlying probability and then naturally extend it; confer [47, Corollary 3.2.3].

Nearly in line with Kolmogorov's conditional probability, von Mises started from the following intuitive, frequentist view: the probability of an event A conditioned on an event B is the frequency of the occurence of the event A given that B happens. In what follows, we build upon this idea, which von Mises called "partition operation" [46, p. 22]. Walley and Fine [48, Section 4.3] have extended this definition to the divergent case of conditional probability on a finite possibility space; we further extend it to conditional upper previsions on arbitrary possibility spaces and link them to the generalized Bayes rule. As a technical preliminary, we define a wrapper function $\Psi\colon \mathrm{PF}(\Omega) \cup \{\bot\} \to \mathrm{PF}(\Omega)$ as:

$$\Psi(P) \coloneqq \begin{cases} P_0 & \text{if } P = \bot, \\ P & \text{otherwise,} \end{cases}$$

where P_0 is an arbitrary finitely additive probability on 2^{Ω}.

4.1. Conditional Probability

Recall our sequence of unconditional finitely additive probabilities $\vec{P}(n) \coloneqq A \mapsto \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right)$. We want to define a similar sequence of *conditional* finitely additive probabilities. A very natural approach is the following: let $A, B \subseteq \Omega$ be such that $\vec{\Omega}(i) \in B$ for at least one $i \in \mathbb{N}$. We write $2^{\vec{\Omega}}_{1+}$ for the set of such events, i.e. events which occur at least once in the sequence. Define a sequence of conditional probabilities $\vec{P}(\cdot|B)\colon \mathbb{N} \to \mathrm{PF}(\Omega)$ by

$$\vec{P}(\cdot|B)(n) \coloneqq \Psi\left(A \mapsto \frac{\sum_{i=1}^{n}(\chi_A \cdot \chi_B)\left(\vec{\Omega}(i)\right)}{\sum_{i=1}^{n} \chi_B\left(\vec{\Omega}(i)\right)}\right),$$

where we consider only those $\vec{\Omega}(i)$ which lie in B, and hence we adapt the relative frequencies to the occurrence of B. Informally, this is simply counting $|A \text{ and } B \text{ occured}|/|B \text{ occured}|$. Throughout, we demand that the event B on which we condition is in $2^{\vec{\Omega}}_{1+}$, i.e. occurs *at least once* in the sequence. Note that this is a much weaker condition than demanding that $P(B) > 0$, if B is precise. Denote by n_B the smallest index so that $\vec{\Omega}(n_B) \in B$. Note that $\vec{P}(A|B)(n) = \vec{P}(n)(A \cap B)/\vec{P}(n)(B)$ for $n \geq n_B$.

Even though the probability is conditional, we deal with a sequence of finitely additive probabilities again. Hence, we can now essentially repeat the argument from Section 2.2. To each $\vec{P}(\cdot|B)(n)$, associate its uniquely corresponding linear prevision $\vec{E}(\cdot|B)(n)$, which is of course given by $(\forall X \in L^{\infty}, \forall n \geq n_B)$:

$$\vec{E}(\cdot|B)(n) = \overrightarrow{\Sigma X}|B(n) \coloneqq X \mapsto \frac{\sum_{i=1}^{n}(X \cdot \chi_B)\left(\vec{\Omega}(i)\right)}{\sum_{i=1}^{n} \chi_B\left(\vec{\Omega}(i)\right)}.$$

It is easy to check that $\vec{E}(\cdot|B)(n)$ is coherent. For $n < n_B$, set $\vec{E}(\cdot|B)(n) = \mathrm{NatExt}(P_0)$. From the weak* compactness of $\mathrm{PF}(\Omega)$, we obtain a non-empty closed set of cluster points $\mathrm{CP}(\vec{E}(\cdot|B))$.

Definition 7 *If $B \in 2^{\vec{\Omega}}_{1+}$, we define the conditional upper prevision and the conditional upper probability as:*

$$\overline{R}(X|B) \coloneqq \sup\left\{E(X)\colon E \in \mathrm{CP}\left(\vec{E}(\cdot|B)\right)\right\}, \ \forall X \in L^{\infty},$$
$$\overline{P}(A|B) \coloneqq \sup\left\{Q(A)\colon Q \in \mathrm{CP}\left(\vec{P}(\cdot|B)\right)\right\}, \ \forall A \subseteq \Omega.$$

Since they are expressed via an envelope representation,[15] $\overline{R}$ and $\overline{P}$ are automatically coherent [47, Theorem 3.3.3]. By similar reasoning as in Section 2.2, we get the following representation.

Corollary 8 *The conditional upper prevision (probability) can be represented as ($B \in 2^{\vec{\Omega}}_{1+}$):*

$$\overline{R}(X|B) = \limsup_{n\to\infty} \overrightarrow{\Sigma X}|B(n), \qquad \forall X \in L^{\infty},$$
$$\overline{P}(A|B) = \limsup_{n\to\infty} \vec{P}(A|B)(n), \quad \forall A \subseteq \Omega.$$

Also, we get the corresponding lower quantities $\underline{R}(X|B) = \liminf_{n\to\infty} \overrightarrow{\Sigma X}|B(n)$ and $\underline{P}(A|B) = \liminf_{n\to\infty} \vec{P}(A|B)(n)$. Note that these definitions also have reasonable frequentist semantics even when B occurs only finitely often; then the sequence $\vec{P}(\cdot|B)$ is eventually constant and we have $\overline{P}(A|B) = |A \text{ and } B \text{ occured}|/|B \text{ occured}|$. For instance, if A and B occur just once, but simultaneously, then $\overline{P}(A|B) = \underline{P}(A|B) =$

[15] An envelope representation expresses a coherent upper prevision as a supremum over a set of linear previsions.

1. This is an advantage over Kolmogorov's approach, where conditioning on events of measure zero is not meaningfully defined.

We now further analyze the conditional upper probability and the conditional upper prevision measure. As a warm-up, we consider the case of precise probabilities. If for some event $A \subseteq \Omega$, we have $\overline{P}(A|B) = \underline{P}(A|B)$, we write $\tilde{P}(A|B) := \lim_{n\to\infty} \overrightarrow{P}(A|B)(n)$.

Proposition 9 *Assume $P(B), P(A \cap B)$ exist for some $A, B \subseteq \Omega$ and $P(B) > 0$. Then it holds that $\tilde{P}(A|B) = P(A|B)$, where $P(\cdot|B)$ is the conditional probability in the sense of Equation 1.*

Thus, when the relative frequencies of B and $A \cap B$ converge, we reproduce the classical definition of conditional probability. Now what happens under non-convergence?

4.2. The Generalized Bayes Rule

We now relax the assumptions of Proposition 9 and only demand that $\underline{P}(B) > 0$.[16] Then we observe that the conditional upper prevision coincides with the *generalized Bayes rule*, which is an important updating principle in imprecise probability (see e.g. [36]). The unconditional set of desirable gambles is:

$$\mathcal{D}_{\vec{\Omega}} := \left\{X \in L^\infty : \overline{R}(X) \le 0\right\}.$$

Definition 10 *For $\underline{P}(B) > 0$, we define the* conditional set of desirable gambles *as:*

$$\mathcal{D}_{\vec{\Omega}|B} := \left\{X \in L^\infty : X\chi_B \in \mathcal{D}_{\vec{\Omega}}\right\}.$$

and a corresponding upper prevision, which we call the generalized Bayes rule*, as:*

$$\begin{aligned}\mathrm{GBR}(X|B) &:= \inf\left\{\alpha \in \mathbb{R} : X - \alpha \in \mathcal{D}_{\vec{\Omega}|B}\right\}\\ &= \inf\left\{\alpha \in \mathbb{R} : \chi_B(X-\alpha) \in \mathcal{D}_{\vec{\Omega}}\right\}\\ &= \inf\left\{\alpha \in \mathbb{R} : \overline{R}\left(\chi_B(X-\alpha)\right) \le 0\right\}.\end{aligned}$$

Remark 11 In fact, Walley [47, Section 6.4] defines the generalized Bayes rule as the solution of $\overline{R}(\chi_B(X-\alpha)) = 0$ for α. It can be checked that this solution coincides with Definition 10.[17]

Proposition 12 *If $\underline{P}(B) > 0$, then $\overline{R}(X|B) = \mathrm{GBR}(X|B)$.*

The proof is in the extended preprint [17].

Remark 13 Note that $X \mapsto \limsup_{n\to\infty} \overrightarrow{\Sigma(X\chi_B)} = \overline{R}(X\chi_B)$ is *not* in general a coherent upper prevision on L^∞, as it can violate UP1. To see this, take for example $X(\omega) = -1$ for a $B \subseteq \Omega$ where $\underline{P}(B) < 1$. In general, we have $\mathrm{GBR}(X|B) \neq \overline{R}(X\chi_B)$.

As a consequence, we can apply the classical representation result for the generalized Bayes rule.

Corollary 14 *If $\underline{P}(B) > 0$, the conditional upper prevision can be obtained by conditioning each linear prevision in the set of cluster points, that is:*

$$\overline{R}(X|B) = \sup\left\{E(X|B) : E \in \mathrm{CP}\left(\vec{E}\right)\right\}, \tag{3}$$

where conditioning of the linear previsions is in the sense of Equation 2.

This follows from [47, Theorem 6.4.2]. Intuitively, it makes no difference whether we consider the cluster points of the sequence of conditional probabilities or whether we condition all probabilities in the set of cluster points in the classical sense.

Closely related to conditional probability is the concept of *statistical independence*, which plays a central role not only in Kolmogorov's [10, p. 37], but more generally in most probability theories (Levin [35]; Fine [13, Sections IIF, IIIG and VH]). In the extended preprint [17] we offer an independence concept for the case of possibly diverging relative frequencies and discuss how it relates to the classical independence notion in Kolmogorov's framework.

5. Related Work and Conclusion

While divergence of relative frequencies has been linked to imprecise probability before, this has almost exclusively been done in settings which are not *strictly* frequentist. Fine [12] was one of the first authors to critically evaluate the hypothesis of statistical stability and observed that this widespread hypothesis is regarded as a "striking instance of order in chaos" in the statistics community. This paper may be seen as a predecessor to a long line of work by Terrence Fine and collaborators, [14, 48, 34, 19, 15, 37, 38, 40, 11]; see also [16] for an introduction. A central motivation behind this work was to develop a frequentist model for the puzzling case of stationary, unstable phenomena with bounded time averages. What differentiates this work from ours is that we take a *strictly frequentist* approach: we explicitly define the upper probability and upper prevision from a given sequence. In contrast, the above works (with the exceptions of [38], [11] and Section 4.3 in [48]) use an imprecise probability to represent a single trial in a

[16] This condition is indispensable in order to make the connection to the generalized Bayes rule.

[17] The conditional set of desirable gambles is considered for instance in [1] and [49], but there the link to the generalized Bayes rule is not made technically clear.

sequence of unlinked repetitions of an experiment, and then induce an imprecise probability via an infinite product space. This can be understood as a generalization of the standard frequentist approach, where one would assume that $X_1, X_2, \ldots$ form an i.i.d. sequence. While we are not against this framework as such, our motivation was to work with a parsimonious set of assumptions. To this end, we took the sequence as the primitive entity, without relying on an underlying "individual" (imprecise) probability.

In order to access a more powerful toolbox, de Cooman, de Bock and Persiau more recently studied the interplay of imprecise probability and randomness in the game-theoretic setup [7, 39] with references to a frequentist perspective, see e.g. [7, Corollary 28]. While we believe there is potential for establishing relations between our approach and theirs, the differences in technical setup make it challenging to do so straightforwardly.

Within the setup of [48], Cozman and Chrisman [6, Theorem 1, Theorem 2] proposed an estimator for the underlying imprecise probability of the sequence. Specifically, they computed relative frequencies along a finite set of selection rules (yet without referring to von Mises) and then took their minimum to obtain bounds on the lower probability. What motivated the authors to do this is an assumption on the data-generating process: at each trial, "nature" may select a different distribution from a set of probability measures; the trials are then independent but not identically distributed. A related approach is that of [11]. They offered the metaphor of an *analytical microscope*. With more and more complex selection rules ("powerful lenses"), along which relative frequencies are computed, more and more structure of the set of probabilities comes to light. The authors also proposed a way to simulate data from a set of probability measures.

In this work, we have extended strict frequentism to the case of possibly divergent relative frequencies and sample averages, tying together threads from [45], [22] and [47] and thus providing a unified account based on coherent upper previsions. In particular, we have recovered the generalized Bayes rule from a strictly frequentist perspective by taking inspiration from von Mises [45] definition of conditional probability. Our converse result (Section 3) provides strictly frequentist semantics for coherent upper previsions; previously, this has been done only for the simpler case of coherent upper probabilities.

Statisticians exclusively assume that their data is part of a stable sequence, but the hypothesis of perfect statistical stability is just a *hypothesis*. A key point is that when one blindly *assumes* convergence of relative frequencies, one will not notice when it is violated — in the practical case, when only a finite sequence is given, such a violation would correspond to instability of relative frequencies even for long observation intervals [18], in the sense that divergence in an infinite sequence is an idealization of instability in the case of finite data. In this work, we have rejected the assumption of perfect statistical stability; furthermore, in contrast to other related work, we have aimed to weaken the set of assumptions by taking the concept of a sequence as the primitive. However, this gives rise to the critique that no finite part of a sequence has any bearing on what the limit is, as has been pointed out by other authors whose studies attempted a frequentist understanding of imprecise probability (e.g. [4]). So what is the empirical content of our theory, what are its practical implications?

The reader may wonder why we have introduced von Mises frequentist account but not further used selection rules afterwards. In von Mises' framework, the set of selection rules expresses randomness assumptions about the sequence, similar to what the i.i.d. assumption achieves in the standard picture. In our view, randomness assumptions are *the* key to empower generalization in the finite data setting. Hence, to supplement our theory with empirical content, the introduction of selection rules is needed. However, multiple directions can be pursued here. For instance, Papamarcou and Fine [38] have defined the concept of an *unstable collective*, where divergence remains unchanged when applying selection rules. By contrast, we could introduce a set of selection rules and assume that relative frequencies converge *within each* selection rule, but to potentially different limits across selection rules.[18] Hence, we view this paper as only the first step of a larger research agenda. The next step is to incorporate randomness assumptions into the picture and explore the connections between various possible approaches, specifically how different ways of relaxing vM1 and vM2 are related.

Acknowledgments

This work was was funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich and Rabanus Derr.

Author Contributions

C. Fröhlich is responsible for most of the content and writing. R. Derr contributed through many discussions and helpful suggestions and made some contributions to writing. The comparison of Ivanenko's and Walley's theories was suggested by R. C. Williamson, who supervised the project and is responsible for the majority of the proof of Theorem 6.

[18] As demonstrated by Examples 2 and 3 in [6], converging relative frequencies within selection rules can lead to both overall convergence or divergence (on the whole sequence).

References

[1] Thomas Augustin, Frank P.A. Coolen, Gert De Cooman, and Matthias C.M. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[2] Émile Borel. *Probability and Certainty*. Walker and Company, 1963. Translation of *Probabilité et Certitude*, Presses Universitaires de France, 1950.

[3] Cristian Calude and Tudor Zamfirescu. Most numbers obey no probability laws. *Publicationes Mathematicae Debrecen*, 54(Supplement):619–623, 1999.

[4] Marco E. G. V. Cattaneo. Empirical interpretation of imprecise probabilities. In Alessandro Antonucci, Giorgio Corani, Inés Couso, and Sébastien Destercke, editors, *Proceedings of the Tenth International Symposium on Imprecise Probability: Theories and Applications*, volume 62 of *Proceedings of Machine Learning Research*, pages 61–72. PMLR, 2017.

[5] Alonzo Church. On the concept of a random sequence. *Bulletin of the American Mathematical Society*, 46(2): 130–135, 1940.

[6] Fabio Cozman and Lonnie Chrisman. Learning convex sets of probability from data. Technical report, Carnegie Mellon University, 1997. CMU-RI-TR 97-25.

[7] Gert De Cooman and Jasper De Bock. Randomness is inherently imprecise. *International Journal of Approximate Reasoning*, 141:28–68, 2022.

[8] Bruno de Finetti. *Theory of probability: A critical introductory treatment*. John Wiley & Sons, 1974/2017.

[9] Rabanus Derr and Robert C. Williamson. Fairness and randomness in machine learning: Statistical independence and relativization. *arXiv preprint arXiv:2207.13596*, 2022.

[10] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

[11] Pablo Ignacio Fierens, Leonardo Chaves Rêgo, and Terrence L. Fine. A frequentist understanding of sets of measures. *Journal of Statistical Planning and Inference*, 139(6):1879–1892, 2009.

[12] Terrence L. Fine. On the apparent convergence of relative frequency and its implications. *IEEE Transactions on Information Theory*, 16(3):251–257, 1970.

[13] Terrence L. Fine. *Theories of probability: An examination of foundations*. Academic Press, 1973.

[14] Terrence L. Fine. A computational complexity viewpoint on the stability of relative frequency and on stochastic independence. In William L. Harper and Clifford Alan Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume 1*, pages 29–40. Springer, 1976.

[15] Terrence L. Fine. Lower probability models for uncertainty and nondeterministic processes. *Journal of Statistical Planning and Inference*, 20(3):389–411, 1988.

[16] Terrence L. Fine. Mathematical alternatives to standard probability that provide selectable degrees of precision. In Alan Hájek and Christopher Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2016.

[17] Christian Fröhlich, Rabanus Derr, and Robert C. Williamson. Strictly frequentist imprecise probability. *arXiv preprint arXiv:2302.03520*, 2023.

[18] Igor I. Gorban. *The Statistical Stability Phenomenon*. Springer, 2017.

[19] Yves L. Grize and Terrence L. Fine. Continuous lower probability-based models for stationary processes with bounded and divergent time averages. *The Annals of Probability*, 15(2):783–803, 1987.

[20] Stanley P. Gudder. *Stochastic methods in quantum mechanics*. North Holland, 1979.

[21] Alan Hájek. Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235, 2009.

[22] Victor I. Ivanenko. *Decision Systems and Nonstochastic Randomness*. Springer, 2010.

[23] Victor I. Ivanenko and Valery A. Labkovskii. On the functional dependence between the available information and the chosen optimality principle. In Vadim I. Arkin, A. Shiraev, and R. Wets, editors, *Stochastic Optimization*, pages 388–392. Springer, 1986.

[24] Victor I. Ivanenko and Valery A. Labkovskii. A class of criterion-choosing rules. *Doklady Akademii Nauk SSSR*, 31(3):204–205, 1986.

[25] Victor I. Ivanenko and Valery A. Labkovskii. A model of nonstochastic randomness. *Doklady Akademii Nauk SSSR*, 310(5):1059–1062, 1990.

[26] Victor I. Ivanenko and Valery A. Labkovskii. On the studying of mass phenomena. In *Proceedings of the 12th IFAC Triennial World Congress*, pages 693–696, 1993.

[27] Victor I. Ivanenko and Valery A. Labkovskii. On regularities of mass phenomena. *Sankhya A*, 77:237–248, 2015.

[28] Victor I. Ivanenko and Bertrand Munier. Decision making in 'random in a broad sense' environments. *Theory and Decision*, 49:127–150, 2000.

[29] Victor I. Ivanenko and Illia Pasichnichenko. Expected utility for nonstochastic risk. *Mathematical Social Sciences*, 86:18–22, 2017.

[30] Andrei Yu Khrennikov. *Probability and randomness: quantum versus classical*. Imperial College Press, 2016.

[31] Andreĭ Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.

[32] Andreĭ Nikolaevich Kolmogorov. *Foundations of the theory of probability: Second English Edition*. Chelsea Publishing Company, 1956.

[33] Andrei Nikolaevich Kolmogorov. On logical foundations of probability theory. In K. Ito and J.V. Prokhorov, editors, *Probability theory and mathematical statistics: Proceedings of the Fourth USSR – Japan Symposium, held at Tbilisi, USSR, August 23-29, 1982*, pages 1–5. Springer, 1983.

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

[35] Leonid A. Levin. A concept of independence with applications in various fields of mathematics. Technical Report MIT/LCS/TR-235, MIT, Laboratory for Computer Science, 1980.

[36] Enrique Miranda and Gert de Cooman. Lower previsions. In *Introduction to Imprecise Probabilities*, chapter 2, pages 28–55. John Wiley & Sons, Ltd, 2014.

[37] Adrian Papamarcou and Terrence L. Fine. Stationarity and almost sure divergence of time averages in interval-valued probability. *Journal of Theoretical Probability*, 4(2):239–260, 1991.

[38] Adrian Papamarcou and Terrence L. Fine. Unstable collectives and envelopes of probability measures. *The Annals of Probability*, 19(2):893–906, 1991.

[39] Floris Persiau, Jasper De Bock, and Gert de Cooman. On the (dis)similarities between stationary imprecise and non-stationary precise uncertainty models in algorithmic randomness. *International Journal of Approximate Reasoning*, 151:272–291, 2022.

[40] Amir Sadrolhefazi and Terrence L. Fine. Finite-dimensional distributions and tail behavior in stationary interval-valued probability models. *The Annals of Statistics*, 22(4):1840–1870, 1994.

[41] Claus P. Schnorr. *Zufälligkeit und Wahrscheinlichkeit: eine algorithmische Begründung der Wahrscheinlichkeitstheorie*. Springer, 2007.

[42] Teddy Seidenfeld, Jessica Cisewski, Jay Kadane, Mark Schervish, and Rafael Stern. When large also is small conflicts between measure theoretic and topological senses of a negligible set. Presented at Pitt Workshop 3/17, 2017.

[43] Alexander Shen. Algorithmic information theory and foundations of probability. *arXiv preprint arXiv:0906.4411v1*, 2009.

[44] Michiel van Lambalgen. Von Mises' definition of random sequences reconsidered. *The Journal of Symbolic Logic*, 52(3):725–755, 1987.

[45] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5(1): 52–99, 1919.

[46] Richard von Mises and Hilda Geiringer. *Mathematical theory of probability and statistics*. Academic press, 1964.

[47] Peter Walley. *Statistical reasoning with imprecise probabilities*. Chapman-Hall, 1991.

[48] Peter Walley and Terrence L. Fine. Towards a frequentist theory of upper and lower probability. *The Annals of Statistics*, 10(3):741–761, 1982.

[49] Gregory Wheeler. A gentle approach to imprecise probability. In Thomas Augustin, Fabio Cozman, and Gregory Wheeler, editors, *Reflections on the Foundations of Statistics: Essays in Honor of Teddy Seidenfeld*. Springer, 2021.