Strictly Frequentist Imprecise Probability*

Christian Fröhlich *christian.froehlich@uni-tuebingen.de*
Department of Computer Science
University of Tübingen

Rabanus Derr *rabanus.derr@uni-tuebingen.de*
Department of Computer Science
University of Tübingen

Robert C. Williamson *bob.williamson@uni-tuebingen.de*
Department of Computer Science
University of Tübingen
and Tübingen AI Center

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 data that exhibit diverging relative frequencies. In doing so, we develop a close connection with the theory of 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. Finally, 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 which demonstrates the naturalness, and arguably completeness, of our theory.

1 Introduction

> *Do other statistical properties, that can not be reduced to stochasticness, exist? This question did not attract any attention until the applications of the probability theory concerned only natural sciences. The situation is definitely [changing], when one studies social phenomena: the stochasticness gets broken as soon as we [deal] with deliberate activity of people.* — Victor Ivanenko and Valery Labkovsky (1993)

It is now (almost surely) universally acknowledged that probability theory ought to be based on Kolmogorov's (1933) mathematical axiomatization (translated in (Kolmogorov, 1956)).[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. Our motivation for the present paper comes from questioning both of these.

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 —

*This paper is an extended and revised version of *Towards a strictly frequentist theory of imprecise probability*, Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications, PMLR 215:230-240, 2023.

[1] An important exception is quantum probability (Gudder, 1979; Khrennikov, 2016).

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 (2017) 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 expect 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; confer (King & Kay, 2020).

On the other hand, there is now ample experimental evidence that relative frequencies can fail to stabilize even under very long observation intervals (Gorban, 2017, 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 (Kolmogorov, 1983).

Building on the works of von Mises & Geiringer (1964), Walley & Fine (1982) and Ivanenko (2010), our goal is to establish a broader theory, which is also applicable to phenomena which are outside of the scope of Kolmogorov's theory by exhibiting unstable relative frequencies.

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 (Schnorr, 2007, p. 8); see also (Calude & Zamfirescu, 1999; Seidenfeld et al., 2017) for critical arguments. Moreover, the examples in (Gorban, 2017, Part II) show that sequences with seemingly non-converging relative frequencies (fluctuating substantially even for long observation intervals) are not "rare" in practice. In Appendix D we examine the question of how pathological or normal such sequences are in more depth.

Conceptually, the underlying problem is that the probability measure P, which is used to measure the event $\{\lim_{n\to\infty} \bar{X}_n = \mathbb{E}[X]\}$ has no clear meaning. Of course, in the subjectivist spirit, one could interpret it as assigning a *belief* in the statement that convergence takes place. But it is unclear what a frequentist interpretation of P would look like. As La Caze (2016) observed:

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

> Importantly, "almost sure convergence" is also given a frequentist interpretation. Almost sure convergence is taken to provide a justification for assuming that the relative frequency of an attribute *would* converge to the probability in actual experiments *were* the experiment to be repeated indefinitely [emphasis in original].

But again, it is unclear on what ground P can be given this interpretation and according to Hájek (2009) this leads to a regress to mysterious "meta-probabilities". Furthermore, the theoretical LLN requires that P be countably additive, which is problematic under a frequency interpretation (Hájek, 2009, pp. 229–230).

Given these complications, we opt for a different approach, namely a *strictly frequentist* one. Reaching back to Richard von Mises' (1919) 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 actually 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 (Borel, 1963, pp. 27ff.) (which begs the question what such "probabilities" mean) we follow the approach of Ivanenko (2010), reformulate his construction of a *statistical regularity* of a sequence, and discover that it is closely connected to the subjectivist theory of *imprecise probability*, more specifically, to the theory of lower and upper previsions. In essence, to each 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 (Hájek, 2009). On an arbitrary (possibly infinite) set of outcomes, the relative frequencies induce a coherent upper probability and the sample averages induce a coherent upper prevision in the sense of Walley (1991). In the convergent case, this reduces to a precise, finitely additive probability and a linear prevision, respectively. 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.

Furthermore, we demonstrate that (for a finite set of outcomes) the reverse direction works, too: given a set of probability measures, we can explicitly construct a sequence, which corresponds to this set in the sense that its relative frequencies have this set of cluster points. Thereby we establish strictly frequentist *semantics* for imprecise probability: a subjective decision maker who uses a set of probability measures to represent their belief can also be understood as assuming an implicit underlying sequence and reasoning in a frequentist way thereon.

We emphasize that we leave the question of randomness in this broader picture for future research (see the discussion in Section 6). Note that throughout the paper we work with a loss-based orientation (see Section 1.2).

1.1 Von Mises — The Frequentist Perspective

Our approach is inspired by Richard von Mises' (1919) axiomatization of probability theory (refined and summarized in (von Mises & Geiringer, 1964)). 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.

> The calculus of probability, i.e. the theory of probabilities, in so far as they are numerically representable, is the theory of definite observable phenomena, repetitive or mass events. Examples are found in games of chance, population statistics, Brownian motion etc. (von Mises, 1981, p. 102).

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 $\mathscr{A} \subseteq 2^{\Omega}$.

Definition 1.1. *Consider a tuple* $\left(\Omega, \overrightarrow{\Omega}, \mathscr{A}, \mathscr{S}\right)$ *with the following data:*

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

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

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

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

We call this limit the probability of A.

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

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

Here, we view $\overrightarrow{\Omega}$ as a sequence of elementary outcomes $\omega \in \Omega$, for some possibility space Ω on which we have a set system of events $\mathscr{A}$. 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. Note it actually comprises *two claims*: 1) the limit *exists* and 2) it is the *same* as the limit in vM1. It is best understood by viewing a selection rule $\overrightarrow{S}_j : \mathbb{N} \to \{0,1\}$ as selecting a subsequence of the original sequence $\overrightarrow{\Omega}$ and then demanding that the limiting relative frequencies thereof coincide with those of the original sequence. Such a selection rule is

[3] In fact, $\mathscr{A}$ does not necessarily has 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 (von Mises & Geiringer, 1964).

[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.

called *admissible*, whereas a selection rule which would give rise to different limiting relative frequencies for at least one $A \in \mathscr{A}$ would be *inadmissible*. 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 is inadmissible, an adversary could use this knowledge to strategically offer a bet on the next outcome and thereby make long-run profit, at the expense of our 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 (van Lambalgen, 1987)). Thus, von Mises explicitly relativizes randomness with respect to a problem-specific set of selection rules (von Mises & Geiringer, 1964, p. 12).[6] A sequence which forms a collective ("is random with respect to") one set of selection rules, might not form a collective with respect to another set.

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, since our focus is the axiom (or hypothesis) of statistical stability. In this spirit, we put aside the question of randomness here.

We are motivated by the following question. What happens to 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 *imprecise probability*. In summary, we establish a strictly frequentist theory of imprecise probability.

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 (1991). 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 a 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 machine learning, we take positive values to represent loss and negative values to represent reward.[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 (1991) proposed a betting interpretation of imprecise probability, which is inspired by de Finetti (1974/2017), 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 smallest number $\overline{R}(X)$ so that $X - \overline{R}(X)$ is a desirable transaction

[6] This class of selection rules necessarily must be specified in advance; confer (Shen, 2009). A prominent line of work aspires to fix the set of selection rules as all partially computable selection rules (Church, 1940), but there is no compelling reason to elevate this to a universal choice; cf. (Derr & Williamson, 2022) for an elaborated critique.

[7] Unfortunately, this introduces tedious sign flips when comparing results to Walley (1991).

to our decision maker, where she incurs the loss $X(\omega)$ but in exchange gets the reward $-\overline{R}(X)$. Formally:

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

where $\mathscr{D}$ is a *set of desirable gambles*. Walley (1991, Section 2.5) argued for a criterion of coherence, which any reasonable functional $\overline{R}$ should satisfy, and consequently obtained the following characterization (Walley, 1991, Theorem 2.5.5), which we shall take here as an axiomatic *definition* instead.[8]

Definition 1.2. *A functional* $\overline{R} \colon L^\infty \to \mathbb{R}$ *is a* coherent upper prevision *if it satisfies* $\forall X, Y \in L^\infty$*:*

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

Together, these properties also imply $\forall X, Y \in L^\infty$ (Walley, 1991, p. 76):

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

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

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

which specifies the highest certain loss α that the decision maker is willing to shoulder in exchange for giving away the loss $X(\omega)$, i.e. receiving the reward $-X(\omega)$. 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 aligns with de Finetti (1974/2017).

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 $\underline{P}(A) := 1 - \overline{P}(A^C) = \underline{R}(\chi_A)$. 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 $\mathscr{A} \subseteq 2^\Omega$ can be extended to a coherent upper prevision $\operatorname{NatExt}(\overline{P})$ on L^∞, which is compatible with $\overline{P}$ in the sense that $\operatorname{NatExt}(\overline{P})(\chi_A) = \overline{P}(A)\ \forall A \in \mathscr{A}$ (cf. (Walley, 1991, Section 3.1)).

2 Unstable Relative Frequencies

Assume that we have some fixed sequence $\overrightarrow{\Omega} \colon \mathbb{N} \to \Omega$ on a possibility set Ω of elementary events, but that for some events $A \in \mathscr{A}$, where $\mathscr{A} \subseteq 2^\Omega$, the limiting relative frequencies do not exist. What can we do then? In a series of papers (Ivanenko & Labkovskii, 1986a;b; 1990; 1993; Ivanenko & Munier, 2000; Ivanenko & Labkovskii, 2015; Ivanenko & Pasichnichenko, 2017) and a monograph (Ivanenko, 2010), 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. They explicitly recognize that "stochasticness gets broken as soon as we deal with deliberate activity of

[8] Here, we need the vector space assumption on the set of gambles. We also note that Walley (1991, pp. 64–65) himself made a similar definition, but then proposed the more general coherence concept.

people" (Ivanenko & Labkovskii, 1993)[9]. 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;[10] instead, we work with linear previsions. Moreover, we establish[11] the connection to imprecise probability and give a different justification for the construction. 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 $\overrightarrow{\Omega}\colon \mathbb{N} \to \Omega$ of elementary events $\overrightarrow{\Omega}(1), \overrightarrow{\Omega}(2), \ldots$, 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. In fact, we will see that this induces a coherent upper probability. Also, our decision maker is interested in assessing a value for each gamble $X\colon \Omega \to \mathbb{R}$, which is evaluated infinitely often over time. Here, the sequence of averages $n \mapsto \frac{1}{n}\sum_{i=1}^{n} X(\overrightarrow{\Omega}(i))$ 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. We provide a justification for using this upper prevision to assess the value of a gamble, which links it to imprecise probability.

2.2 Ivanenko's Argument — Formally

Let Ω be an arbitrary (finite, countably infinite or uncountably infinite) set of outcomes and fix $\overrightarrow{\Omega}\colon \mathbb{N} \to \Omega$, 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. $\exists K \in \mathbb{R}\colon |X(\omega)| \le K \ \forall \omega \in \Omega$ 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 $E\colon L^\infty \to \mathbb{R}$. 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 $(L^\infty)^*$:

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

Due to the Alaoglu-Bourbaki theorem, this set is compact under the weak* topology, see Appendix A.1.

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

[9] This not only occurs because of non-equilibrium effects, but also from feedback loops, what has become known as "performativity" (MacKenzie et al., 2007) or "reflexivity" (Soros, 2009). See the epigraph at the beginning of the present paper.

[10] The two well-known accounts of the theory of integrals with finitely additive measures are (Rao & Rao, 1983) and (Dunford & Schwartz, 1988). The theory of linear previsions as laid out in (Walley, 1991) appears to be an easier approach for our purposes.

[11] Ivanenko & Labkovskii (2015) mention in passing that sets of probabilities also appear in (Walley, 1991).

PF1. $P(\Omega) = 1$.
PF2. $P(A \cup B) = P(A) + P(B)$ whenever $A \cap B = \emptyset$ and $A, B \in \mathscr{A}$.

We induce a sequence of finitely additive probabilities $\vec{P}$ where $\vec{P}(n) := A \mapsto \frac{1}{n}\sum_{i=1}^{n} \chi_A(\vec{\Omega}(i))$ for each $n \in \mathbb{N}$. 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*. Due to (Walley, 1991, 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 (Walley, 1991, 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 finite 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 (Walley, 1991, Corollary 3.2.3) that this linear prevision is in fact the natural extension of $\vec{P}(n)$, i.e. $\vec{E}(n) = X \mapsto \frac{1}{n}\sum_{i=1}^{n} X\left(\vec{\Omega}(i)\right) = X \mapsto \overrightarrow{\Sigma X}(n)$. 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 $\vec{S} \colon \mathbb{N} \to \mathscr{T}$, where $\mathscr{T}$ is any topological space, if:

$$\forall N, \text{ where } N \text{ is any neighbourhood of } z \text{ with respect to } \mathscr{T},\ \forall n_0 \in \mathbb{N} \colon \exists n \geq n_0 \colon \vec{S}(n) \in N.$$

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 $\mathrm{CP}(\vec{E})$. Equivalently, by applying these linear previsions to indicator gambles, we obtain the set of finitely additive probabilities $\mathscr{P} := \left\{A \mapsto E(\chi_A) \colon E \in \mathrm{CP}(\vec{E})\right\}$. Due to the one-to-one relationship, we might work with either $\mathrm{CP}(\vec{E})$ or $\mathscr{P}$. Following Ivanenko, we call $\mathscr{P}$ the *statistical regularity* of the sequence $\vec{\Omega}$; in the language of imprecise probability, it is a *credal set*.

We further define

$$\overline{R}(X) := \sup\left\{E(X) \colon E \in \mathrm{CP}(\vec{E})\right\} = \sup\left\{E_P(X) \colon P \in \mathscr{P}\right\}, \quad X \in L^\infty,$$

where $E_P := \mathrm{NatExt}(P)$, and

$$\overline{P}(A) := \sup\left\{E(\chi_A) \colon E \in \mathrm{CP}(\vec{E})\right\} = \sup\left\{P(A) \colon P \in \mathscr{P}\right\}, \quad A \subseteq \Omega.$$

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 $P \in \mathscr{P}$ is a finitely additive probability on 2^Ω. We further observe that $\overline{R}$ is a coherent upper prevision on L^∞ or equivalently, a coherent risk measure in the sense of Artzner et al. (1999).[13] 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 (Walley, 1991, 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.

[12] This would hold under sequential compactness, which is not fulfilled here in general, but it is for finite Ω.
[13] For the close connection of coherent upper previsions and coherent risk measures we refer to (Fröhlich & Williamson, 2024).

Proposition 2.1. *The sequence of averages $\overrightarrow{\Sigma X}$ has the set of cluster points*

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

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

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

We use the following result from (Ivanenko & Pasichnichenko, 2017, Lemma 3).[14]

Lemma 2.2. *Let $f \colon Y \to \mathbb{R}$ be a continuous function on a compact space Y and $\overrightarrow{y}$ a Y-valued sequence. Then $\mathrm{CP}\left(n \mapsto f\left(\overrightarrow{y}(n)\right)\right) = f\left(\mathrm{CP}\left(\overrightarrow{y}\right)\right)$.*

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

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

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

A similar statement holds for the coherent upper probability.

Corollary 2.3. *For any $A \subseteq \Omega$, $\overline{P}(A) = \limsup_{n\to\infty}\left(\overrightarrow{P}(n)(A)\right) = \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\overrightarrow{\Omega}(i)\right)$.*

Proof. Just observe that $\overrightarrow{P}(n)(A) = \overrightarrow{\Sigma\chi_A}(n)$ 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 :

$$\underline{R}(X) = \inf\left\{E(X) \colon E \in \mathrm{CP}\left(\overrightarrow{E}\right)\right\} = \liminf_{n\to\infty} \overrightarrow{\Sigma X}(n), \quad \forall X \in L^{\infty},$$
$$\underline{P}(A) = \inf\{P(A) \colon P \in \mathscr{P}\} = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\overrightarrow{\Omega}(i)\right), \quad A \subseteq \Omega,$$

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.

Example 2.4. von Mises & Geiringer (1964, p. 11) considered a binary sequence ($\Omega = \{0,1\}$) given by

$$\overrightarrow{\Omega} = \left\langle 0^{[1]}, 1^{[1]}, 0^{[2]}, 1^{[2]}, 0^{[4]}, 1^{[4]}, \ldots 0^{[2^i]}, 1^{[2^i]}, \ldots \right\rangle,$$

for $i \to \infty$. The notation $i^{[j]}$ here means j repetitions of i and $\langle\cdot\rangle$ forms a finite or infinite sequence. It is a straightforward calculation to show that the induced relative frequencies of ones have all elements of $\left[\frac{1}{3}, \frac{1}{2}\right]$ as cluster points.

[14] A subtle point in the argument, which Ivanenko & Pasichnichenko (2017) 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.

Example 2.5. Let $\Omega = \{0,1,2,3\}$ and

$$\overrightarrow{\Omega} = \left\langle \langle 3,0\rangle^{[1]}, \langle 1,2\rangle^{[1]}, \langle 3,0\rangle^{[2]}, \langle 1,2\rangle^{[2]}, \langle 3,0\rangle^{[4]}, \langle 1,2\rangle^{[4]}, \ldots, \langle 3,0\rangle^{[2^i]}, \langle 1,2\rangle^{[2^i]}, \ldots \right\rangle,$$

for $i \to \infty$. Similarly, the notation $\langle a,b\rangle^{[j]}$ means j repetitions of the tuple $\langle a,b\rangle$, e.g. $\langle a,b\rangle^{[2]} = a,b,a,b$. It is easy to see that the events $\{0,2\}$ and $\{1,3\}$ each have the limiting relative frequency 0.5. But for the elementary events the cluster points of the relative frequencies are: $\left[\frac{1}{4},\frac{1}{3}\right]$ for $\omega = 3$ and $\omega = 0$; $\left[\frac{1}{6},\frac{1}{4}\right]$ for $\omega = 1$ and $\omega = 2$. This illustrates that the set system of events for which precise probabilities exist need not form a *field*, but in fact a *pre-Dynkin system* (see Derr & Williamson (2023)).

Example 2.6. The arising of finite additivity is not a feature of the divergence, but is due to the strictly frequentist setting. Consider for instance $\Omega = \mathbb{N}$, $\overrightarrow{\Omega}(i) = i$. Then the relative frequencies for each elementary converge to 0. Hence $\overline{P}(\{\omega\}) = \underline{P}(\{\omega\}) = 0\ \forall \omega \in \mathbb{N}$ and $\overline{P} = \underline{P}$ is a finitely additive probability, but $\overline{P}(\Omega) = 1$ in violation of countable additivity.

2.3 The Induced Upper Prevision

We have seen that the upper prevision $\overline{R}$, as we have just defined it, has the property that it is induced by the statistical regularity of the sequence, and at the same time corresponds to taking the supremum over the cluster points of the sequence of averages of a gamble over time. The set of cluster points is in general a complicated object, hence it is unclear why one should take the supremum to reduce it to a single number in a decision making context. Our goal in this section is to provide some intuition *why* it is reasonable to use $\overline{R}$, as we defined it, to assess the risk inherent in a sequence. Ivanenko & Munier (2000) argued that $\overline{R}$ is the unique object which satisfies a certain list of axioms, which are similar to those for an upper prevision, but including a so-called "principle of guaranteed result", which appears rather mysterious to us.

Our setup is as follows. We imagine an individual decision maker, who is faced with a fixed sequence $\overrightarrow{\Omega}\colon \mathbb{N} \to \Omega$ and various gambles $X\colon \Omega \to \mathbb{R}$. The question to the decision maker is how to value this gamble in light of the sequence, i.e. imagining that the gamble is evaluated at each $\overrightarrow{\Omega}(1), \overrightarrow{\Omega}(2), \ldots$, infinitely often. Here, $X(\overrightarrow{\Omega}(i))$ represents a loss for positive values, and a gain for negative values. We can think of the $\overrightarrow{\Omega}(i)$ as the states of nature, and the sequence determines which are realized and how often. Importantly, we view our decision maker as facing a *mass decision*, i.e. the gamble will not only be evaluated once, but instead infinitely often.

In the imprecise probability literature, a key concept for decision making is *desirability*. A coherent set of almost-desirable gambles is a set $\mathscr{D} \subset L^\infty$, satisfying the following properties (Walley, 1991, p. 152):

- **D1.** If $\inf(X) > 0$ then $X \notin \mathscr{D}$ (avoiding sure loss)
- **D2.** If $\sup(X) < 0$ then $X \in \mathscr{D}$ (accepting sure gains)
- **D3.** If $X \in \mathscr{D}$ and $\lambda > 0$, then $\lambda X \in \mathscr{D}$ (positive homogeneity)
- **D4.** If $X, Y \in \mathscr{D}$ then $X + Y \in \mathscr{D}$ (addition)
- **D5.** If $X - \varepsilon \in \mathscr{D}$ for any $\varepsilon > 0$, then $X \in \mathscr{D}$ (closure).

Almost-desirable gambles are those which the decision maker would accept even without any reward for it. We remark the existence of a subtle controversy about the role of 0 that distinguishes "almost" from "strict" desirability (Walley, 1991, Section 3.7), see also (Couso & Moral, 2011). In this paper, we focus on almost-desirability, and from now on drop writing "almost". Which gambles are *desirable* to our decision maker? We argue that an appropriate set of desirable gambles is given by:

$$\mathscr{D}_{\overrightarrow{\Omega}} := \left\{ X \in L^\infty \colon \limsup_{n\to\infty} \overrightarrow{\Sigma X} \le 0 \right\}.$$

It is easy to check that this satisfies D1–D5. Consider what $X \in \mathscr{D}_{\vec{\Omega}}$ means. If the limes superior of the gamble's average sequence, i.e. the growth rate of the accumulated loss, is negative or zero, then we are guaranteed that there is no strictly positive accumulated loss which we will face infinitely often. The choice of the average as the aggregation functional is justified from the mass decision character of the setting, since we assume that our decision maker does not care about individual outcomes, but only about long-run capital. Now, given this set of desirable gambles, we seek a functional $\overline{R}(X)\colon L^\infty \to \mathbb{R}$, so that when at each time step i, the transaction $X(\vec{\Omega}(i)) - \overline{R}(X)$ takes place, this results in a desirable gamble for our decision maker. Our decision maker shoulders the loss $X(\vec{\Omega}(i))$, while at the same time asking for $-\overline{R}(X)$ in advance. Intuitively, $\overline{R}(X)$ is supposed to be the certainty equivalent of the "uncertain" loss X, in the sense that $X(\vec{\Omega}(i))$ will vary over time. Therefore we define, in correspondence with $\mathscr{D}_{\vec{\Omega}}$, the upper and lower previsions ($\forall X \in L^\infty$):

$$
\begin{aligned}
\overline{R}(X) &:= \inf\left\{\alpha \in \mathbb{R}\colon\ X - \alpha \in \mathscr{D}_{\vec{\Omega}}\right\} \qquad (1)\\
\underline{R}(X) &:= -\overline{R}(-X) = \sup\left\{\alpha \in \mathbb{R}\colon\ \alpha - X \in \mathscr{D}_{\vec{\Omega}}\right\}.
\end{aligned}
$$

When a set of desirable gambles and an upper (lower) prevision are in this correspondence, it holds that $X - \overline{R}(X) \in \mathscr{D}_{\vec{\Omega}}$ and furthermore $\overline{R}(X)$ is the least (smallest) functional for which this holds. We can now observe that in fact $\overline{R}(X) = \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n)$, since

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

is the general correspondence for a set of desirable gambles and a coherent upper prevision. Hence, we have motivated the definition of $\overline{R}(X)$ in Section 2.2. It is easy to see explicitly (cf. Appendix A.2) that $X - \overline{R}(X) \in \mathscr{D}_{\vec{\Omega}}$ and that $\overline{R}(X) = \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n)$ is in fact the smallest number such that the relation in Equation 1 holds.

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 $\mathrm{CP}(\vec{E})$. In this section, 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 space Ω.

Theorem 3.1. *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:*

$$
\overline{R}(X) = \overline{R}_{\vec{\Omega}}(X) := \sup\left\{E(X)\colon E \in \mathscr{E}_{\vec{\Omega}}\right\}, \quad \mathscr{E}_{\vec{\Omega}} := \mathrm{CP}\left(\vec{E}_{\vec{\Omega}}\right) \quad \forall X \in L^\infty,
$$

where we now make the dependence on the sequence $\vec{\Omega}$ explicit in the notation, i.e. $\vec{E}_{\vec{\Omega}}(n) := X \mapsto \frac{1}{n}\sum_{i=1}^n X\left(\vec{\Omega}(i)\right)$.

The proof is constructive and deterministically yields a sequence. The significance of this result is that it establishes strictly frequentist *semantics* for imprecise probability. 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.

The key to prove this is Theorem 3.3, for which we introduce some convenient notation. For $k \in \mathbb{N}$, let $[k] := \{1, \ldots, k\}$ and define the $(k-1)$-simplex as

$$
\Delta^k := \left\{d = (d_1, \ldots, d_k) \in \mathbb{R}^k\colon\ \sum_{i=1}^n d_i = 1,\ d_i \ge 0\ \forall i \in [k]\right\}.
$$

It is also helpful to have a dual notation for sequences $\vec{x}: \mathbb{N} \to [k]$, whereby we write either $\vec{x}(i)$ or $\vec{x}_i$ to mean the same thing.

Definition 3.2. *Suppose $k \in \mathbb{N}$ and $\vec{x}: \mathbb{N} \to [k]$. For any $n \in \mathbb{N}$ define the* relative frequency of $\vec{x}$ with respect to i at n, $\vec{r}_i^{\vec{x}}: \mathbb{N} \to [0,1]$ *via*

$$\vec{r}_i^{\vec{x}}(n) := \tfrac{1}{n}|\{j \in [n] : \vec{x}_j = i\}|$$

and the relative frequency of $\vec{x}$ at n, $\vec{r}^{\vec{x}}: \mathbb{N} \to \Delta^k$ *as*

$$\vec{r}^{\vec{x}}(n) := \vec{r}^{\vec{x}}_{[k]}(n) = \begin{pmatrix} \vec{r}_1^{\vec{x}}(n) \\ \vdots \\ \vec{r}_k^{\vec{x}}(n) \end{pmatrix}. \tag{2}$$

Theorem 3.3. *Suppose $k \in \mathbb{N}$ and C is a rectifiable closed curve in Δ^k. There exists $\vec{x}: \mathbb{N} \to [k]$ such that $\mathrm{CP}(\vec{r}^{\vec{x}}) = C$.*

The proof (which is constructive) is in Appendix B along with an example. From this, we obtain the following Corollary (proven in Appendix B.10). Denote the topological boundary of a set D as ∂D.

Corollary 3.4. *Suppose $k \in \mathbb{N}$ and $D \subseteq \Delta^k$ is a non-empty convex set. There exists $\vec{x}: \mathbb{N} \to [k]$ such that $\mathrm{CP}(\vec{r}^{\vec{x}}) = \partial D$.*

Since we have a finite possibility space Ω, we can identify each linear prevision E with a point in the simplex, by assigning coordinates to its underlying finitely additive probability; in the case of $\vec{E}_{\vec{\Omega}}(n)$, this is the relative frequency $\vec{r}^{\vec{\Omega}}(n)$. This is formalized in the following.

Proposition 3.5. *Let $\vec{E}(n): \mathbb{N} \to \mathrm{PF}(\Omega)$ be a sequence of linear previsions with underlying probabilities $\vec{P}(n) := A \mapsto \vec{E}(n)(A)$. Then $E \in \mathrm{CP}\left(\vec{E}(n)\right)$ with respect to the weak* topology if and only if the sequence $\vec{D}: \mathbb{N} \to \Delta^k$, $\vec{D}(n) := \left(\vec{P}(n)(\omega_1), \ldots, \vec{P}(n)(\omega_k)\right)$ has as cluster point $d_E = \left(E\left(\chi_{\{\omega_1\}}\right), \ldots, E\left(\chi_{\{\omega_k\}}\right)\right)$ with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$.*

For the proof, see Appendix B.1. Here, the assumption of finite Ω is crucial; we leave it to future research to (if possible) extend Theorem 3.1 to infinite Ω. Combining Corollary 3.4 and Proposition 3.5 allows us to now prove Theorem 3.1.

Proof of Theorem 3.1. Let $\Omega = [k]$. If $\overline{R}$ is a coherent upper prevision on L^∞, we can write it as (Walley, 1991, Theorem 3.6.1):

$$\overline{R}(X) = \sup\{E(X) : E \in \mathscr{E}\}, \quad \forall X \in L^\infty,$$

for some weak* compact and convex set $\mathscr{E} \subseteq \mathrm{PF}(\Omega)$. From (Walley, 1991, Theorem 3.6.2) we further know that

$$\overline{R}(X) = \sup\{E(X) : E \in \mathscr{E}\} = \sup\{E(X) : E \in \operatorname{ext}\mathscr{E}\}, \quad \forall X \in L^\infty,$$

where ext denotes the set of extreme points of $\mathscr{E}$.[15] Then:

$$\begin{aligned} \overline{R}(X) &= \sup\{E(X) : E \in \mathscr{E}\} \\ &= \sup\{E(X) : E \in \operatorname{ext}\mathscr{E}\} \\ &\le \sup\{E(X) : E \in \partial\mathscr{E}\} \\ &\le \sup\{E(X) : E \in \mathscr{E}\} = \overline{R}(X), \end{aligned}$$

[15] A point $E \in \mathscr{E}$ is an extreme point of $\mathscr{E}$ if it cannot be written as a convex combination of any other elements in $\mathscr{E}$.

since $\text{ext}\mathscr{E} \subseteq \partial\mathscr{E}$ and $\partial\mathscr{E} \subseteq \mathscr{E}$; note that $\mathscr{E}$ is closed. In summary, $\overline{R}(X) = \sup\{E(X) \colon E \in \partial\mathscr{E}\}$.

Now choose $D := \left\{\left(E(\chi_{\omega_1}), \dots, E(\chi_{\omega_k})\right) : E \in \mathscr{E}\right\}$, which is a non-empty convex set in Δ^k. We then obtain from Corollary 3.4 a sequence $\vec{\Omega} \colon \mathbb{N} \to [k]$ with $\text{CP}(\vec{r}^{\,\vec{\Omega}}) = \partial D$. But then it follows from Proposition 3.5 that the sequence $\vec{E}_{\vec{\Omega}}$ has cluster points $\text{CP}\left(\vec{E}_{\vec{\Omega}}\right) = \partial\mathscr{E}$. Thus

$$\overline{R}(X) = \sup\left\{E(X) \colon E \in \text{CP}\left(\vec{E}_{\vec{\Omega}}\right)\right\}, \quad \forall X \in L^\infty,$$

which concludes the proof. □

Ivanenko (2010) offers a somewhat similar result to Theorem 3.1 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." (Ivanenko, 2010, Theorem 4.2). 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 (Walley & Fine, 1982) and Theorem 2.2 in (Papamarcou & Fine, 1991b). On the one hand, our result is stronger since it holds for upper previsions, whereas Theorem 4.2 in (Walley & Fine, 1982) and Theorem 2.2 in (Papamarcou & Fine, 1991b) hold for upper probabilities only; note that upper previsions are more expressive than upper probabilities.[16] On the other hand, Theorem 2.2 in (Papamarcou & Fine, 1991b) is stronger in the sense that it guarantees that the same upper probability is induced when applying selection rules.

We observe that two sequences $\vec{\Omega}_1, \vec{\Omega}_2$, might have different sets of cluster points $\text{CP}\left(\vec{E}_{\vec{\Omega}_1}\right), \text{CP}\left(\vec{E}_{\vec{\Omega}_2}\right)$, but when their convex hull coincides, the same upper probability and prevision is induced.[17] Thus, in light of the argument in Section 2.3, *for the purpose of mass decision making*, we may consider these sequences equivalent. While in the classical case, relative frequencies are the relevant description of a sequence, the statistical regularity provides an analogous description in the general case; moreover, we differentiate only "up to the same convex hull" for decision making.

4 Unstable Conditional Probability

An interesting aspect of the strictly frequentist approach is that there is a natural way of introducing conditional probability 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.

Recall that for a countably or finitely additive probability Q, we can define conditional probability as:

$$Q(A|B) := \frac{Q(A \cap B)}{Q(B)}, \quad A, B \subseteq \Omega, \text{ if } Q(B) > 0. \tag{3}$$

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.

[16] Indeed, the proof of Theorem 4.2 in (Walley & Fine, 1982) exploits this simplification by assuming that the credal set has a finite number of extreme points.

[17] Assume $\overline{R}(X) := \sup\{E(X) \colon E \in \mathscr{E}\}$. Then indeed $\overline{R}(X) = \sup\{E(X) \colon E \in \overline{\text{co}}\mathscr{E}\}$, where $\overline{\text{co}}$ denotes the weak* closure of the convex hull; cf. (Walley, 1991, Section 3.6).

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

$$E(X|B) := \frac{E(\chi_B X)}{E(\chi_B)}, \quad \text{if } E(\chi_B) > 0, B \subseteq \Omega. \tag{4}$$

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 (Walley, 1991, 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" (von Mises & Geiringer, 1964, p. 22). Walley & Fine (1982, 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:

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

where P_0 is an arbitrary finitely additive probability on 2^{Ω}, and $\bot$ represents "undefined".

4.1 Conditional Probability

Recall our sequence of unconditional finitely additive probabilities $\vec{P}(n) := A \mapsto \frac{1}{n}\sum_{i=1}^{n} \chi_A\left(\vec{\Omega}(i)\right)$. 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

$$\vec{P}(\cdot|B)(n) := \Psi\left(A \mapsto \frac{\sum_{i=1}^{n} (\chi_A \cdot \chi_B)\left(\vec{\Omega}(i)\right)}{\sum_{i=1}^{n} \chi_B\left(\vec{\Omega}(i)\right)}\right), \tag{5}$$

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 $|A \text{ and } B \text{ occured}| / |B \text{ occured}|$. Until B occurs for the first time, the denominator will be 0 and thus the mapping undefined (returning the falsum $\bot$). 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 $n \geq n_B$.

Proposition 4.1. *Assume $B \in 2^{\vec{\Omega}}_{1+}$. Then $\vec{P}(\cdot|B)$ is a sequence of finitely additive probabilities.*

Proof. For $n < n_B$, this is clear due to Ψ. Now let $n \geq n_B$.

PF1: $\vec{P}(\Omega|B)(n) = 1$: obvious.

PF2: If $A, C \subseteq \Omega$, $A \cap C = \emptyset$, then we show that $\overrightarrow{P}(A \cup C|B)(n) = \overrightarrow{P}(A|B)(n) + \overrightarrow{P}(C|B)(n)$.

$$\begin{aligned}\overrightarrow{P}(A \cup C|B)(n) &= \frac{\sum_{i=1}^{n} (\chi_{A\cup C} \cdot \chi_B)\left(\overrightarrow{\Omega}(i)\right)}{\sum_{i=1}^{n} \chi_B\left(\overrightarrow{\Omega}(i)\right)} \\ &= \frac{\sum_{i=1}^{n} (\chi_A \cdot \chi_B)\left(\overrightarrow{\Omega}(i)\right) + \sum_{i=1}^{n} (\chi_C \cdot \chi_B)\left(\overrightarrow{\Omega}(i)\right)}{\sum_{i=1}^{n} \chi_B\left(\overrightarrow{\Omega}(i)\right)} \\ &= \overrightarrow{P}(n)(A \cap B)/\overrightarrow{P}(n)(B) + \overrightarrow{P}(n)(C \cap B)/\overrightarrow{P}(n)(B) \\ &= \overrightarrow{P}(A|B)(n) + \overrightarrow{P}(C|B)(n).\end{aligned}$$

Noting that since A and C are disjoint, $\overrightarrow{\Omega}(i)$ cannot lie in both at the same time for any i. □

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 $\overrightarrow{P}(\cdot|B)(n)$, associate its uniquely corresponding linear prevision $\overrightarrow{E}(\cdot|B)(n)$, which is of course given by $(\forall X \in L^\infty,\ n \geq n_B)$:

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

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

Definition 4.2. *If $B \in 2^{\overrightarrow{\Omega}}_{1+}$, we define the conditional upper prevision and the conditional upper probability as:*

$$\overline{R}(X|B) := \sup\left\{\tilde{E}(X) \colon \tilde{E} \in \text{CP}\left(\overrightarrow{E}(\cdot|B)\right)\right\}; \quad \overline{P}(A|B) := \sup\left\{Q(A) \colon Q \in \text{CP}\left(\overrightarrow{P}(\cdot|B)\right)\right\}, \quad A \subseteq \Omega.$$

Since they are expressed via an envelope representation,[18] $\overline{R}$ and $\overline{P}$ are automatically coherent (Walley, 1991, Theorem 3.3.3). By similar reasoning as in Section 2.2, we get the following representation.

Proposition 4.3. *Assume $B \in 2^{\overrightarrow{\Omega}}_{1+}$. The conditional upper prevision (probability) can be represented as:*

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

Also, we obtain the corresponding lower quantities $\overline{R}(X|B) = \liminf_{n\to\infty} \overrightarrow{\Sigma X}|B(n)$ and $\underline{P}(A|B) = \liminf_{n\to\infty} \overrightarrow{P}(A|B)(n)$. Note that these definitions also have reasonable frequentist semantics even when B occurs only finitely often; then the sequence $\overrightarrow{P}(\cdot|B)$ is eventually constant and we have $\overrightarrow{P}(A|B) = |A \text{ and } B \text{ occured}|/|B \text{ occured}|$. For instance, if A and B occur just once, but simultaneously, then $\overline{P}(A|B) = \underline{P}(A|B) = 1$. 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. 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 4.4. *Assume $\overrightarrow{\Omega}$ is such that $P(B), P(A \cap B)$ exist for some $A, B \subseteq \Omega$ and $P(B) > 0$, that is, relative frequencies converge to those values. Then it holds that $\tilde{P}(A|B) = P(A|B)$, where $P(\cdot|B)$ is the conditional probability in the sense of Equation 3.*

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

Proof.

$$\begin{aligned}
\tilde{P}(A|B) &= \lim_{n\to\infty} \overrightarrow{P}(A|B)(n) \\
&= \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n}(\chi_A\cdot\chi_B)\left(\overrightarrow{\Omega}(i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\chi_B\left(\overrightarrow{\Omega}(i)\right)} \\
&\overset{(1)}{=} \frac{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\chi_A\cdot\chi_B)\left(\overrightarrow{\Omega}(i)\right)}{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\chi_B\left(\overrightarrow{\Omega}(i)\right)} \\
&= \frac{P(A\cap B)}{P(B)} \\
&\overset{(2)}{=} P(A|B).
\end{aligned}$$

(1) The limits exist by assumption and the denominator is > 0.
(2) In the sense of Equation 3.

□

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 4.4 and only demand that $\underline{P}(B) > 0$.[19] 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. (Miranda & Cooman, 2014)). The unconditional set of desirable gambles is:

$$\mathscr{D}_{\overrightarrow{\Omega}} := \left\{X \in L^\infty \colon \limsup_{n\to\infty} \overrightarrow{\Sigma X} \le 0\right\} = \left\{X \in L^\infty \colon \overline{R}(X) \le 0\right\}.$$

Definition 4.5. *For $\underline{P}(B) > 0$, we define the* conditional set of desirable gambles *as:*

$$\mathscr{D}_{\overrightarrow{\Omega}|B} := \{X \in L^\infty \colon X\chi_B \in \mathscr{D}_{\overrightarrow{\Omega}}\} = \left\{X \in L^\infty \colon \limsup_{n\to\infty} \overrightarrow{\Sigma(X\chi_B)} \le 0\right\},$$

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

$$\begin{aligned}
\mathrm{GBR}(X|B) &:= \inf\left\{\alpha \in \mathbb{R} \colon X - \alpha \in \mathscr{D}_{\overrightarrow{\Omega}|B}\right\} \\
&= \inf\{\alpha \in \mathbb{R} \colon \chi_B(X-\alpha) \in \mathscr{D}_{\overrightarrow{\Omega}}\} \\
&= \inf\left\{\alpha \in \mathbb{R} \colon \overline{R}\left(\chi_B(X-\alpha)\right) \le 0\right\}.
\end{aligned} \tag{6}$$

Remark 4.6. In fact, Walley (1991, 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 4.5,[20] see Appendix A.3.

Proposition 4.7. *Let $\underline{P}(B) > 0$. It holds that $\overline{R}(X|B) = \mathrm{GBR}(X|B)$.*

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

[20] The conditional set of desirable gambles is considered for instance in (Augustin et al., 2014) and (Wheeler, 2021), but there the link to the generalized Bayes rule is not made technically clear.

Proof. It is not hard to check that $\overline{R}(\cdot|B)$ is a coherent upper prevision on L^∞, hence we can represent it as (Walley, 1991, Theorem 3.8.1):

$$\overline{R}(X|B) = \inf\left\{\alpha \in \mathbb{R} \colon X - \alpha \in \mathscr{D}_{\overline{R}(\cdot|B)}\right\}, \quad \text{where } \mathscr{D}_{\overline{R}(\cdot|B)} := \left\{X \in L^\infty \colon \overline{R}(X|B) \le 0\right\}.$$

We show that $\overline{R}(X|B) = \text{GBR}(X|B)$ by showing that $\mathscr{D}_{\overline{R}(\cdot|B)} = \mathscr{D}_{\overrightarrow{\Omega}|B}$.

Let $X \in L^\infty$. On the one hand, we know

$$\begin{aligned} X \in \mathscr{D}_{\overrightarrow{\Omega}|B} \iff X\chi_B \in \mathscr{D}_{\overrightarrow{\Omega}} \iff \overline{R}\left(X\chi_B\right) \le 0 \\ \iff \limsup_{n\to\infty} \sum_{i=1}^{n} \frac{(X\chi_B)\left(\overrightarrow{\Omega}(i)\right)}{n} \le 0. \end{aligned} \tag{7}$$

On the other hand,

$$X \in \mathscr{D}_{\overline{R}(\cdot|B)} \iff \overline{R}(X|B) \le 0 \iff \limsup_{n\to\infty} \frac{\frac{1}{n}\sum_{i=1}^{n}(X\chi_B)\left(\overrightarrow{\Omega}(i)\right)}{\frac{1}{n}\sum_{i=1}^{n}\chi_B\left(\overrightarrow{\Omega}(i)\right)} \le 0. \tag{8}$$

It remains to show that the two limit statements (Equation 7 and Equation 8) are equivalent. Due to the limit operation we can neglect the terms $n = 1, \ldots, n_B - 1$. Furthermore, we know that $\overrightarrow{b}(n) \in (0,1]$, $n \ge n_B$, and also $0 < \liminf_{n\to\infty} \overrightarrow{b}(n)$. Thus, defining $\overrightarrow{a}(n) := \frac{1}{n}\sum_{i=1}^{n}(X\chi_B)\left(\overrightarrow{\Omega}(i)\right)$ and $\overrightarrow{b}(n) := \frac{1}{n}\sum_{i=1}^{n}\chi_B\left(\overrightarrow{\Omega}(i)\right)$, we can leverage Lemma A.2,[21] included in Appendix A.4 to show:

$$\limsup_{n\to\infty} \overrightarrow{a}(n) \le 0 \iff \limsup_{n\to\infty} \frac{\overrightarrow{a}(n)}{\overrightarrow{b}(n)} \le 0.$$

□

Remark 4.8. Note that $X \mapsto \limsup_{n\to\infty} \overrightarrow{\Sigma(X\chi_B)} = \overline{R}\left(X\chi_B\right)$ is *not* in general a coherent upper prevision on L^∞, as it can violate UP1; see Appendix A.5. In general, we have $\text{GBR}(X|B) \ne \overline{R}\left(X\chi_B\right)$.

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

Corollary 4.9. *If $\underline{P}(B) > 0$, the conditional upper prevision can be obtained by updating each linear prevision in the set of cluster points, that is:*

$$\overline{R}(X|B) = \sup\left\{E(X|B) \colon E \in \text{CP}\left(\overrightarrow{E}\right)\right\},$$

where conditioning of the linear previsions is in the sense of Definition 4.

This follows from (Walley, 1991, 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 (Durrett, 2019, p. 37), but more generally in most probability theories (Levin (1980); Fine (1973, Sections IIF, IIIG and VH)). In Appendix C 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.

[21] To rigorously apply the Lemma, we would again introduce a wrapper for the sequence $\overrightarrow{b}(n)$ to ensure strict positivity, since finitely many terms $i = 1, \ldots, n_B - 1$ might be zero.

5 Related Work

In this section we examine previous research at the intersection of frequentism and imprecise probability. 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 (1970) was one of the first authors to critically evaluate the hypothesis of statistical stability. Fine (1970) observed that this widespread hypothesis is regarded as a "striking instance of order in chaos" in the statistics community, and sought to challenge its nature as an empirical "fact". In contrast to our approach, Fine (1970) was concerned with finite sequences and the question what it means for such a sequence to be random. While Fine did mention von Mises, Fine (1970) opted for a randomness definition based on computational complexity. Intuitively, one can consider a sequence random if it cannot be generated by a short computer program (i.e. Turing machine). Fine then showed that statistical stability ("apparent convergence") occurs *because of*, and not in spite of, high randomness of the sequence. In contrast, a sequence for which relative frequencies diverge has low computational complexity. We consider these findings surprising, and believe that an interesting avenue for future research with respect to statistical stability lies in the comparison of the computational complexity approach to von Mises randomness notion based on selection rules. We agree with Fine (1970) that apparent convergence is not some law of nature, but rather a consequence of data handling.

The previously mentioned paper may be seen as a predecessor to a long line of work by Terrence Fine and collaborators, (Fine, 1976; Walley & Fine, 1982; Kumar & Fine, 1985; Grize & Fine, 1987; Fine, 1988; Papamarcou & Fine, 1991a;b; Sadrolhefazi & Fine, 1994; Fierens et al., 2009); see also (Fine, 2016) 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 Section 4.3 in (Walley & Fine, 1982), (Papamarcou & Fine, 1991b) and (Fierens et al., 2009)) use an imprecise probability to represent a single trial in a sequence of unlinked repetitions of an experiment, and then induce an imprecise probability via an infinite product space. This is in the spirit of, and 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 of random variables; here, there is both an "individual P," as well as an induced "aggregate P" on the infinite product space, which can be used to measure an event such as convergence or divergence of relative frequencies.

When a single trial is assumed to be governed by an imprecise probability, how can this be interpreted? And what is the interpretation of the mysterious "aggregate imprecise probability"? This model falls prey to similar criticisms as we outlined in the Introduction (Section 1) concerning the theoretical law of large numbers. In fact, Walley & Fine (1982) subscribed to a frequency-propensity interpretation (specifically, they were inspired by Giere (1973)), where the imprecise probability of a single trial represents its propensity, that is, its tendency or disposition to produce a certain outcome. Consequently, one obtains a propensity for compound trials in terms of an imprecise probability and thus one can ascribe a lower and upper probability to events such as divergence of relative frequencies. To us, the meaning of such a propensity is unclear. While we are not against a propensity interpretation 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.

Closely related to our work is (Papamarcou & Fine, 1991b), who were also inspired by von Mises. The authors proved that, for any set of probability measures $\mathscr{P}$ on $(\Omega, 2^{\Omega})$, $|\Omega| < \infty$, and any countable set of place selection rules $\mathscr{S}$,[22] the existence of a sequence $\overrightarrow{\Omega}: \mathbb{N} \to \Omega$ with the following property can be guaranteed (Papamarcou & Fine, 1991b, Theorem 2.2):

$$\forall \overrightarrow{S}_j \in \mathscr{S}: \forall A \subseteq \Omega: \limsup_{n\to\infty} \frac{\sum_{i=1}^{n} \chi_A\left(\overrightarrow{\Omega}(i)\right) \cdot \overrightarrow{S}_j(i)}{\sum_{i=1}^{n} \overrightarrow{S}_j(i)} = \sup\{P(A) : P \in \mathscr{P}\}.$$

That is, the sequence has the specified upper probability (take $\overrightarrow{S}_j(i) = 1\ \forall i \in \mathbb{N}$) and this property is stable under subselection. Note that this claim is in one sense weaker than our Proposition 3.1, where we construct a sequence for which the set of cluster points is exactly a prespecified one — coherent upper previsions are more expressive than coherent upper probabilities; on the other hand it is stronger, since the property holds also when applying selection rules.

Within the setup of (Walley & Fine, 1982), Cozman & Chrisman (1997, Theorem 1, Theorem 2) proposed an estimator for the underlying imprecise probability of the sequence. Specifically, they computed relative frequencies along a set of selection rules (however without referring to von Mises) and then took their minimum to obtain a lower probability; in a specific technical sense, this estimation succeeds. 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. This viewpoint also motivated Fierens et al. (2009), who restricted themselves to finite sequences. 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.

Cattaneo (2017) investigated an empirical, frequentist interpretation of imprecise probability in a similar setting, where $X_1, X_2, ..$ is a sequence of precise Bernoulli random variables, but $p_i := P(X_i = 1)$ is chosen by nature and may differ from trial to trial, hence $p_i \in [\underline{p_i}, \overline{p_i}]$. The author drew the sobering conclusion that "imprecise probabilities do not have a generally valid, clear empirical meaning, in the sense discussed in this paper".

Works which proposed extensions (modifications) of the law of large numbers to imprecise probabilities include (Marinacci, 1999; Maccheroni & Marinacci, 2005; De Cooman & Miranda, 2008; Chen et al., 2013; Peng, 2019).

In order to access a more powerful toolbox, De Cooman & De Bock (2022) and Persiau et al. (2022) more recently studied the interplay of imprecise probability and randomness in the game-theoretic setup with references to a frequentist perspective, see e.g. (De Cooman & De Bock, 2022, 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.

Separate from the imprecise probability literature, Gorban (2017) studied the phenomenon of statistical stability and its violations in depth, including theory and experimental studies. Similarly, the work of Ivanenko (2010), a major motivation for our work, does not appear to be known in the imprecise probability literature.

6 Conclusion

In this work, we have extended strict frequentism to the case of possibly divergent relative frequencies and sample averages, tying together threads from (von Mises, 1919), (Ivanenko, 2010) and (Walley, 1991).

[22] See (Papamarcou & Fine, 1991b) for the definition. Intuitively, a place selection rule is causal, i.e. depends only on past values.

In particular, we have recovered the generalized Bayes rule from a strictly frequentist perspective. Furthermore, we have established strictly frequentist *semantics* for imprecise probability, by demonstrating that (under the mild assumption that $|\Omega| < \infty$) we can explicitly construct a sequence for which the relative frequencies have a prespecified set of cluster points, corresponding to the coherent upper prevision. Previous results only covered the more restrictive case of upper probabilities (Walley & Fine, 1982; Papamarcou & Fine, 1991b).

Statisticians exclusively assume that their data is part of a stable sequence, but the hypothesis of perfect statistical stability is just a *hypothesis*; see Appendix D for an elaboration of this point. Importantly, 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 amounts to instability of relative frequencies even for long observation intervals (Gorban, 2017). In this work, we have rejected the assumption of 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. (Cattaneo, 2017)). 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 & Fine (1991b) 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.[23] 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. As a consequence, we would obtain potentially multiple ways of bridging the finite and the infinite case, which enables practical applications[24].

Finally, we remark that an interesting avenue for future research may investigate the use of *nets*, which generalize the concept of a sequence. Indeed, *fraction-of-time probability* (Gardner, 1986; Leśkow & Napolitano, 2006; Napolitano & Gardner, 2022; Gardner, 2022) is a theory of probability with remarkable parallels to von Mises' (1919). Instead of sequences, this theory is based on continuous time, hence a net $\vec{\Omega}\colon \mathbb{R} \to \Omega$.[25] Sample averages are then given by integration instead of summation. In essence, this amounts to using a different *relative measure* than the counting measure, which is implicit in the work of von Mises (1919). However, fraction-of-time probability was so far developed only for the convergent case; we expect that a similar construction as in Section 2.2 could be used to extend it to the case of divergence.

[23] As demonstrated by Examples 2 and 3 in (Cozman & Chrisman, 1997), converging relative frequencies within selection rules can lead to both overall convergence or divergence (on the whole sequence).

[24] Note that precise probability in the framework of von Mises also faces this bridging problem, which is a special case of ours; a point explicitly recognised by Kolmogorov long ago: "Here I insist on the view, expressed by Mr. von Mises himself (von Mises, 1931, pp. 21–26) that 'collectives' are finite (though very large) in real practice" (Kolmogorov, 1939). Kolmogorov reiterated von Mises' view that the infinite sample case is indeed a "mathematical idealization" of the practical situation. Approximately a century of effort has produced some reasonable refinements of that idealization in the finite *stable* case. We hope it will not take quite that long in the finite *unstable* case, but recognise the many difficulties that will need addressing.

[25] We remark that the notion of a sampling net in (Ivanenko, 2010) is a different one, that is, fraction-of-time probability does not fit this concept, although it is based on the usage of a net.

Acknowledgments

Robert Williamson thanks Jingni Yang for a series of discussions over several years about Ivanenko's work which provided much inspiration for the present work. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — EXC number 2064/1 — Project number 390727645. Thanks to the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christian Fröhlich and Rabanus Derr.

Contributions

C. Fröhlich is responsible for most of the content and writing. The comparison of Ivanenko's and Walley's theories was suggested by R. C. Williamson, who supervised the project and provided large parts of the constructive existence proof of arbitrary diverging frequency sequences. R. Derr minorly contributed through discussions and helpful suggestions; he wants to emphasize the effort which C. Fröhlich and R. C. Williamson put into this project.

References

Sergio Albeverio, Mykola Pratsiovytyi, and Grygoriy Torbin. Topological and fractal properties of real numbers which are not normal. *Bulletin des Sciences Mathématiques*, 129(8):615–630, 2005.

Sergio Albeverio, Iryna Garko, Muslem Ibragim, and Grygoriy Torbin. Non-normal numbers: Full Hausdorff dimensionality vs zero dimensionality. *Bulletin des Sciences Mathématiques*, 141(2):1–19, 2017.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Miroslav D. Ašić and Dušan D. Adamović. Limit points of sequences in metric spaces. *The American Mathematical Monthly*, 77(6):613–616, 1970.

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

Andrea Aveni and Paolo Leonetti. Most numbers are not normal. In *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 1–11. Cambridge University Press, 2022.

Ralph Alger Bagnold. The nature and correlation of random distributions. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 388(1795):273–291, 1983.

Heinz H. Bauschke, Minh N. Dao, and Walaa M. Moursi. On Fejér monotone sequences and nonexpansive mappings. *Linear and Nonlinear Analysis*, 1(2):287–295, 2015.

Laurent Bienvenu, Glenn Shafer, and Alexander Shen. On the history of martingales in the study of randomness. *Electronic Journal for History of Probability and Statistics*, 5(1):1–40, 2009.

Christopher J. Bishop and Yuval Peres. *Fractals in probability and analysis*. Cambridge University Press, 2017.

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

Efim M. Bronshteyn and L.D. Ivanov. The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853, 1975.

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

Cristian S. Calude. *Information and randomness: an algorithmic perspective*. Springer Science & Business Media, 2002.

Cristian S. Calude, Solomon Marcus, and Ludwig Staiger. A topological characterization of random sequences. *Information Processing Letters*, 88(5):245–250, 2003.

Georges Canguilhem. *On the Normal and the Pathological*. D. Riedel, 1978.

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

Ping Chen. Nonequilibrium and nonlinearity — a bridge between the two cultures. In George P. Scott (ed.), *Time, rhythms, and chaos in the new dialogue with nature*, pp. 67–85. Iowa State University Press, 1991.

Zengjing Chen, Panyu Wu, and Baoming Li. A strong law of large numbers for non-additive probabilities. *International Journal of Approximate Reasoning*, 54(3):365–377, 2013.

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

Jessi Cisewski, Joseph B. Kadane, Mark J. Schervish, Teddy Seidenfeld, and Rafael Stern. Standards for modest Bayesian credences. *Philosophy of Science*, 85(1):53–78, 2018.

Inés Couso and Serafín Moral. Sets of desirable gambles: conditioning, representation, and precise probabilities. *International Journal of Approximate Reasoning*, 52(7):1034–1055, 2011.

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

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

Gert De Cooman and Enrique Miranda. Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*, 138(8):2409–2432, 2008.

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

Abraham de Moivre. *The Doctrine of Chances: A Method of Calculating the Probabilities of Events in Play*. Frank Cass and Company Limited, 2nd edition, 1738/1967.

Anton Deitmar. *Analysis*. Springer, 2016.

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

Rabanus Derr and Robert C. Williamson. Systems of precision: Coherent probabilities on pre-Dynkin systems and coherent previsions on linear subspaces. *Entropy*, 25(9), 2023.

Alain Desrosières. *The politics of large numbers: A history of statistical reasoning.* Harvard University Press, 1998.

Nelson Dunford and Jacob T. Schwartz. *Linear operators, part 1: general theory.* John Wiley & Sons, 1988.

Detlef Dürr. Bohmian mechanics. In Jean Bricmont, Detlef Dürr, Maria Carla Galavotti, Giancarlo Ghirardi, Francesco Petruccione, and Nino Zanghi (eds.), *Chance in Physics: Foundations and Perspectives*, pp. 115–131. Springer, 2001.

Detlef Dürr and Ward Struyve. Typicality in the foundations of statistical physics and Born's rule. *Do Wave Functions Jump? Perspectives of the Work of GianCarlo Ghirardi*, pp. 35–43, 2021.

Detlef Dürr and Stefan Teufel. *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer, 2009.

Detlef Dürr, Sheldon Goldstein, and Nino Zanghi. Quantum equilibrium and the origin of absolute uncertainty. *Journal of Statistical Physics*, 67:843–907, 1992.

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

Harold Gordon Eggleston. The fractional dimension of a set defined by decimal properties. *The Quarterly Journal of Mathematics*, 20:31–36, 1949.

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.

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

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

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 (eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Volume 1*, pp. 29–40. Springer, 1976.

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

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

Christian Fröhlich and Robert C. Williamson. Risk measures and upper probabilities: Coherence and stratification. *Journal of Machine Learning Research (to appear)*, 2024.

Francis Galton. *Natural Inheritance.* Macmillan and Company, 1889.

Bruno Galvan. Bohmian mechanics and typicality without probability. *arXiv preprint quant-ph/0605162*, 2006.

Bruno Galvan. Typicality vs. probability in trajectory-based formulations of quantum mechanics. *Foundations of Physics*, 37(11):1540–1562, 2007.

Bruno Galvan. Non-probabilistic typicality, with application to quantum mechanics. *arXiv preprint arXiv:2209.14985*, 2022.

William A. Gardner. *Statistical Spectral Analysis: A Nonprobabilistic Theory.* Prentice-Hall, Inc., USA, 1986.

William A. Gardner. Transitioning away from stochastic process models, 2022. URL https://cyclostationarity.com/wp-content/uploads/2022/07/Transitioning-Away-from-Stochastic-Process-Models.pdf. Accessed: 2023-01-27.

Ronald N. Giere. Objective single-case probabilities and the foundations of statistics. In *Studies in Logic and the Foundations of Mathematics*, volume 74, pp. 467–483. Elsevier, 1973.

Sheldon Goldstein. Typicality and notions of probability in physics. In *Probability in physics*, pp. 59–71. Springer, 2012.

Igor I. Gorban. Statistical instability of physical processes. *Radioelectronics and Communications Systems*, 54(9):499, 2011.

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

Igor I. Gorban. *Randomness and Hyper-randomness.* Springer, 2018.

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.

Xiaoyang Gu and Jack H. Lutz. Effective dimensions and relative frequencies. *Theoretical computer science*, 412(48):6696–6711, 2011.

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

Yair M. Guttmann. *The concept of probability in statistical physics.* Cambridge University Press, 1999.

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

Richard Holmes. *Geometric functional analysis and its applications.* Springer, 1975.

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

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 (eds.), *Stochastic Optimization*, pp. 388–392. Springer, 1986a.

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

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

Victor I. Ivanenko and Valery A. Labkovskii. On the studying of mass phenonmena. In *Proceedings of the 12th IFAC Triennial World Congress, Sydney, Australia*, pp. 693–696, 1993.

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

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

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

Konstantinos V. Katsikopoulos, Özgür Şimşek, Marcus Buckmann, and Gerd Gigerenzer. *Classification in the wild: The science and art of transparent decision making.* MIT Press, 2021.

Andrei Y. Khrennikov. *Interpretations of Probability – Second revised and extended edition.* Walter de Gruyter, 2009.

Andrei Y. Khrennikov. *p-Adic valued distributions in mathematical physics.* Springer, 2013.

Andrei Y. Khrennikov. *Probability and randomness: quantum versus classical.* Imperial College Press, 2016.

Mervyn King and John Kay. *Radical uncertainty: Decision-making for an unknowable future.* Hachette UK, 2020.

Andreĭ Nikolaevich Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, 1933.

Andreĭ Nikolaevich Kolmogorov. Letter to Maurice Fréchet. 3 August 1939; reproduced in (Bienvenu et al., 2009), 1939.

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

Andreĭ Nikolaevich Kolmogorov. On logical foundations of probability theory. In K. Ito and J.V. Prokhorov (eds.), *Probability theory and mathematical statistics: Proceedings of the Fourth USSR – Japan Symposium, held at Tbilisi, USSR, August 23-29, 1982*, pp. 1–5. Springer, 1983.

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

Adam La Caze. Frequentism. In Alan Hájek and Christopher Hitchcock (eds.), *The Oxford Handbook of Probability and Philosophy*. Oxford University Press, 2016.

J. Dennis Lawrence. *A Catalog of Special Plane Curves.* Dover, 1972.

Johannes Lenhard and Wolfgang Krohn. The law of large numbers. edgar zilsel's attempt at the foundation of physical and socio-historical laws. In *Edgar Zilsel: Philosopher, Historian, Sociologist*, pp. 121–142. Springer, 2022.

Jacek Leśkow and Antonio Napolitano. Foundations of the functional approach for signal analysis. *Signal processing*, 86(12):3796–3825, 2006.

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.

Edward H. Lockwood. *A Book of Curves.* Cambridge University Press, 1961.

Fabio Maccheroni and Massimo Marinacci. A strong law of large numbers for capacities. *The Annals of Probability*, 33(3):1171–1178, 2005.

Donald A. MacKenzie, Fabian Muniesa, and Lucia Siu (eds.). *Do economists make markets?: On the performativity of economics*. Princeton University Press, 2007.

Massimo Marinacci. Limit laws for non-additive probabilities and their frequentist interpretation. *Journal of Economic Theory*, 84(2):145–195, 1999.

Celestino G. Méndez. On the law of large numbers, infinite games, and category. *The American Mathematical Monthly*, 88(1):40–42, 1981.

Mary Midgley. *Science as salvation: A modern myth and its meaning*. Routledge, 2013.

John R. Milton. The origin and development of the concept of the 'laws of nature'. *European Journal of Sociology/Archives Européennes de Sociologie*, 22(2):173–195, 1981.

Enrique Miranda. A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*, 48(2):628–658, 2008.

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

Gabriel Nagy. Topological vector spaces III: Finite dimensional spaces – Notes from the Functional Analysis Course (Fall 07 - Spring 08), 2007. URL https://www.math.ksu.edu/~nagy/func-an-2007-2008/top-vs-3.pdf. Accessed: 2023-01-13.

Antonio Napolitano and William A. Gardner. Fraction-of-time probability: Advancing beyond the need for stationarity and ergodicity assumptions. *IEEE Access*, 10:34591–34612, 2022.

Grégoire Nicolis and Ilya Prigogine. *Self-organization in nonequilibrium systems*. John Wiley & Sons, 1977.

Lars Olsen. Extremely non-normal numbers. *Mathematical Proceedings of the Cambridge Philosophical Society*, 137(1):43–53, 2004.

John C. Oxtoby. *Measure and category: A survey of the analogies between topological and measure spaces (2nd edition)*. Springer, 1980.

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, 1991a.

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

Shige Peng. *Nonlinear expectations and stochastic calculus under uncertainty: with robust CLT and G-Brownian motion*. Springer Nature, 2019.

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.

Graeme M. Philip and D.F. Watson. Some speculations on the randomness of nature. *Mathematical Geology*, 19(6):571–573, 1987.

Alois Pichler. The natural Banach space for version independent risk measures. *Insurance: Mathematics and Economics*, 53(2):405–415, 2013.

Itamar Pitowsky. Typicality and the role of the Lebesgue measure in statistical mechanics. In *Probability in physics*, pp. 41–58. Springer, 2012.

Ilya Prigogine. Time, structure, and fluctuations (Nobel prize lecture). *Science*, 201(4358):777–785, 1978.

Ilya Prigogine. *From Being to Becoming: Time and Complexity in the Physical Sciences*. W.H. Freeman and Company, 1980.

Ilya Prigogine and Isabelle Stengers. *Order out of chaos: Man's new dialogue with nature*. Bantam Books, 1985.

K.P.S. Bhaskara Rao and M. Bhaskara Rao. *Theory of charges: a study of finitely additive measures*. Academic Press, 1983.

Ángel Rivas. On the role of joint probability distributions of incompatible observables in Bell and Kochen–Specker theorems. *Annals of Physics*, 411:167939, 2019.

Todd Rose. *The end of average: How to succeed in a world that values sameness*. Penguin UK, 2016.

Jane E. Ruby. The origins of scientific"law". *Journal of the History of Ideas*, 47(3):341–359, 1986.

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.

Eric Schechter. *Handbook of Analysis and its Foundations*. Academic Press, 1997.

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

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. In *Presented at Pitt Workshop 3/17*, 2017.

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

Lawrence Sklar. *Physics and chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press, 1995.

Lawrence Sklar. Topology versus measure in statistical mechanics. *The Monist*, 83(2):258–273, 2000.

George Soros. General theory of reflexivity. *The Financial Times*, 27 October 2009.

Anastasios Stylianou. A typical number is extremely non-normal. *arXiv preprint arXiv:2006.02202*, 2020.

Terence Tao. 245B, notes 11: The strong and weak topologies. https://terrytao.wordpress.com/2009/02/21/245b-notes-11-the-strong-and-weak-topologies/, 2009. Accessed: 2023-01-13.

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

Elart Von Collani. A note on the concept of independence. *Economic Quality Control*, 21(1):155–164, 2006.

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

Richard von Mises. *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und Theoretischen Physik*. Deuticke, 1931.

Richard von Mises. *Probability, Statistics, and Truth*. Dover, 1981.

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

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

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

Friedel Weinert (ed.). *Laws of nature: Essays on the philosophical, scientific and historical dimensions*, volume 8. Walter de Gruyter, 1995.

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

Edgar Zilsel. *Das Anwendungsproblem. Ein philosophischer Versuch über das Gesetz der großen Zahlen und die Induktion. Leipzig*. Barth, 1916.

Edgar Zilsel. The genesis of the concept of physical law. *The Philosophical Review*, 51(3):245–279, 1942.

A Proofs

A.1 Weak* compactness of PF(Ω)

Walley (1991, Appendix D4) states that the set of linear previsions, PF(Ω), is compact due to the Alaoglu-Bourbaki theorem, but does not explain how this follows. For completeness, we provide an argument. First, we can observe, like Walley (1991), that the set is weak* closed. We use the following well known Lemma, see e.g. (Deitmar, 2016, Lemma 12.3.4.).

Lemma A.1. *Let $\mathscr{X}$ be a topological space, $K \subseteq \mathscr{X}$ be compact and $L \subseteq K$ be closed. Then L is compact.*

Thus we will show that PF(Ω) is a subset of some weak* compact set in $(L^\infty)^*$, hence it is in fact weak* compact. From the Alaoglu-Bourbaki theorem (see e.g. (Holmes, 1975, p. 70)), we know that the unit ball of the dual norm is weak* compact in $(L^\infty)^*$. By definition of the dual norm $\|\cdot\|^*$ of the L^∞ norm $\|X\| := \sup_{\omega\in\Omega}\{|X(\omega)|\}$, this is the following set:

$$
\begin{aligned}
B &:= \left\{X^* \in (L^\infty)^* : \|X^*\|^* \leq 1\right\} \\
&= \left\{X^* \in (L^\infty)^* : \sup\left\{|X^*(X)| : X \in L^\infty \text{ for which } \|X\| \leq 1\right\} \leq 1\right\} \\
&= \left\{X^* \in (L^\infty)^* : \sup\left\{|X^*(X)| : X \in L^\infty \text{ for which } \sup_{\omega\in\Omega}\{|X(\omega)|\} \leq 1\right\} \leq 1\right\}.
\end{aligned}
$$

Thus, to show that PF(Ω) is weak* compact, it suffices to show that PF(Ω) ⊆ B. That is, given some $E \in$ PF(Ω), we show that if $X \in L^\infty$ is such that $\sup_{\omega\in\Omega}\{|X(\omega)|\} \leq 1$, then $|E(X)| \leq 1$, since then the supremum over all such X is also ≤ 1. Thus it suffices to show that $|E(X)| \leq \sup_{\omega\in\Omega}\{|X(\omega)|\}$. We know that $E(|X|) \leq \sup_{\omega\in\Omega}\{|X(\omega)|\}$, since E is a coherent upper prevision, cf. Walley (1991, 2.6.1a). But we also have that $|E(X)| \leq E(|X|)$ from monotonicity and $E(0) = 0$, see e.g. Pichler (2013, Proposition 5), and thus $|E(X)| \leq E(|X|) \leq \sup_{\omega\in\Omega}\{|X(\omega)|\} \leq 1$, which concludes the proof.

A.2 Properties of the Induced Upper Prevision

To see that $X - \overline{R}(X) \in \mathscr{D}_{\vec{\Omega}}$:

$$\begin{aligned} X - \overline{R}(X) \in \mathscr{D}_{\vec{\Omega}} &\iff \lim_{n\to\infty} \sup \frac{1}{n} \sum_{i=1}^{n} \left((X - \overline{R}(X)) \left(\vec{\Omega}(i) \right) \right) \leq 0 \\ &\iff \limsup_{n\to\infty} \overrightarrow{\Sigma X} - \overline{R}(X) \leq 0, \quad \text{since } \overline{R}(X) \text{ is constant} \\ &\iff \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n) - \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n) = 0 \leq 0. \end{aligned}$$

Now we show that $\overline{R}(X) = \limsup_{n\to\infty} \overrightarrow{\Sigma X}(n)$ is in fact the smallest number such that the relation in Equation 1 holds. Suppose there exists $\varepsilon > 0$ such that $\overline{R}(X) - \varepsilon$ (with $\overline{R}$ defined as before) makes $X - (\overline{R}(X) - \varepsilon)$ desirable, that is

$$\limsup_{n\to\infty} \overrightarrow{\Sigma X} - \overline{R}(X) + \varepsilon \leq 0,$$

which is a contradiction due to our choice of $\overline{R}$.

A.3 Proof of Remark 4.6

In the literature, the generalized Bayes rule is defined as the solution α^* of $\overline{R}\left(\chi_B(X-\alpha)\right) = 0$. We show that $\alpha^* = \inf\left\{\alpha \in \mathbb{R} \colon \overline{R}\left(\chi_B(X-\alpha)\right) \leq 0\right\}$. Of course, we get for $\alpha := \alpha^*$ that equality holds (= 0). We just have to exclude the possibility that there exists $\tilde{\alpha} < \alpha^*$ so that $\overline{R}\left(\chi_B(X-\tilde{\alpha})\right) \leq 0$.

Assume such an $\tilde{\alpha}$ exists, so $\overline{R}\left(\chi_B X - \chi_B \tilde{\alpha}\right) \leq 0$. Write $\tilde{\alpha} + \varepsilon \leq \alpha^*$ for some $\varepsilon > 0$. Since $\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon \geq \chi_B X - \chi_B \alpha^*$, it follows from monotonicity that $\overline{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) \geq \overline{R}(\chi_B X - \chi_B \alpha^*) = 0$. But from subadditivity, $\overline{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) \leq \overline{R}(\chi_B X - \chi_B \tilde{\alpha}) + \overline{R}(-\chi_B \varepsilon)$ and since $\varepsilon \overline{R}(-\chi_B) < 0$ due to coherence and $\underline{P}(B) > 0$, we have $\overline{R}(\chi_B X - \chi_B \tilde{\alpha} - \chi_B \varepsilon) < \overline{R}(\chi_B X - \chi_B \tilde{\alpha})$. Taking this together, we obtain $\overline{R}(\chi_B X - \chi_B \tilde{\alpha}) > 0$, a contradiction to the assumption.

Thus we have shown that $\alpha^* = \inf\left\{\alpha \in \mathbb{R} \colon \overline{R}\left(\chi_B(X-\alpha)\right) \leq 0\right\}$. The other expressions in Definition 4.5 follow by simple manipulations.

A.4 Supplement for Proof of Proposition 4.7

Lemma A.2. *Let $\vec{a} \colon \mathbb{N} \to \mathbb{R}$ be a sequence and $\vec{b} \colon \mathbb{N} \to (0,1]$ be a nonnegative sequence such that $\liminf_{n\to\infty} \vec{b}(n) > 0$. Then:*

$$\limsup_{n\to\infty} \vec{a}(n) \leq 0 \iff \limsup_{n\to\infty} \frac{\vec{a}(n)}{\vec{b}(n)} \leq 0.$$

Proof. For brevity we simply write a_n and b_n for the sequences $\vec{a}(n)$ and $\vec{b}(n)$.

$$\begin{aligned} \limsup_{n\to\infty} a_n \leq 0 &\iff \limsup_{n\to\infty} \frac{a_n}{b_n} \leq 0 \\ \lim_{n\to\infty} \left(\sup_{k\geq n} a_k \right) \leq 0 &\iff \lim_{n\to\infty} \left(\sup_{k\geq n} \frac{a_k}{b_k} \right) \leq 0. \end{aligned}$$

where we know that $b_n \in (0,1]$ and furthermore $0 < \liminf_{n\to\infty} b_n \leq \limsup_{n\to\infty} b_n$. If the sequence b_n would actually converge, then the statement is clearly true, since we can then pull out the limit of b_n (this is allowed).

We begin by showing that $LHS \le 0 \implies RHS \le 0$. Our assumption is that

$$\forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : \forall n \ge n_0 : \sup_{k \ge n} a_k < \varepsilon. \tag{9}$$

Our aim is to show that

$$\forall \varepsilon' > 0 : \exists n_0' \in \mathbb{N} : \forall n \ge n_0' : \sup_{k \ge n} \frac{a_k}{b_k} < \varepsilon'.$$

So let some $\varepsilon' > 0$ be given and fixed. We have to exhibit some n_0' such that the above statement holds. Choose $\varepsilon := \varepsilon' \cdot \lim_{n\to\infty} \inf_{k \ge n} b_k \cdot \frac{1}{\kappa}$, for an arbitrary $\kappa > 1$. Then $\varepsilon > 0$ by our assumption that $\lim_{n\to\infty} \inf_{k\ge n} b_k > 0$, i.e. that $\underline{P}(B) > 0$. Note that $\varepsilon \le \varepsilon'$. Then, we know that $\exists n_0(\varepsilon)$ such that $\forall n \ge n_0(\varepsilon)$ $\sup_{k \ge n} a_k < \varepsilon$.

Also, we know that $\forall \kappa' > 1 : \exists n_0'' \in \mathbb{N} : \forall n \ge n_0''$:

$$\frac{\lim_{n\to\infty} \inf_{k\ge n} b_k}{\inf_{k \ge n} b_k} \le \kappa'. \tag{10}$$

The numerator is the limit of the denominator (which exists) and furthermore $\inf_{k\ge n} b_k$ is monotone increasing in n, that is, $\forall n \in \mathbb{N}$: $\lim_{n\to\infty} \inf_{k\ge n} b_k \ge \inf_{k \ge n} b_k$. Thus, the ratio approaches 1 from above for large n.

Now choose $\kappa' := \kappa$ and $n_0' := \max(n_0(\varepsilon), n_0''(\kappa'))$. That is, we know that then both Equation 10 and Equation 9 hold. Then we want to show:

$$\sup_{k\ge n} \frac{a_k}{b_k} = \max\left(\sup_{k \ge n, a_k \ge 0} \frac{a_k}{b_k}, \sup_{k\ge n, a_k < 0} \frac{a_k}{b_k} \right) \overset{!}{<} \varepsilon',$$

which is a legitimate decomposition of the supremum into the "negative" and "nonnegative" subsequences. But look at the second term ($a_k < 0$) and observe that since $b_k > 0$, clearly $\sup_{k \ge n, a_k < 0} \frac{a_k}{b_k} \le 0 < \varepsilon'$. Thus we only have to consider the first term. Further observe that

$$\sup_{k \ge n, a_k \ge 0} \frac{a_k}{b_k} \le \sup_{k \ge n, a_k \ge 0} a_k \cdot \sup_{k \ge n, a_k \ge 0} \frac{1}{b_k} = \sup_{k\ge n, a_k \ge 0} a_k \cdot \frac{1}{\inf_{k \ge n, a_k \ge 0} b_k},$$

due to nonnegativity of the $a_k \ge 0$ and a general rule for the supremum/infimum, which applies since b_k is strictly positive. Now by assumption,

$$\sup_{k \ge n, a_k \ge 0} a_k \cdot \frac{1}{\inf_{k\ge n, a_k \ge 0} b_k} < \varepsilon' \lim_{n\to\infty} \inf_{k \ge n} b_k \frac{1}{\kappa} \frac{1}{\inf_{k \ge n, a_k \ge 0} b_k} = \varepsilon' \cdot \underbrace{\frac{\lim_{n\to\infty} \inf_{k\ge n} b_k}{\inf_{k\ge n, a_k\ge 0} b_k}}_{\le \kappa} \cdot \frac{1}{\kappa} \le \varepsilon'.$$

Noting that $\inf_{k\ge n} b_k \le \inf_{k \ge n, a_k \ge 0} b_k$ and therefore

$$\frac{\lim_{n\to\infty} \inf_{k\ge n} b_k}{\inf_{k\ge n, a_k \ge 0} b_k} \le \frac{\lim_{n\to\infty} \inf_{k \ge n} b_k}{\inf_{k\ge n} b_k} \le \kappa'.$$

Altogether, we have shown that

$$\sup_{k \ge n} \frac{a_k}{b_k} < \varepsilon',$$

and therefore $LHS \le 0 \implies RHS \le 0$.

It remains to show that $RHS \le 0 \implies LHS \le 0$. Our assumption is that

$$\forall \varepsilon' > 0 : \exists n_0' \in \mathbb{N} : \forall n \ge n_0' : \sup_{k \ge n} \frac{a_k}{b_k} < \varepsilon. \tag{11}$$

and our aim is to show that then

$$\forall \varepsilon > 0 : \exists n_0 \in \mathbb{N} : \forall n \geq n_0 : \sup_{k \geq n} a_k < \varepsilon.$$

So let $\varepsilon > 0$ be fixed. Choose $\varepsilon' := \varepsilon$ and set $n_0 := n_0'$. Then we want to show that $\forall n \geq n_0$:

$$\sup_{k \geq n} a_k = \max\left(\sup_{k \geq n, a_k \geq 0} a_k, \sup_{k \geq n, a_k < 0} a_k \right) \overset{!}{<} \varepsilon.$$

As to the second term, it is obviously negative, in particular $\sup_{k \geq n, a_k < 0} a_k < \varepsilon$. For the first term, where the a_k are nonnegative, observe that then $a_k \leq \frac{a_k}{b_k}$ since $b_k \in (0,1]$, consequently we have $\forall n \geq n_0$:

$$\sup_{k \geq n, a_k \geq 0} a_k \leq \sup_{k \geq n, a_k \geq 0} \frac{a_k}{b_k} \leq \sup_{k \geq n} \frac{a_k}{b_k} < \varepsilon = \varepsilon'.$$

by our assumption Equation 11. And thus we have shown that $RHS \leq 0 \implies LHS \leq 0$. □

A.5 Proof of Remark 4.8

Take for example $X(\omega) = -1$ for a $B \subseteq \Omega$ where $\underline{P}(B) < 1$. Then $\sup X = -1$, but

$$\overline{R}\left(X\chi_B\right) = \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(X\chi_B\right)\left(\vec{\Omega}(i)\right) = \limsup_{n\to\infty} -\frac{1}{n} \sum_{i=1}^{n} \chi_B\left(\vec{\Omega}(i)\right) = -\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \chi_B\left(\vec{\Omega}(i)\right) = -\underline{P}(B),$$

and $\sup X = -1 < -\underline{P}(B)$, hence UP1 does not hold. Thus $X \mapsto \overline{R}\left(X\chi_B\right)$ is not a coherent upper prevision on L^∞ in general (it is of course for $B = \Omega$).

B Existence of Sequences with Prespecified Relative Frequency Cluster Points

In this section we prove Theorem 3.3 and thus demonstate the existence of sequences $\vec{x} : \mathbb{N} \to [k]$ whose corresponding relative frequencies $\vec{r}^{\,\vec{x}} : \mathbb{N} \to \Delta^k$ have the property that their set of cluster points $\mathrm{CP}(\vec{r}^{\,\vec{x}}) = C$, where C is an arbitrary closed rectifiable curve in Δ^k. We do so constructively by providing an explicit procedure which takes a chosen C and constructs a suitable $\vec{x}$. We illustrate our method with some examples. The question of how common sequences with non-convergent relative frequencies are is addressed in Section D.

B.1 Sufficient to Work With Topology Induced by Euclidean Norm on the Simplex

Let $|\Omega| = k < \infty$. A linear prevision E on L^∞ is in a one-to-one correspondence with a finitely additive probability P, which we can represent as a point in the $(k-1)$-simplex (see Lemma B.2 below). It is convenient to then consider the cluster points of a sequence of such probabilities in the $(k-1)$-simplex, with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$. We show that the notion of such a cluster point coincides with a cluster point in $(L^\infty)^*$ with respect to the weak* topology. This is the goal of this section; in particular, we prove Proposition 3.5.

Since $|\Omega| = k < \infty$, we can represent any $X \in L^\infty$ as:

$$X(\omega) = c_1 \chi_{\{\omega_1\}} + \cdots + c_k \chi_{\{\omega_k\}},$$

where $c_i = X(\omega_i)$.

Similarly, any $Z \in (L^\infty)^*$ can be represented as:

$$\begin{aligned} Z(X) &= Z\left(c_1\chi_{\{\omega_1\}} + \cdots + c_k\chi_{\{\omega_k\}}\right) \\ &= c_1 Z\left(\chi_{\{\omega_1\}}\right) + \cdots + c_k Z\left(\chi_{\{\omega_k\}}\right), \end{aligned}$$

since the $Z \in (L^\infty)^*$ are linear functionals and the coefficients c_i depend on X. Intuitively, $Z\left(\chi_{\{\omega_i\}}\right) = P(\omega_i)$ if $Z \in \mathrm{PF}(\Omega)$. Define $d_i := Z\left(\chi_{\{\omega_i\}}\right)$ $\forall i \in 1,\dots,n$ and consequently define

$$\|Z\| := \sqrt{d_1^2 + \cdots + d_k^2}.$$

For a given $Z \in (L^\infty)^*$, call $d_Z := (d_1,..,d_k) \in \mathbb{R}^k$ the coordinate representation of Z.

Lemma B.1. *$\|\cdot\|$ is a norm on $(L^\infty)^*$.*

Proof. Point-separating property: if and only if $Z = 0$, where $0 \in (L^\infty)^*$ is given by $0(X) = 0$, $\forall X \in L^\infty$. But this is easily observed, due to the similar property holding for the Euclidean norm: $\|Z\| = 0$ if and only if $d_i = 0$ $\forall i = 1,\dots,n$. Since $d_i = Z\left(\chi_{\{\omega_i\}}\right)$, this is the case exactly if $Z = 0$.

Subadditivity: $\|Y + Z\| \le \|Y\| + \|Z\|$. For the input $Y + Z$ we get $d_i = (Y+Z)\left(\chi_{\{\omega_i\}}\right) = Y\left(\chi_{\{\omega_i\}}\right) + Z\left(\chi_{\{\omega_i\}}\right)$ and then subadditivity follows from the similar property for the Euclidean norm.

Absolute homogeneity: $\|\lambda Z\| = |\lambda|\|Z\|$, $\forall \lambda \in \mathbb{R}$. This follows easily.

With these properties, $\|\cdot\|$ is a valid norm on $(L^\infty)^*$. □

We now show that the $(k-1)$-simplex is in a one-to-one correspondence with the set of linear previsions $\mathrm{PF}(\Omega)$ via the coordinate representation.

Lemma B.2. *Let $d \in \Delta^k$. Then $Z(X) := c_1 d_1 + ... + c_k d_k \in \mathrm{PF}(\Omega)$. Conversely, let $Z \in \mathrm{PF}(\Omega)$. Then the corresponding $d_Z \in \Delta^k$.*

Proof. Let $d \in \Delta^k$ and $Z(X) := c_1 d_1 + \cdots + c_k d_k$. Since $\sum_{i=1}^k d_i = 1$ we have immediately that $Z\left(\chi_\Omega\right) = 1$, noting that $\chi_\Omega = 1\chi_{\{\omega_1\}} + \cdots + 1\chi_{\{\omega_k\}}$. Also, if $X \ge 0$, i.e. $c_i \ge 0$, $\forall i \in [k]$, then $Z(X) \ge 0$ since $d_i \ge 0$ $\forall i$. Thus, $Z \in \mathrm{PF}(\Omega)$.

Conversely, let Z be a linear prevision, i.e. $Z\left(\chi_\Omega\right) = 1$ and $Z(X) \ge 0$ if $X \ge 0$. From $Z\left(\chi_\Omega\right) = 1$ we can deduce that $\sum_{i=1}^k d_i = 1$. If $X \ge 0$, we know that $c_i \ge 0$ $\forall i \in [k]$, hence $Z(X) \ge 0$ can only be true if all $d_i \ge 0$. Thus $d_Z \in \Delta^k$. □

We here restate Proposition 3.5 for convenience.

Proposition B.3. *Let $\vec{E}(n) : \mathbb{N} \to \mathrm{PF}(\Omega)$ be a sequence of linear previsions with underlying probabilities $\vec{P}(n) := A \mapsto \vec{E}(n)(A)$. Then $E \in \mathrm{CP}\left(\vec{E}(n)\right)$ with respect to the weak* topology if and only if the sequence $\vec{D} : \mathbb{N} \to \Delta^k$, $\vec{D}(n) := \left(\vec{P}(n)(\omega_1),\dots,\vec{P}(n)(\omega_k)\right)$ has as cluster point $d_E = \left(E\left(\chi_{\{\omega_1\}}\right),\dots,E\left(\chi_{\{\omega_k\}}\right)\right)$ with respect to the topology induced by the Euclidean norm on $\mathbb{R}^k$.*

First note that if $E \in \mathrm{PF}(\Omega)$, then $d_E = \left(E\left(\chi_{\{\omega_1\}}\right),\dots,E\left(\chi_{\{\omega_k\}}\right)\right) = (P(\omega_1),\dots,P(\omega_k))$, where P is the underlying probability of E, and hence $\|E\| = \sqrt{P(\omega_1)^2 + \cdots + P(\omega_k)^2}$. To complete the proof, we need some further statements first.

Definition B.4. *A vector space $\mathscr{X}$ is called topological vector space if the topology on $\mathscr{X}$ is such that $(x, y) \mapsto x + y$ is continuous with respect to the product topology on $\mathscr{X} \times \mathscr{X}$ and $(\lambda, x) \mapsto \lambda x$ is continuous with respect to the product topology on $\mathbb{R} \times \mathscr{X}$. We call a topology which makes $\mathscr{X}$ a topological vector space a* linear topology.

Remark B.5. *The weak* topology makes $(L^\infty)^*$ a topological vector space with a Hausdorff topology,*[26] *where vector addition and scalar product are defined pointwise: $Y+Z := X \mapsto Y(X)+Z(X)\ \forall X \in L^\infty$, $\lambda Z := X \mapsto \lambda Z(X)\ \forall X \in L^\infty$.*

Remark B.6 (well-known)**.** *A vector space whose topology is induced by a norm is a topological vector space.*

Proposition B.7. *On every finite dimensional vector space X there is a unique topological vector space structure. In other words, any two Hausdorff linear topologies on X coincide (Nagy, 2007).*

Now Proposition B.3 directly follows.

Proof of Proposition B.3. Our norm $\|\cdot\|$ makes $(L^\infty)^*$ a topological vector space due to Remark B.6, and any topology induced by a norm is Hausdorff; but the weak* topology also makes $(L^\infty)^*$ a topological vector space, and the weak* topology is Hausdorff. Hence we can conclude from Proposition B.7 that they coincide. But then the two notions of what a cluster point is of course also coincide, since this depends just on the topology. □

Thus, for the proof of Theorem 3.1, we will work exclusively with the topology induced by the Euclidean metric restricted to Δ^k. For $z \in \Delta^k$ and $\epsilon > 0$ define the ϵ-neighbourhood (ball)

$$N_\epsilon(z) := \left\{p \in \Delta^k \colon \|p - z\| < \epsilon\right\},$$

where $\|\cdot\|$ is the Euclidean norm (restricted to the simplex). Then from (Schechter, 1997, p. 430) we have an equivalent definition of a cluster point:

Definition B.8. *Say that $z \in \Delta^k$ is a cluster point of a sequence $\vec{x} \colon \mathbb{N} \to \Delta^k$ (and denote by $\mathrm{CP}(\vec{x})$ the set of all cluster points of $\vec{x}$) if for all $\epsilon > 0$, $\left|\{n \in \mathbb{N} \colon \vec{x}(n) \in N_\epsilon(z)\right| = \aleph_0$, where $\aleph_0 := |\mathbb{N}|$ is the cardinality of the natural numbers.*

B.2 Further Notation

We will work solely with the topology on Δ^k induced by the Euclidean metric; by the argument in the previous subsection the cluster points w.r.t. this topology coincide with those w.r.t. the weak* topology.

We introduce further notation to assist in stating our algorithm. For terser notation, we drop the $\vec{\cdot}$ symbol for sequences $\vec{x} \colon \mathbb{N} \to [k]$ throughout this appendix and simply write x. The ith canonical unit vector in Δ^k is denoted $e_i := (0, \ldots, 1, \ldots, 0)$, where the 1 is in the ith position. The boundary of the simplex is

$$\partial\Delta^k := \{(z_1, \ldots, z_k) \colon z_1, \ldots, z_k \geq 0,\ z_1 + \cdots + z_k = 1\}.$$

If $p_1, p_2 \in \Delta^k$ then $l(p_1, p_2) := \{\lambda p_1 + (1-\lambda) p_2 \colon \lambda \in [0,1]\}$ is the *line segment connecting p_1 and p_2*. If $C \subset \Delta^k$ is a rectifiable closed curve parametrised by $c \colon [0,1] \to \Delta^k$, its length is $\mathrm{length}(C) = \int_0^1 |c'(t)| dt$. For $y \in \mathbb{R}$, $\lfloor y \rceil$ is the nearest integer to y: $\lfloor y \rceil := \lfloor y + \frac{1}{2} \rfloor$. We apply certain operations T elementwise. For example, if $z = \langle z_1, \ldots, z_k \rangle \in \Delta^k$, then $\lfloor Tz \rceil := \langle \lfloor Tz_1 \rceil, \ldots, \lfloor Tz_k \rceil \rangle$ and (overloading notation) for $T \in \mathbb{N}$ and $\iota \in \mathbb{N}^k$, $\iota / T \in \mathbb{R}^n$ is simply $\langle \iota_1/T, \ldots, \iota_k/T \rangle$. If $i < j \in \mathbb{N}$ the "interval" is $[i, j] := \{m \in \mathbb{N} \colon i \leq m \leq j\}$.

To avoid confusion, we will reserve "sequence" for the infinitely long $x \colon \mathbb{N} \to [k]$ and use "segment" to denote finite length strings $z \colon [n] \to [k]$ which we will write explicitly as $\langle z_1, \ldots, z_n \rangle$. We construct the sequence x attaining the desired behavior of r^x by iteratively appending a series of segments. We denote the empty segment as $\langle\rangle$. If x^1 and x^2 are two finite segments of lengths ℓ_1 and ℓ_2 then their concatenation is the length $\ell_1 + \ell_2$ segment $x^1 x^2 := \left\langle x^1_1, \ldots, x^1_{\ell_1}, x^2_1, \ldots, x^2_{\ell_2} \right\rangle$. We extend the $i^{[j]}$ notation to segments: if

[26] See for instance Exercise 13 and 14 in Tao (2009).

$z = \langle z_1, \ldots, z_\ell \rangle$, then $z^{[\iota]} := \langle z, z \ldots, z \rangle$ is the length $\ell\iota$ segment formed by concatenating ι copies of z. Given $n \in \mathbb{N}$ and a sequence $x \colon \mathbb{N} \to [k]$, the shifted sequence $x^{+n} \colon \mathbb{N} \to [k]$ is defined via $x^{+n}(i) := x(i+n)$ for $i \in \mathbb{N}$.

B.3 Properties of Relative Frequency Sequences

Our construction of x relies upon the following elementary property of relative frequency sequences.

Lemma B.9. *Suppose $k, n, m \in \mathbb{N}$, $x \colon \mathbb{N} \to [k]$. Then*

$$r^x(n+m) = \frac{n}{n+m} r^x(n) + \frac{m}{n+m} r^{x^{+n}}(m). \tag{12}$$

Proof. For any $i \in [k]$ we have

$$\begin{aligned} r_i^x(n+m) &= \tfrac{1}{n+m} |\{j \in [n+m] \colon x(j) = i\}| \\ &= \tfrac{1}{n+m} \left(|\{j \in [n] \colon x(j) = i\}| + |\{j \in [n+m] \setminus [n] \colon x(j) = i\}| \right) \\ &= \tfrac{1}{n+m} \tfrac{n}{n} |\{j \in [n] \colon x(j) = i\}| + \tfrac{1}{n+m} \tfrac{m}{m} |\{t \in [m] \colon x(n+t) = i\}| \\ &= \tfrac{n}{n+m} r_i^x(n) + \tfrac{m}{n+m} \tfrac{1}{m} |\{t \in [m] \colon x^{+n}(t) = i\}| \\ &= \tfrac{n}{n+m} r_i^x(n) + \tfrac{m}{n+m} r_i^{x^{+n}}(m). \end{aligned}$$

Since this holds for all $i \in [k]$ we obtain Equation 12. □

Observe that (12) also holds when $x \colon [n+m] \to [k]$ is a segment, in which case $x^{+n} = \langle x_{n+1}, \ldots, x_{n+m} \rangle$. Furthermore note that (12) is a convex combination of the two points $r^x(n)$ and $r^{x^{+n}}(m)$ in Δ^k since $\frac{n}{n+m} + \frac{m}{n+m} = 1$ and both coefficients are positive. These two points are (respectively) the relative frequency of x at n, and the relative frequency of x^{+n} at m. This latter sequence will be the piece "added on" at each stage of our construction and forms the basis of our piecewise linear construction of r^x such that its cluster points are a given $C \subset \Delta^k$.

The set of cluster points of any sequence is closed. In addition, we have

Lemma B.10. *For any $k \in \mathbb{N}$ and $x \colon \mathbb{N} \to [k]$, $\mathrm{CP}(r^x)$ is a connected set.*

This follows immediately from (Bauschke et al., 2015, Lemma 2.6) upon observing that $\lim_{n\to\infty} \|r^x(n) - r^x(n+1)\| = 0$ since $\|r^x(n) - r^x(n+1)\| = \|r^x(n) - \frac{n}{n+1} r^x(n) - \frac{1}{n+1} e_{(x(n+1))}\| = \frac{1}{n+1} \|r^x(n) - e_{x(n+1)}\| \leq \frac{2}{n+1}$. The boundedness of r^x is essential for this to hold — for unbounded sequences the set of cluster points need not be connected (Ašić & Adamović, 1970).

B.4 Logic of the Construction

The idea of our construction is as follows (see Figure 1 below for a visual aid). In order to satisfy the definition of cluster points, we need to return to each neighbourhood of each point in C infinitely often. To that end we iterate through an infinite sequence of generations indexed by $g \in \mathbb{N}$. For each g, we approximate C by a polygonal approximation C^g comprising V^g-many seperate segments. We choose the sequence $(V^g)_{g\in\mathbb{N}}$ so that C^g approaches C in an appropriate sense. Then for generation g we append elements to x to ensure the sequence of relative frequencies makes another cycle approximately following C^g. We control the approximation error of this process and ensure its error is of a size that also decreases with increasing g.

We now describe the construction of a single generation. Thus suppose g is now fixed and suppose the current partial sequence (segment) x has length n. We suppose (and will argue this is ok later) that $r^x(n)$ is close to one of the vertices of C^{g-1}. We then choose a finer approximation C^g of C (since $V^g > V^{g-1}$).

Algorithm 1 Construction of x such that $\mathrm{CP}(r^x) = C$

Require: $C \subset \Delta^k$, a rectifiable closed curve parametrized as $c\colon [0,1] \to \Delta^k$
Require: $V\colon \mathbb{N} \to \mathbb{N}$ ▷ Number of segments at generation g; as a function of n
Require: $T\colon \mathbb{N} \to \mathbb{N}$ ▷ Controls quantization of angle; needs to be increasing
$x \leftarrow \langle 1 \rangle$ ▷ Arbitrary initialization $x_1 = 1$
$p_{\text{old}} \leftarrow e_1$ ▷ $p_{\text{old}} = r^{\langle 1 \rangle}(1) = e_1$
$n \leftarrow 1$ ▷ n is always updated to correspond to the current length of x
$g \leftarrow 1$
while true **do** ▷ Iterate over repeated generations g; V is chosen at start of generation
    $V \leftarrow V(g)$ ▷ Choose V for generation g
    $p_v \leftarrow c(v/V)$ for $v = 0, \ldots, V$ ▷ Vertices of $C^g := \bigcup_{v \in [V]} l(p_{v-1}, p_v)$
    $v \leftarrow 0$
    while $v \le V$ **do** ▷ For all vertices of C^g
        $T \leftarrow T(n)$ ▷ Quantization of angle; chosen per segment
        $p_{\text{new}} \leftarrow p_{v+1}$ ▷ The next vertex of C^g
        $\gamma_i \leftarrow p_{\text{old},i}/(p_{\text{old},i} - p_{\text{new},i})$ for $i \in [k]$ ▷ Will have $p_{\text{old}} \approx p_v$
        $\gamma \leftarrow \min\{\gamma_i : i \in [k], \gamma_i > 0\}$ ▷ See (15)
        $p^* \leftarrow \gamma(p_{\text{new}} - p_{\text{old}}) + p_{\text{old}}$ ▷ Determine $p^* \in \partial\Delta^k$
        $\iota \leftarrow \lfloor T p^* \rceil$ ▷ Elementwise; $\iota = (\iota_1, \ldots, \iota_k)$
        $\tilde{p}^* \leftarrow \iota/T$ ▷ Elementwise; quantized version of p^*
        $\tilde{T} \leftarrow \sum_{i=1}^k \iota_i$ ▷ Will have $\tilde{T} \approx T$
        $y \leftarrow \langle 1^{[\iota_1]}, \ldots, k^{[\iota_k]} \rangle$ ▷ The string y is thus of length $\tilde{T}$
        $\tilde{\ell} \leftarrow \lceil \frac{n}{T(\gamma-1)} \rceil$ ▷ Integer number of repetitions of y needed
        $x \leftarrow x\, y^{[\tilde{\ell}]}$ ▷ Construct new x by appending z, comprising $\tilde{\ell}$ copies of y
        $n \leftarrow n + \tilde{\ell}\tilde{T}$ ▷ Length of x now
        $p_{\text{old}} \leftarrow r^x(n)$ ▷ Relative frequency at current n
        $v \leftarrow v + 1$ ▷ Move onto next vertex of C^g
    end while
    $g \leftarrow g + 1$ ▷ Move onto next generation of the construction
end while ▷ Procedure never terminates

For each vertex p_v^g, $v \in [V^g]$ we append elements to x resulting in a segment of length n'. We do this in a manner such that we move the relative frequency from $r^x(n)$ to $r^x(n') \approx p_v^g$. We do so by appending multiple copies of a vector z to x where $r^z(m)$ points in the same direction as the direction one needs to go from p_{old} to p_{new}. This can only be done approximately because with a finite length segment, the set of directions one can move the relative frequencies is quantized. We choose the fineness of the quantization to be fine enough to achieve the accuracy we need. That is governed by the parameter $T \in \mathbb{N}$. We then append $\tilde{\ell}$ copies of z to x where $\tilde{\ell}$ is the integer closest to the real number ℓ that would be the ideal number of steps needed to get to the desired point p_{new}. We also control the error incurred by approximating ℓ by $\tilde{\ell}$. The upshot of this is that with the resulting extension to x we have $r^x(n')$ is sufficiently close to p_{new}. We then repeat this operation for all the vertices p_v^g for $v \in [V^g]$. This completes generation g. We show below that for each generation g, *all* the points in the relative frequency sequence are adequately close to C^g, where "adequately close" is quantified and increases in accuracy as g increases.

We consistently use the following terminology in describing our algorithm:

generation These are indexed by g and entail an entire pass around the curve C, or more precisely its polygonal approximation $C^g := \bigcup_{v \in [V^g]} l(p_{v-1}, p_v)$

segment Corresponds to a single line segment $l(p_{v-1}, p_v)$ of the gth polygonal approximation.

piece Corresponds to appending $z = \langle 1^{[\iota_1]}, \ldots, k^{[\iota_k]} \rangle$ to x, which results in moving $r^x(n)$ in the direction $\hat{p}^*$.

step The appending of a single element of z, which will always move $r^x(n)$ towards one of the vertices of the simplex e_i $(i \in [k])$.

The end result is that we have constructed a procedure (Algorithm 1) which runs indefinitely (g increases without bound), and which has the property that for any choice of $\epsilon > 0$, if one waits long enough, there will be a sufficiently large g such that all the relative frequencies associated with generated x are within ϵ_g of C, and $(\epsilon_g)_{g \in \mathbb{N}}$ is a null sequence. We will thus conclude that $\mathrm{CP}(r^x) \supseteq C$. We will also argue that $\mathrm{CP}(r^x) \subseteq C$ completing the proof.

B.5 Construction of p^* and its Approximation $\tilde{p}^*$

The basic idea of the construction is to exploit Lemma B.9. Suppose $n \in \mathbb{N}$ (and suppose it is "large") and fix $m = 1$ in (12) to obtain

$$r^x(n+1) = \tfrac{n}{n+1} r^x(n) + \tfrac{1}{n+1} r^{x^{+n}}(1). \tag{13}$$

Now $r^{x^{+n}}(1) = e_{x(n+1)}$ and so $r^x(n+1) = \frac{n}{n+1} r^x(n) + \frac{1}{n+1} e_{x(n+1)}$. When n is large $\frac{n}{n+1} \approx 1$ and $\frac{1}{n+1}$ is small, and so this says that appending $x(n+1)$ to the length n segment $x([n])$ moves the relative frequency r^x from $r^x(n)$ in the direction of $e_{x(n+1)}$ by a small amount. Observe that the *only* directions which the point $r^x(n)$ can be moved is towards one of the vertices of the k-simplex, $e_1, \ldots, e_k$. Thus if we had, for a fixed n that $r^x(n) = p_{\mathrm{old}}$ and we wished to append m additional elements z to x to produce xz such that $r^{xz}(n+m) = p_{\mathrm{new}}$, we need to figure out a way of heading in the direction $d = p_{\mathrm{new}} - p_{\mathrm{old}}$ when at each step we are constrained to move a small amount to one of the vertices. The solution is to approximate the direction d by a quantized choice that can be obtained by an integer number of elements of $[k]$.

Given arbitrary $p_{\mathrm{old}} \neq p_{\mathrm{new}} \in \operatorname{relint} \Delta^k$, we define p^* to be the intercept by $\partial \Delta^k$ of the line segment starting at p_{old} and passing through p_{new}. (If $p_{\mathrm{new}} \in \partial \Delta^k$ set $p^* = p_{\mathrm{new}}$.) The intercept on the boundary of Δ^k is denoted p^* and is given by

$$p^* := \gamma(p_{\mathrm{new}} - p_{\mathrm{old}}) + p_{\mathrm{old}} \tag{14}$$

for some $\gamma > 0$. We can determine γ as follows. The choice of γ can not take p^* outside the simplex. Thus let γ_i $(i \in [k])$ satisfy $\gamma_i(p_{\mathrm{new}_i} - p_{\mathrm{old}_i}) + p_{\mathrm{old}_i} = 0$. Thus $\gamma_i = \frac{p_{\mathrm{old}_i}}{p_{\mathrm{old}_i} - p_{\mathrm{new}_i}}$. Any $\gamma_i < 0$ points in the wrong direction and so we choose

$$\gamma := \min\{\gamma_i \colon i \in [k] \text{ and } \gamma_i > 0\}. \tag{15}$$

Such a choice of γ guarantees that $p^* \in \partial \Delta^k$. Observe that the requirement that $\gamma_i > 0$ means the denominator in the definition of γ_i is positive and less than the numerator, and thus all γ_i which are positive exceed 1, and consequently $\gamma > 1$.

We can now take p^* to be the direction we would like to move $r^x(n)$ towards. However our only control action is to choose a sequence $z \in [k]^m$. To that end we suppose we quantize the vector p^* so that it has rational components with denominator $T \in \mathbb{N}$ (which will be strategically chosen henceforth). As we will shortly show, this will allow us to move (approximately) towards p^*. Thus let $\iota_i := \lfloor T p_i^* \rceil$ for $i \in [k]$ and set

$$\tilde{p}^* := \left(\frac{\iota_1}{T}, \ldots, \frac{\iota_k}{T} \right). \tag{16}$$

Observe that $\tilde{p}^*$ is not guaranteed to be in Δ^k because there is no guarantee that $\sum_{i=1}^k \tilde{p}_i^* = 1$.

We have

Lemma B.11. *Let p^* and $\tilde{p}^*$ be defined by (14) and (16) respectively. Then for all $i \in [k]$, $\left|\tilde{p}_i^* - p_i^*\right| \le \frac{1}{2T}$.*

Proof. $\left|\tilde{p}_i^* - p_i^*\right| = \frac{1}{T}\left|T\tilde{p}_i^* - Tp_i^*\right| = \frac{1}{T}\left|\lfloor Tp_i^*\rceil - Tp_i^*\right| \le \frac{1}{2T}$, by definition of the rounding operator $\lfloor\cdot\rceil$. □

B.6 Determining the Number of Steps to Take

Observe that (12) can be written as

$$r^x(n+m) = (1-\alpha)r^x(n) + \alpha r^{x^{+n}}(m). \tag{17}$$

where $\alpha = \frac{m}{n+m}$ and thus $1-\alpha = \frac{n}{n+m}$. This suggests that we can engineer the construction of x by requiring a suitable α such that

$$(1-\alpha)p_{\text{old}} + \alpha p^* = p_{\text{new}}.$$

Recall we want the sequence r^x to move from p_{old} to p_{new} which can be achieved by a taking a suitable convex combination of p_{old} and p^*, which corresponds to appending a suitable number of copies of z to x, where z is chosen to move $r^x(n)$ in the direction of p^*. If we substitute the definition of p^* from (14) we obtain the problem:

$$\begin{aligned}
&\text{Find } \alpha \text{ such that } (1-\alpha)p_{\text{old}} + \alpha[\gamma(p_{\text{new}} - p_{\text{old}}) + p_{\text{old}}] = p_{\text{new}} \\
\Leftrightarrow\ &\text{Find } \alpha \text{ such that } (1-\alpha)p_{\text{old}} + \alpha\gamma p_{\text{new}} - \alpha\gamma p_{\text{old}} + \alpha p_{\text{old}} - p_{\text{new}} = 0_k \in \mathbb{R}^k \\
\Leftrightarrow\ &\text{Find } \alpha \text{ such that } p_{\text{old}}(1-\alpha\gamma) + p_{\text{new}}(\alpha\gamma - 1) = 0_k,
\end{aligned}$$

which is only true when either $p_{\text{old}} = p_{\text{new}}$ (which is a trivial case) or when $1-\alpha\gamma = 0$ and thus $\alpha := 1/\gamma$, which we take as a definition. Since $\gamma > 1$ this implies $\alpha < 1$, which is consistent with our original motivation for taking convex combinations.

We will append the segment z to x, where $z = y^{[\ell]}$ and $y = \langle 1^{[\iota_1]}, \ldots, k^{[\iota_k]}\rangle$. Now if each y is of length T and we notionally made ℓ repetitions, we would have $m = \ell T$. From the definition of α this means

$$\alpha = \frac{\ell T}{n + \ell T}. \tag{18}$$

We presume n is given (at a particular stage of construction) and $T \in \mathbb{N}$ is a fixed design parameter. We can thus solve for ℓ to obtain

$$\ell = \frac{\alpha n}{T(1-\alpha)} = \frac{(1/\gamma)n}{T(1-1/\gamma)} = \frac{n}{T(\gamma-1)}.$$

Observe that ℓ is not guaranteed to be an integer, a complication we will deal with later. If it was an integer, we would create z by concatenating ℓ copies of y which is of length T. The vector y moves p_{old} towards p^*. By appending ℓ copies we should move r^x to p_{new} as desired.

However we do not head exactly in the direction of p^*, since we worked with a quantized version $\tilde{p}^*$ instead, and we can not always take ℓ copies because ℓ is not guaranteed to be an integer; instead we will take

$$\tilde{\ell} := \left\lceil \frac{\alpha n}{T(1-\alpha)} \right\rceil = \left\lceil \frac{n}{T(\gamma-1)} \right\rceil$$

copies of y which will move $r^x(n)$ towards $\hat{p}^*$ instead of p^*. We now proceed to analyse the effect of these approximations on our construction.

B.7 Analyzing the Effect of Approximations

The parameter $T \in \mathbb{N}$ is a design variable. Our construction will utilize

$$\tilde{T} := \sum_{j=1}^{k} \iota_j, \tag{19}$$

where $\iota_j = \lfloor T p_j^* \rceil$ for $j \in [k]$. Then $\tilde{T} \approx T$, a claim which we quantify below.

Lemma B.12. *Suppose $T \in \mathbb{N}$, and $\tilde{T}$ is defined as above. Then $T - \frac{k}{2} \leq \tilde{T} \leq T + \frac{k}{2}$.*

Proof. By definition of the rounding operator $\lfloor \cdot \rceil$, we have that

$$\left| \iota_j - \lfloor T p_j^* \rceil \right| \leq \tfrac{1}{2} \quad \forall j \in [k].$$

Thus

$$\begin{aligned} & T p_j^* - \frac{1}{2} \leq \iota_j \leq T p_j^* + \frac{1}{2} \quad \forall j \in [k] \\ \Rightarrow & \sum_{j=1}^{k} \left(T p_j^* - \frac{1}{2} \right) \leq \sum_{j=1}^{k} \iota_j \leq \sum_{j=1}^{k} \left(T p_j^* + \frac{1}{2} \right) \\ \Rightarrow & T - \frac{k}{2} \leq \tilde{T} \leq T + \frac{k}{2}. \end{aligned}$$

□

Ideally we move $r^x(n)$ to

$$p_{\text{new}} = (1-\alpha) p_{\text{old}} + \alpha p^*. \tag{20}$$

by appending z (i.e. we hope that $r^{xz}(n+m) = p_{\text{new}}$). But in fact the segment z which we will append to x will move r^x from p_{old} instead to

$$\hat{p}_{\text{new}} := (1-\tilde{\alpha}) p_{\text{old}} + \tilde{\alpha} \hat{p}^*, \tag{21}$$

where

$$\tilde{\alpha} := \frac{\tilde{\ell} T}{\tilde{\ell} T + m}. \tag{22}$$

and

$$\hat{p}^* := \left(\frac{\iota_1}{\tilde{T}}, \ldots, \frac{\iota_k}{\tilde{T}} \right). \tag{23}$$

We now determine the error incurred from these approximations. We first need the following Lemma:

Lemma B.13. *Suppose $n, T \in \mathbb{N}$, α is defined by (18) and $\tilde{\alpha}$ is defined by (22). Then*

$$|\alpha - \tilde{\alpha}| \leq \frac{T}{n}.$$

Proof. By definition of $\tilde{\alpha}$ we have

$$\begin{aligned} |\alpha - \tilde{\alpha}| &= \left| \frac{\tilde{\ell} T}{\tilde{\ell} T + n} - \frac{\ell T}{\ell T + n} \right| \\ &= \left| \frac{\tilde{\ell} T (\ell T + n) - \ell T (\tilde{\ell} T + n)}{(\tilde{\ell} T + n)(\ell T + n)} \right| \\ &= \left| \frac{\tilde{\ell} T n - \ell T n}{(\tilde{\ell} T + n)(\ell T + n)} \right|. \end{aligned}$$

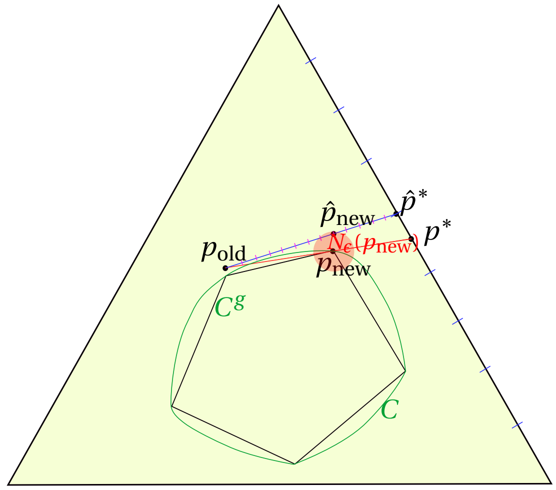


Figure 1: Illustration of construction of the sequence x. The figure shows how a single segment is created. We have the desired curve C (in dark green) and a polygonal approximation C^g using 5 segments (thus $V^g = 5$). We assume that we have already constructed the first n elements of x and thus we can compute $r^x(n)$. Denote this by p_{old}, an element of the simplex, itself shown in pale green. We hope to append a segment y of length m to x such $r^{xy}(n+m) = p_{\text{new}}$. However we can only choose elements of y from $[k]$ and that means that at each step we move towards one of the vertices of the simplex. We note that ideally we would move from p_{old} towards $p^* \in \partial\Delta^k$. In order to deal with the restrictions on directions we can head, we quantize p^* as $\hat{p}^* = \lfloor Tp^* \rceil / T$ where in this example we have chosen $T = 9$. As argued in the main text this means we will thus be restricted to heading towards one of a fixed set of points on the boundary (marked in blue). Observe $\hat{p}^*$ is located at one of these points in the diagram, but in general it might not even be on the boundary of the simplex. We then construct y to move r^x towards $\hat{p}^*$ and take sufficient steps to move to $\hat{p}_{\text{new}}$. This too is done in repeated steps by setting $y = z^{[\tilde{\ell}]}$ where z is a shorter segment (marked by purple ticks on the line segment $l(p_{\text{old}}, \hat{p}^*)$) which will move towards $\hat{p}^*$ by a small amount. The end result is that we get $r^{xy}(n+m) = \hat{p}_{\text{new}}$ which is contained within $N_\epsilon(p_{\text{new}})$, an ϵ-ball centered at p_{new}.

Since $(\tilde{\ell}T + n)(\ell T + n) > 0$ and $\tilde{\ell} = \lceil \ell \rceil \geq \ell$, we have

$$(\tilde{\ell}T + n)(\ell T + n) \geq (\ell T + n)(\ell T + n) \geq n^2$$

and since $|\tilde{\ell} - \ell| \leq 1$,

$$|\alpha - \tilde{\alpha}| = \frac{|\tilde{\ell} - \ell| \cdot Tn}{n^2} \leq \frac{T}{n}.$$

□

Our construction does not move $r^x(n)$ towards $\tilde{p}^*$ but an approximation of it, namely $\hat{p}^*$ defined in (23). We exploit the fact that repeating a segment does not change its relative frequencies, which we state formally as

Lemma B.14. *Let $z = \langle 1^{[\iota_1]}, \ldots, k^{[\iota_k]} \rangle$. Then $r^z(\tilde{T}) = \hat{p}^* = r^{z^{[\tilde{\ell}]}}(\tilde{\ell}\tilde{T})$.*

Proof. For any $i \in [k]$ we have $r_i^z(\tilde{T}) = \frac{1}{\tilde{T}} \left| \{ j \in [\tilde{T}] : z_j = i \} \right| = \frac{1}{\tilde{T}} \iota_i$. The first equality is immediate. For the second, similarly we have $r_i^{z^{[\tilde{\ell}]}}(\tilde{\ell}\tilde{T}) = \frac{1}{\tilde{\ell}\tilde{T}} \left| \left\{ j \in [\tilde{\ell}\tilde{T}] : z_j^{[\tilde{\ell}]} = i \right\} \right| = \frac{1}{\tilde{\ell}\tilde{T}} \cdot \tilde{\ell}\iota_i = \frac{1}{\tilde{T}}\iota_i$ by definition of $z^{[\tilde{\ell}]}$. □

Since $\tilde{p}^* = \left(\frac{\iota_1}{T}, \ldots, \frac{\iota_k}{T} \right)$ we have that $\tilde{p}^* = \frac{\tilde{T}}{T} \hat{p}^*$. This allows us to show:

Lemma B.15. *Suppose $T \in \mathbb{N}$ and $\hat{p}^*$ is defined via (23). Then $\| \hat{p}^* - \tilde{p}^* \| \leq \frac{k}{2T}$.*

Proof. We have $\|\hat{p}^* - \tilde{p}^*\| = \left\|\hat{p}^* - \frac{\tilde{T}}{T}\hat{p}^*\right\| = \left|1 - \frac{\tilde{T}}{T}\right| \cdot \|\hat{p}^*\| \leq \left|1 - \frac{\tilde{T}}{T}\right|$. Suppose $\tilde{T} < T$, then $1 - \frac{\tilde{T}}{T} > 0$ and $\left|1 - \frac{\tilde{T}}{T}\right| = 1 - \frac{\tilde{T}}{T} \leq 1 - \frac{T-k/2}{T} = \frac{k}{2T}$ by Lemma B.12. Similarly if $\tilde{T} > T$, then $1 - \frac{\tilde{T}}{T} < 0$ and $\left|1 - \frac{\tilde{T}}{T}\right| = \frac{\tilde{T}}{T} - 1 \leq \frac{T+k/2}{T} - 1 = \frac{k}{2T}$ completing the proof. □

Lemma B.16. *Suppose $k, T \in \mathbb{N}$ and p_{new} and $\hat{p}_{\text{new}}$ are defined as above. Then*

$$\|\hat{p}_{\text{new}} - p_{\text{new}}\| \leq \frac{4T}{n} + \frac{k}{T}. \tag{24}$$

Proof. From (20) and (21) we have

$$\begin{aligned}
\|\hat{p}_{\text{new}} - p_{\text{new}}\| &= \|(1-\tilde{\alpha})p_{\text{old}} + \tilde{\alpha}\hat{p}^* - (1-\alpha)p_{\text{old}} - \alpha p^*\| \\
&= \|[(1-\tilde{\alpha}) - (1-\alpha)]p_{\text{old}} + (\tilde{\alpha}\hat{p}^* - \alpha p^*)\| \\
&\leq \|(\alpha - \tilde{\alpha})p_{\text{old}}\| + \|\tilde{\alpha}\hat{p}^* - \alpha p^*\| \\
&\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \|\tilde{\alpha}\hat{p}^* - \alpha p^*\|. \qquad (25)
\end{aligned}$$

The second term in (25) can be bounded as follows:

$$\begin{aligned}
\|\tilde{\alpha}\hat{p}^* - \alpha p^*\| &= \|(\tilde{\alpha} - \alpha + \alpha)\hat{p}^* - \alpha p^*\| \\
&= \|(\tilde{\alpha} - \alpha)\hat{p}^* + (\alpha\hat{p}^* - \alpha p^*)\| \\
&\leq \|(\tilde{\alpha} - \alpha)\hat{p}^*\| + \|\alpha\hat{p}^* - \alpha p^*\| \\
&= |\tilde{\alpha} - \alpha| \cdot \|\hat{p}^*\| + \alpha\|\hat{p}^* - p^*\| \\
&= |\tilde{\alpha} - \alpha| \cdot \|\hat{p}^*\| + \alpha\|(\hat{p}^* - \tilde{p}^*) + (\tilde{p}^* - p^*)\| \\
&\leq |\tilde{\alpha} - \alpha|\sqrt{2} + \alpha\|\hat{p}^* - \tilde{p}^*\| + \alpha\|\tilde{p}^* - p^*\| \\
&\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha\left(\textstyle\sum_{i=1}^k \left(\tilde{p}_i^* - p_i^*\right)^2\right)^{1/2},
\end{aligned}$$

by Lemma B.15 and the fact that $\|\hat{p}^*\| \leq 1$,

$$\begin{aligned}
&\leq \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha\left(\textstyle\sum_{i=1}^k \left(\frac{1}{2T}\right)^2\right)^{1/2} \\
&= \sqrt{2}|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \alpha\frac{\sqrt{k}}{2T}, \qquad (26)
\end{aligned}$$

where we used Lemma B.11 in the penultimate step. Since $\alpha \leq 1$, combining (25) and (26) we have

$$\|\hat{p}_{\text{new}} - p_{\text{new}}\| \leq 2\sqrt{2}\,|\tilde{\alpha} - \alpha| + \frac{k}{2T} + \frac{\sqrt{k}}{2T} \leq 4\,|\tilde{\alpha} - \alpha| + \frac{k}{T}.$$

Appealing to Lemma B.13 gives us (24). □

The above arguments control the errors at the end of a *piece* (and thus in a *segment*). But for later purposes we need control at each *step*. This follows immediately by the fact that we make small steps:

Lemma B.17. *For $n \in \mathbb{N}$ and $m \in [\tilde{T}]$ and any $x\colon \mathbb{N} \to [k]$,*

$$\left\|r^x(n) - r^x(n+m)\right\| \leq \frac{2T+k}{n}.$$

Proof. By Lemma B.9,

$$\begin{aligned}\left\|r^x(n)-r^x(n+m)\right\| &= \left\|r^x(n)-\tfrac{n}{n+m}r^x(n)-\tfrac{m}{n+m}r^{x^{+n}}(m)\right\| \\ &= \left\|(1-\tfrac{n}{n+m})r^x(n)-\tfrac{m}{n+m}r^{x^{+n}}(m)\right\| \\ &= \frac{m}{n+m}\left\|r^x(n)-r^{x^{+n}}(m)\right\| \\ &\le \frac{2m}{n+m} \\ &\le \frac{2m}{n} \\ &\le \frac{2T+k}{n},\end{aligned}$$

where the first inequality holds since $\|r^x(n)\|, \|r^{x^{+n}}(m)\| \le 1$ and the last step follows from Lemma B.12. □

B.8 Completing the Proof of Theorem 3.3

The remaining piece of the argument concerns the piecewise linear approximation of C by C^g. For $A, B \subseteq \Delta^k$ and $a \in \Delta^k$ define $d(a,B) := \min\{\|a-b\|: b \in B\}$ and the Hausdorff distance

$$d(A,B) := \max\{d(a,B): a \in A\}. \tag{27}$$

Let us write V^g (the number of vertices in the piecewise linear approximation at generation g) in functional form as $V(g)$. Let $n(g)$ denote the length of the segment of x that has been constructed at the beginning of generation g, and let $g(n) := \inf\{g \in \mathbb{N}: n \le n(g)\}$ denote its quasi-inverse. Clearly $n(g)$ is strictly increasing in g and $g(n)$ is increasing, but often constant. With these definitions, we have $V = V(g) = V(g(n))$.

Denote by $\tilde{C}(V)$ the best piecewise linear approximation of C with V vertices, in the sense of minimizing $\psi_C(V) := d(C, \tilde{C}(V))$. Since every rectifiable curve C has a Lipschitz continuous parametrisation, we have that $V \mapsto \psi_C(V)$ is decreasing in V and $\lim_{V\to\infty}\psi_C(V) = 0$. Thus $\lim_{g\to\infty}\psi_C(V(g)) = 0$ and $\lim_{n\to\infty}\psi_C(V(g(n))) = 0$, although the convergence could be very slow (in n) and its speed will depend on the choice of C. Denote by $\bar{C}(n) := \tilde{C}(V(g(n)))$ the sequence of best possible piecewise linear approximations of C indexed by n, and let $\bar{\psi}_C(n) := d(C, \bar{C}(n))$. We have thus shown:

Lemma B.18. *Let $C \subset \Delta^k$ be a rectifiable curve. Then*

$$\lim_{n\to\infty}\bar{\psi}_C(n) = 0.$$

We summarize what we know so far.

1. For all generations g, $\|\hat{p}_{\text{new}} - p_{\text{new}}\| \le \frac{4T}{n} + \frac{k}{T}$, where $p_{\text{new}} = p_v$ for $v \in [V^g]$ (Lemma B.16). This means the following. Suppose at the beginning of segment v in generation g we have $n = \text{length}(x)$. By definition, we have $r^x(n) = p_{\text{old}}$ and $r^{xz^{[\tilde{\ell}]}}(n+\tilde{\ell}\tilde{T}) = \hat{p}_{\text{new}}$. Furthermore, for $m = i\tilde{T}$, $i \in [\tilde{\ell}]$ we have $r^{xz^{[i]}}(n+i\tilde{T}) \in l(p_{\text{old}}, \hat{p}_{\text{new}})$.
2. Furthermore, (by Lemma B.17) for all $j \in [\tilde{\ell}\tilde{T}]$, $d\left(r^{xz^{[\tilde{\ell}]}}(j), l(p_{\text{old}}, \hat{p}_{\text{new}})\right) \le \frac{2T+k}{n}$ — the relative frequencies for all points in the segment are close to the line segment $l(p_{\text{old}}, \hat{p}_{\text{new}})$.

Combining these facts, and appealing to the triangle inequality, we conclude that the sequence x constructed by Algorithm 1 satisfies

$$d(r^x(n), C) \le \frac{4T}{n} + \frac{k}{T} + \frac{2T+k}{n} + \bar{\psi}_C(n) \quad \forall n \in \mathbb{N}. \tag{28}$$

We now choose $T = T(n)$ and $V = V(g(n))$ appropriately. One choice is to choose $T(n) = \sqrt{n}$. Equation 28 then implies

$$d(r^x(n), C) \le \frac{4\sqrt{n}}{n} + \frac{k}{\sqrt{n}} + \frac{2\sqrt{n}+k}{n} + \bar{\psi}_C(n) \quad \forall n \in \mathbb{N}$$

which implies that for all $n \in \mathbb{N}$ with $\sqrt{n} > k$,

$$d(r^x(n), C) \le \frac{4}{\sqrt{n}} + \frac{k}{\sqrt{n}} + \frac{4}{\sqrt{n}} + \bar{\psi}_C(n) = \frac{8+k}{\sqrt{n}} + \bar{\psi}_C(n),$$

and thus Lemma B.18 implies $\lim_{n\to\infty} d(r^x(n), C) = 0$. By definition of the Hausdorff distance this means that for any $\epsilon > 0$ and any point $y \in C$, there exists an n such that $\|r^x(n) - y\| \le \epsilon$. Furthermore, by the generational nature of our construction, if $r^x(n)$ is ϵ-close to y then for each subsequent generation g, there exists n^g such that $r^x(n^g)$ is also ϵ-close to y. Since there are an infinite number of generations g, the sequence r^x visits an ϵ-neighbourhood of y infinitely often. Since ϵ was arbitrary, y is thus a cluster point of r^x. Since $y \in C$ was arbitrary, every element of C is a cluster point of r^x.

Finally, since in each generation the bounds above constrain r^x more and more tightly, there cannot exist cluster points that are not in C; that is, $\mathrm{CP}(r) \subseteq C$. We have thus proved Theorem 3.3.

B.9 Remarks on the Construction

We make a few remarks on the construction.

1. By the definition of $\tilde{\ell}$ we are guaranteed that $\tilde{\ell} \ge 1$ for each segment and each generation. For the construction to approximate well, we need $\tilde{\ell} \gg 1$, which it will inevitably be when n gets large enough.
2. Recall $n(g)$ is the length of x at the beginning of generation g. Let $L_C^g := \mathrm{length}(C^g) > 0$. Since each step which r^x moves is of size less than $1/n(g)$ (see Equation 13), we require at least $L_C^g \cdot n(g)$ steps for r^x to traverse the whole of C in generation g. Thus at the end of generation g and the beginning of generation $g+1$ we have

$$n(g+1) \ge n(g) + L_C^g \cdot n(g) = \lambda_C^g \cdot n(g).$$

   Furthermore, L_C^g is increasing in g and approaches length(C). Thus for all g, $\lambda_C^g > 1$ (and is in fact increasing in g). Hence $n(g)$ grows exponentially with g.
3. The growth of the length of x is controlled in a complex manner by the nature of the curve C. In particular if C is very complex, then $\bar{\psi}_C$ must decay slowly. Furthermore, if C has parts close to $\partial\Delta^k$, in particular if some vertices v of the piecewise linear approximation C^g are close to $\partial\Delta^k$, then $\tilde{\ell}$ can end up very large for that segment, meaning that the length of the sequence x grows more rapidly. See Lemma B.19 for an illustration of this observation.

Finally note that since (by Lemma B.10) $\mathrm{CP}(r^x)$ must always be connected, we have in Theorem 3.3 what appears to be the most general result possible (under the restriction that x takes values only in a finite set $[k]$). We do not know what the appropriate generalization is to sequences x that can take values in an infinite (or uncountable) set[27].

[27] Although we do not pursue this in any detail, we remark that one could design an algorithm to construct x such that $\mathrm{CP}(r^x)$ is *any* subset $D \subseteq \Delta^k$ by using our algorithm as a subroutine. The idea would be to construct a space filling curve that fills D, each generation of which is a rectifiable curve. One would appeal to our algorithm for each generation, and then once within a suitable tolerance, change the target to be the next generation of the space filling curve. A suitable method would be to simply intersect an extant families of closed space filling curves for k dimensional cubes with the $(k-1)$-simplex (e.g. generalisations of the Moore curve), attaching joins on ∂D where necessary. Since (see the main body of the paper) it ends up being only the convex hull of $\mathrm{CP}(r^x)$ that matters, such an exotic construction is of little direct interest.

B.10 From Boundary to Curves

Proof of Corollary 3.4. Bronshteyn & Ivanov (1975) show that if D is a convex set contained in the unit ball (w.r.t. the Euclidean norm) in $\mathbb{R}^n$ and $\epsilon < 10^{-3}$, then there exists a set of at most $K_\epsilon := 3\sqrt{n}(9/\epsilon)^{(n-1)/2}$ points whose convex hull is at most ϵ away from D. Thus for any convex $D \subseteq \Delta^k$, and any $\epsilon < 10^{-3}$ there is a polyhedron D_ϵ comprising the convex hull of $K_\epsilon < \infty$ points $Q_\epsilon := \{q_i \in \Delta^k \colon i \in [K_\epsilon]\}$ such that $d(D, D_\epsilon) < \epsilon$, where d is the Hausdorff distance (27). Hence for any $\epsilon < 10^{-3}$ there exists a closed rectifiable curve C_ϵ (constructed by linearly connecting successive points in Q_ϵ) such that $d(\operatorname{co} C_\epsilon, D) < \epsilon$. One can then construct a sequence x such that $\mathrm{CP}(r^x) = \partial D$ as follows. Start with $\epsilon_0 = 10^{-3}$. Pick and construct C_{ϵ_0} as above. Construct x according to the previous procedure (Appendix B.4–B.9) to get an entire generation of r^x within ϵ_0 of C_{ϵ_0}. Then let $\epsilon_1 = \epsilon_0/2$ and repeat the procedure, appending the constructed sequence. Continue iterating (dividing the ϵ in half each phase) and one achieves that in the limit $\mathrm{CP}(r^x) = \partial D$. □

B.11 Illustration

We illustrate our construction for $k = 3$ for two (identical) generations and thus a single piecewise linear (in fact polygonal) approximation C^g of C. We take for C (a scaled version of) the lemniscate of Bernoulli (Lockwood, 1961, Chapter 12), (Lawrence, 1972, Section 5.3) mapped onto the 2-simplex as the space curve $\{(z_1(t), z_2(t), z_3(t) \colon t \in [0, 2\pi]\}$, where

$$
\begin{aligned}
z_1(t) &= \frac{1}{3} + \frac{1}{12}\frac{2\cos(t)}{1+\sin^2(t)}\\
z_1(t) &= \frac{1}{3} + \frac{1}{12}\frac{2\sin(t)\cos(t)}{1+\sin^2(t)}\\
z_3(t) &= 1 - z_1(t) - z_2(t).
\end{aligned}
$$

We set $T = 12$ and $V = 30$ (number of nodes in the polygonal approximation C^g for both $g = 1, 2$ to make the figure less cluttered) and we iterated long enough to go around the lemniscate twice, which resulted in a sequence of length 85677. As the construction proceeded, $\tilde{T}$ went from 9 up to 164 and $\tilde{\ell}$ went from 1 or 2 for the first few segments up to 69 for the last (with the largest value being 104). The results can be seen in Figure 2 which plots the achieved relative frequencies at different zoom levels. The small red squares are the vertices of C^g.

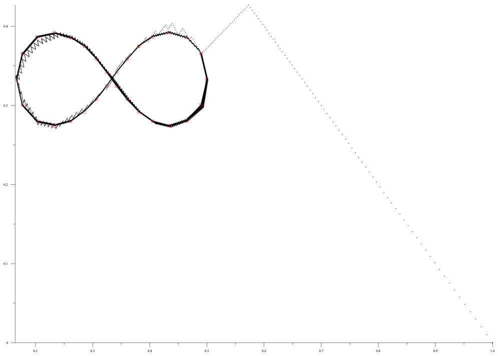
B.12 Construction of x such that $\operatorname{co}\mathrm{CP}(r^x) = \Delta^k$

How much of the simplex can we cover with $\mathrm{CP}(r^x)$? This question is poorly posed as (it seems) we can only ever construct one dimensional sets that are the set of cluster points of r^x. However, for inducing an upper prevision, all that matters is the convex hull of $\mathrm{CP}(r^x)$. We now show the convex hull of $\mathrm{CP}(r^x)$ can be made as large as conceivable with simpler and more explicit construction:

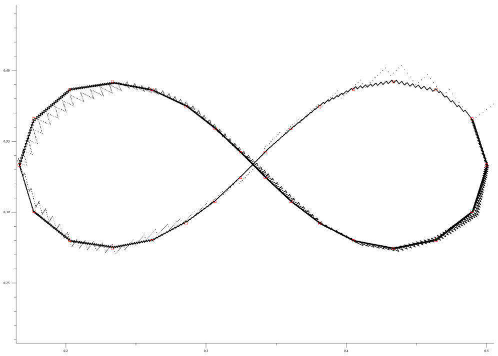
Lemma B.19. *Suppose $k \in \mathbb{N}$. There exists a sequence $x \colon \mathbb{N} \to [k]$ such that* $\operatorname{co}(\mathrm{CP}(r^x)) = \Delta^k$.

Proof. As before, our proof is constructive. Recall $e_1, \ldots, e_k$ are the vertices (and extreme points) of the $(k-1)$-simplex, and $\operatorname{co}\{e_1, \ldots, e_k\} = \Delta^k$. We will construct a sequence x such that for all $\epsilon > 0$, and all $i \in [k]$, r^x visits $N_\epsilon(e_i)$ infinitely often. Since $\lim_{\epsilon\to 0} \operatorname{co}\bigcup_{i\in[k]} N_\epsilon(e_i) = \Delta^k$ we will have achieved the desired result.

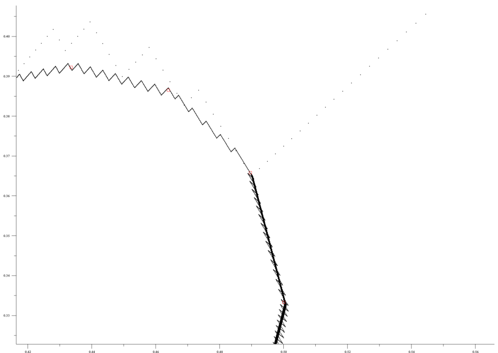
We again make use of (12). We will construct a sequence x by adding segments (s) such that for each successive s we drive r^x closer and closer towards one of the vertices e_i $(i \in [k])$. In order to do this, at each s we append m copies of i to the current x. Specifically, we construct x as follows:



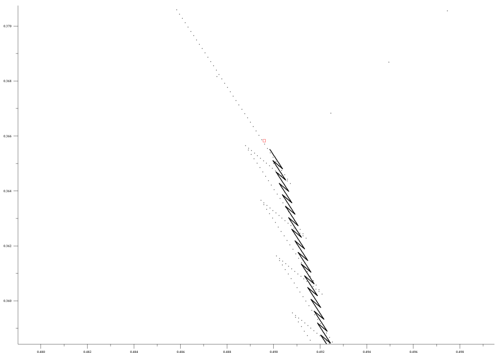
(a) Overall view, showing initialisation.



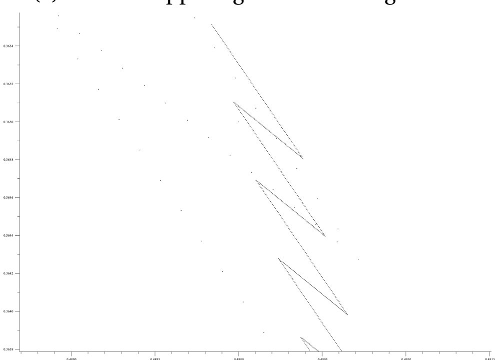
(b) Closer view of just the region containing C.



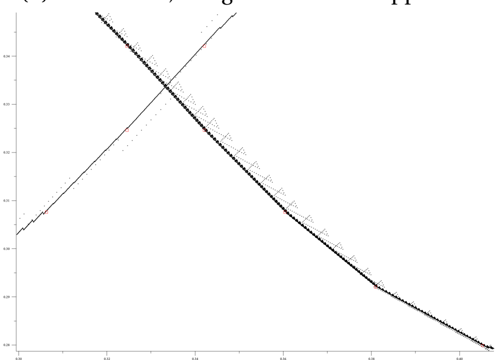
(c) Zoom of upper right corner of Figure 2b.



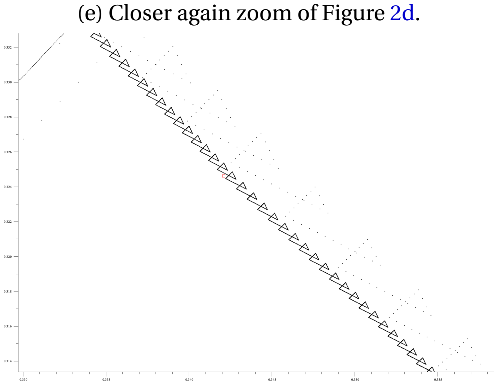
(d) Closer view; two generations are apparent.



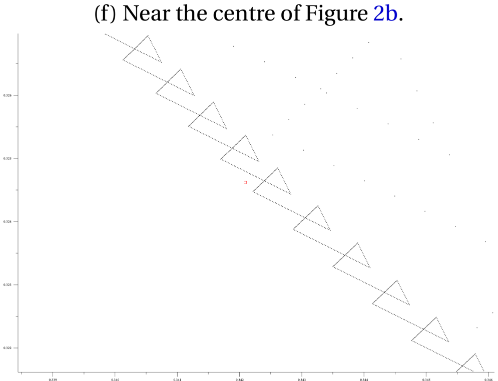
(e) Closer again zoom of Figure 2d.



(f) Near the centre of Figure 2b.



(g) Further zoom of Figure 2f.



(h) An even closer view of Figure 2g.

Figure 2: Illustration of approximation of the polygonal curve C^g by the relative frequencies of the sequence constructed according to Algorithm 1. Two generations were used. Red squares are vertices of C^g. See Section B.11 for details.

```
Require: ϕ: ℕ → ℕ
x ← ⟨⟩
s ← 1
while true do
    i ← s mod k                        ▷ Cycle around the vertices of the simplex
    m ← ϕ(s+1) − ϕ(s)
    x ← x i^m                          ▷ Append m copies of i to x
    s ← s+1
end while
```

We need to make m large enough so that the convex combination coefficient $\frac{m}{n+m}$ approaches 1. To that end, consider an increasing function $\phi\colon \mathbb{N} \to \mathbb{N}$ which will further restrict later. The role of ϕ is to control n as a function of segment number s; that is $n = \phi(s)$ and thus $m = \phi(s+1) - \phi(s)$. With this choice, we have

$$\frac{n}{n+m} = \frac{\phi(s)}{\phi(s+1)} \quad \text{and} \quad \frac{m}{n+m} = 1 - \frac{\phi(s)}{\phi(s+1)}.$$

and thus for all $s \in \mathbb{N}$

$$r^x(\phi(s+1)) = \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) + \left(1 - \frac{\phi(s)}{\phi(s+1)}\right) r^{x^{+\phi(s)}}\left(\phi(s+1) - \phi(s)\right). \tag{29}$$

We demand that $\lim_{s\to\infty} \frac{\phi(s)}{\phi(s+1)} = 0$ so that as s increases, the second term in (29) dominates. For any $s \in \mathbb{N}$, $r^x(\phi(s)) \in \Delta^k$ and thus

$$\begin{aligned}
\|e_{s \bmod k} - r^x(\phi(s+1))\| &= \left\| e_{s \bmod k} - \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) - \left(1 - \frac{\phi(s)}{\phi(s+1)}\right) r^{x^{+\phi(s)}}\left(\phi(s+1) - \phi(s)\right) \right\| \\
&= \left\| \frac{\phi(s)}{\phi(s+1)} e_{s \bmod k} - \frac{\phi(s)}{\phi(s+1)} r^x(\phi(s)) \right\| \\
&= \frac{\phi(s)}{\phi(s+1)} \left\| e_{s \bmod k} - r^x(\phi(s)) \right\| \\
&\le \frac{\phi(s)}{\phi(s+1)} \cdot 2,
\end{aligned}$$

where the second line follows from the fact that we constructed x such that $r^{x^{+\phi(s)}}\left(\phi(s+1) - \phi(s)\right) = e_{s \bmod k}$. But by assumption, $\lim_{s\to\infty} \frac{\phi(s)}{\phi(s+1)} = 0$ and hence for any $\epsilon > 0$ there exists s_ϵ such that for all $i \in [k]$,

$$\left|\left\{s \in \mathbb{N}\colon\ s > s_\epsilon,\ s \bmod k = i,\ \left\|e_i - r^x(\phi(s+1))\right\| \le \epsilon\right\}\right| = \aleph_0.$$

That is, for each $i \in [k]$, each ϵ-neighbourhood of e_i is visited infinitely often by the sequence r^x and hence $\{e_1, \ldots, e_k\} \subseteq \mathrm{CP}(r^x)$. But since $r^x(n) \in \Delta^k$ for all $n \in \mathbb{N}$ we conclude that indeed $\mathrm{co}(\mathrm{CP}(r^x)) = \mathrm{co}(\{e_1, \ldots, e_k\}) = \Delta^k$ as required. □

A suitable choice of ϕ is $\phi(s) = \lceil \exp(s^\alpha) \rceil$ for some $\alpha > 1$, in which case $\frac{\phi(s)}{\phi(s+1)} \approx \exp\left(-\alpha s^{\alpha-1}\right)$. An argument as in the proof of Theorem 3.3 would show that $S_k \coloneqq \bigcup_{i\in[k]} l\left(e_i, e_{(i+1) \bmod k}\right) \subseteq \mathrm{CP}(r^x)$. Observe that when $k = 3$, $S_k = \partial\Delta^k$, but for $k > 4$, that is not true, even though $\mathrm{co}(S_k) = \partial\Delta^k$.

C Unstable Independence

Closely related to conditional probability is the concept of *statistical independence*. Independence plays a central role not only in Kolmogorov's (Durrett, 2019, p. 37), but more generally in most probability theories (Levin (1980); Fine (1973, Section IIF, IIIG and VH)). Already de Moivre (1738/1967, Introduction, p. 6) nicely summarized a pre-theoretical, probabilistic notion of real-world independence:

> Two events are independent, when they have no connexion one with the other, and that the happening of one neither forwards nor obstructs the happening of the other.

This intuitive conception was then formalized by Kolmogorov (1933) (translated in (Kolmogorov, 1956)) in the following classical definition.

Definition C.1. *Let $(\Omega, \mathscr{F}, P)$ be a probability space.*[28] *We call events $A \in \mathscr{F}$ and $B \in \mathscr{F}$ classically independent if $P(A \cap B) = P(A)P(B)$.*

If $P(B) > 0$, then we can equivalently express this condition as $P(A|B) = P(A)$ by using the definition of conditional probability.

Kolmogorov's definition is formal and it has been questioned whether it is an adequate expression of what we mean by independence in a statistical context (Von Collani, 2006). As it is stated in purely measure-theoretic terms, it is unclear whether it has reasonable frequentist semantics. In our framework, we construct an intuitive definition of independence, where the independence of events is based on an independence notion of *processes* (cf. (von Mises & Geiringer, 1964, p. 35-39)). Therefore, our definition is thoroughly grounded in the frequentist setting. Furthermore, we shall generalize the independence concept to the case of possible divergence, where new subtleties come into play. We will then consider how our definitions relate to the classical case when relative frequencies converge. Assume that a sequence $\overrightarrow{\Omega}$ is given and we have constructed an upper probability $\overline{P}$ as in Section 2.2.

Definition C.2. *We call an event $B \in 2^{\overrightarrow{\Omega}}_{1+}$ irrelevant to another event $A \subseteq \Omega$ if:*

$$\overline{P}(A|B) = \overline{P}(A).$$

This definition captures the concept of *epistemic irrelevance* in the imprecise probability literature (Miranda, 2008). Why does this definition possess reasonable frequentist semantics? Consider what $\overline{P}(A|B)$ means (see Section 4.1): we are considering a subsequence, induced by the indicator gamble χ_B, that is, we condition (in an intuitive sense) on the occurence of B; and on this subsequence, we then consider an unconditional upper probability. If this then coincides with the orginal upper probability, our decision maker values A just the same whether B occurs or not. Thus B is irrelevant for putting a value on A. In contrast to the classical, precise case, irrelevance is not necessarily symmetric. Hence, we define independence as follows.

Definition C.3. *Let $A, B \in 2^{\overrightarrow{\Omega}}_{1+}$. We call A and B independent if $\overline{P}(A|B) = \overline{P}(A)$ and $\overline{P}(B|A) = \overline{P}(B)$.*

Thus, we have obtained a grounded concept of independence for events. We note that Definition C.3 is similar to a condition proposed by Walley & Fine (1982) for independence of joint experiments[29]; they did not propose an independence concept for gambles.

How can we extend this to an irrelevance and independence concept for gambles? First, we briefly recall how this is done in the classical case.

[28] Here, Ω is the possibility space, $\mathscr{F}$ is a σ-algebra and P is a countably additive probability measure.

[29] Walley & Fine (1982) considered outcomes of "joint experiments" in $\Omega \times \Omega$. They furthermore demanded that lower limits of relative frequencies factorize; to us it is not clear from a strictly frequentist perspective why this condition should be introduced.

Definition C.4. *Let* $(\Omega, \mathscr{F}, P)$ *be a probability space and fix the Borel* σ*-algebra* $\mathscr{B}$ *on* $\mathbb{R}$*. Given two gambles* $X, Y : \Omega \to \mathbb{R}$*, we say that they are* classically independent *if:*

$$P(A \cap B) = P(A)P(B) \quad \forall A \in \sigma(X), B \in \sigma(Y),$$

where the σ*-algebra generated from a gamble* X*,* $\sigma(X)$*, is defined as the smallest* σ*-algebra which* X *is measurable with respect to:*

$$\sigma(X) := \sigma\left(X^{-1}(\mathscr{B})\right),$$

and $\sigma(\mathscr{H})$ *is the smallest* σ*-algebra containing all sets* $H \in \mathscr{H}$*,* $\mathscr{H} \subseteq 2^{\Omega}$*.*

Thus independence of gambles is reduced to independence of events. But note that this definition inherently depends on the choice of the Borel σ-algebra on $\mathbb{R}$. In our case, this is similar: to define irrelevance and independence on gambles, we need to fix a set system on $\mathbb{R}$, but we leave the choice open in general.

Definition C.5. *Assume a set system* $\mathscr{H} \subseteq 2^{\mathbb{R}}$ *and two gambles* $X, Y : \Omega \to \mathbb{R}$ *are given. We call* Y *irrelevant to* X *with respect to* $\mathscr{H}$ *if*

$$\overline{P}(X^{-1}(A)|Y^{-1}(B)) = \overline{P}(X^{-1}(A)) \quad \forall A, B \in \mathscr{H} \text{ if } Y^{-1}(B) \in 2^{\vec{\Omega}}_{1+}.$$

Similarly, we call them independent when both directions hold.

Observe that if $\mathscr{H} = \mathscr{B}$ and $\overline{P}$ was actually a precise P on $\sigma(X)$ and $\sigma(Y)$, this definition would be equivalent to Definition C.4 (modulo the subtlety regarding conditioning on measure zero events), due to the following.

Lemma C.6. *Given set systems* $\mathscr{A}, \mathscr{B} \subseteq \Omega$*, in the precise case, the following statements are equivalent.*

PI1. $P(A|B) = P(A)\ \forall A \in \mathscr{A}, B \in \mathscr{B}$ *and* $P(B) > 0$.
PI2. $P(A \cap B) = P(A)P(B)\ \forall A \in \mathscr{A}, B \in \mathscr{B}$.

Proof. Obviously PI2 implies PI1 by the definition of conditional probability. One only has to check that when PI1 holds, that PI2 holds even if $P(B) = 0$. But if $P(B) = 0$, then also $P(A \cap B) = 0$ due to monotonicity of P in the sense of a capacity. □

Example C.7. Choose $\mathscr{H} := \{(-\infty, a] : a \in \mathbb{R}\}$ in Definition C.5. Such an $\mathscr{H}$ is called a Π-system, which is a non-empty set system that is closed under finite intersections. This particular Π-system can in fact be used to define independence in the classical case, which is done in terms of the joint cumulative distribution function.

D Pathological or Normal?

Much of the confusion about probability arises because the true depth of the law of large numbers as an extremely hard analytical assertion is not appreciated at all. — Detlef Dürr and Stefan Teufel (2009, p. 62)

When one looks at *finite* sequences $x : [n] \to [2]$, there is a simple counting argument using the binomial theorem that illustrates that the vast majority of the 2^n possible sequences have roughly equal numbers of elements with values of 1 and 2. If one *assumes* that an infinite sequence $x : \mathbb{N} \to [2]$ is generated i.i.d. then this argument can be used to prove the law of large numbers, which ensures "most" sequences have relative frequencies which converge.

Hence the construction, as illustrated in the present paper, of sequences $x : \mathbb{N} \to [k]$ with *divergent* relative frequencies naturally raises the question of how contrived they are. That is, are we examining a rare pathology, or something "normal" that we might actually encounter in the world? We will refer to sequences whose relative frequencies converge as "stochastic sequences" and sequences whose relative frequencies do not converge as "non-stochastic sequences"[30].

The classical law of large numbers suggests that indeed "almost all" sequences are stochastic, and therefore, by such reasoning, the non-stochastic sequences with which we have concerned ourselves in the present paper are indeed pathological exceptions. In this appendix we will argue:

1. This very much depends upon what one means by "rare" or "almost all" and there are many choices, and the only real argument in favour of the usual ones (which declare non-stochastic sequences rare) is familiarity — different notions of "typicality" (for that is what is at issue) lead to very different conclusions. Specifically, there are choices (arguably just as "natural" as the familiar ones) which imply that rather than non-stochastic sequences being rare, they are in fact the norm in a very strong sense.
2. Nevertheless, none of the mathematical nuances of the previous point allow one to conclude *anything* about the empirical prevalence of stochastic or non-stochastic sequences in the world. Indeed, no purely mathematical reasoning allows one to draw such conclusions, unless one wishes to appeal to some conception of a Kantian "synthetic a priori."

We will first explore what can said from a purely mathematical perspective, illustrating that there is a surprising amount of freedom of choice in precisely posing the problem, and that the choices are consequential. Then in Subsection D.3 we examine the question of prevalence of non-stochastic sequences actually in the world.

D.1 The Mathematical Argument — The Choices to be Made

The classical Law of Large Numbers says "almost all sequences" are stochastic. But the "almost all" claim comes from the mapping of sequences to real numbers in $[0,1]$ and then making a claim that "almost all" numbers correspond to stochastic sequences. Thus there are at least three choices being made here:

Mapping from Sequences to Real Numbers The choice of mapping from sequences to real numbers, to enable to use of some notion of typicality on $[0,1]$ to gauge how common stochastic sequences are.

Notion of Typicality The notion of typicality to be used (e.g. Cardinality, Hausdorff dimension, Category or Measure).

Specific Index of Typicality Within the above choice of notion of typicality, the particular choice of typicality index, e.g. the measure or topology that underpins the notion of typicality.

The choices for the classical law of large numbers are 1) k-ary positional representation; 2) a σ-additive measure on $[0,1]$; 3) The Lebesgue measure. As we shall summarize below, each of these three choices substantially affects the theoretical preponderance of non-stochastic sequences.

That there are alternate choices that lead to the unusual conclusion that non-stochastic sequences are "typical" has been known for some time: "This result may be interpreted to mean that the category analogue of the strong law of large numbers is false" (Oxtoby, 1980, p. 85); see also (Méndez, 1981). The significance of this fact has been stressed recently (Seidenfeld et al., 2017; Cisewski et al., 2018). And it has been observed that the introduction of alternate topologies can change whether sequences are stochastic (Khrennikov, 2013). However, the strongest results arise in number theory, motivated by the notion of a "normal number."

[30] This dichotomy would appear clear-cut, but there is a subtlety: there exist sequences such that $(x_i)_{i\in\mathbb{N}}$ and $(y_i)_{i\in\mathbb{N}}$ are both stochastic, but the *joint* sequence $((x_i, y_i))_{i\in\mathbb{N}}$ is *non*-stochastic (Rivas, 2019); that is, the marginal relative frequencies converge, but the joint relative frequencies do not!

D.2 Notions of Typicality — Cardinality, Dimension, Comeagreness, and Measure

Let $\mathscr{S}$ (resp. $\mathscr{N}$) denote the set of stochastic (resp. non-stochastic) sequences $\mathbb{N} \to [k]$. That is, $\mathscr{S} := \{x \colon \mathbb{N} \to [k] \colon \lim_{n\to\infty} r^x(n) \text{ exists}\}$ and $\mathscr{N} := [k]^{\mathbb{N}} \setminus \mathscr{S}$. (For simplicity, and alignment with Appendix B, we restrict ourselves to sequences whose domain is $[k]$.)

In order to make a claim regarding the relative preponderance of stochastic versus non-stochastic sequences, they are often mapped onto the unit interval.[31] In such cases, the question of relative preponderance of classes of sequences is reduced to that of a question concerning the relative preponderance of classes of subsets of $[0,1]$. The question then arises of how to measure the size of such subsets. Unlike in the finite case mentioned above, merely counting (i.e. determining the cardinality of the respective subsets) is hardly adequate, as it is easy to argue that $|\mathscr{S}| = |\mathscr{N}| = \aleph_1$. There are three notions that have been used to compare the size of $\mathscr{S}$ and $\mathscr{N}$:

Measure A countably additive measure, usually the Lebesgue measure on $[0,1]$.

Meagre / Comeagre A subset S of a topological space X is meagre if it is a countable union of nowhere dense sets (i.e. sets whose closure has empty interior). A set S is comeagre (residual) if $X \setminus S$ is meagre.

Dimension A variety of fractal dimensions, such as the Hausdorff dimension, have also been used to judge the size of non-stochastic sequences (and the numbers they induce); however for space considerations we omit discussion of these results[32].

Some of the results obtained in the literature are summarized below. The object is not to state them in an entirely formal manner, or even to describe them in their full generality. Rather we simply wish to show the diversity of conclusions available by tweaking the three choices enumerated above. If no representation is mentioned, the usual k-ary positional representation is used, whereby $\tilde{x} \in [0,1]$ is constructed from $x \colon \mathbb{N} \to [k]$ via $\tilde{x} := \sum_{i\in\mathbb{N}} (x_i - 1)k^{-i}$ (the $(x_i - 1)$ term is required because our sequences map to $[k] = \{1,\ldots,k\}$). Obviously every $x \in [k]^{\mathbb{N}}$ maps to some $\tilde{x} \in [0,1]$; and every $z \in [0,1]$ corresponds to at least one $x \in [k]^{\mathbb{N}}$ (recalling we have to handle the situation that, when $k = 10$ for example, $0.4\overline{9} = 0.5\overline{0}$, where $\overline{i}$ means that i is repeated infinitely, and thus there are two sequences $x_1, x_2 \in [k]^{\mathbb{N}}$ such that $1/2 = \tilde{x}_1 = \tilde{x}_2$). [33] Let $\tilde{\mathscr{S}} := \{\tilde{x} \in [0,1] \colon x \in \mathscr{S}\}$ and $\tilde{\mathscr{N}} := \{\tilde{x} \in [0,1] \colon x \in \mathscr{N}\}$.

Most (Lebesgue measure) sequences are stochastic This is the classical strong law of large numbers. If μ_{leb} denotes the Lebesgue measure on $[0,1]$, then the claim is that $\mu_{\text{leb}}(\tilde{\mathscr{S}}) = 1$.

Most (comeagre) sequences are non-stochastic Let $X_i = [2]$ and $X = \times_{i\in\mathbb{N}} X_i$ equipped with the product topology. As a set $X \cong [2]^{\mathbb{N}}$. Then $\mathscr{N} \subset X$ is comeagre (Oxtoby, 1980).

Most (comeagre) sequences are stochastic With different choices of topology, the opposite conclusion holds — there are topologies such that $\mathscr{S}$ is comeagre (Calude et al., 2003).

Most (comeagre) sequences are extremely non-stochastic Let $\tilde{\mathscr{N}}^\star$ denote the subset of $[0,1]$ of $\tilde{x}$ corresponding to $x \in [k]^{\mathbb{N}}$ which satisfy $\forall i \in [k]$, $\liminf_{n\to\infty} r_i^x(n) = 0$ and $\limsup_{n\to\infty} r_i^x(n) = 1$. These sequences are (justifiably) called extremely non-stochastic; the sequence constructed in Subsection B.12 is an example. Then the set $\tilde{\mathscr{N}}^\star$ is comeagre in the usual topology of real numbers (Calude & Zamfirescu, 1999); confer (Calude, 2002, section 7.3).

Most (comeagre) sequences are perversely non-stochastic Denote the set of perversely nonstochastic sequences $\mathscr{N}^{\star\star} := \{x \in [k]^{\mathbb{N}} \colon \mathrm{CP}(r^x) = \Delta^k\}$. Observe $\mathscr{N}^{\star\star} \subset \mathscr{N}^\star$. Let $\tilde{\mathscr{N}}^{\star\star} := \{\tilde{x} \in [0,1] \colon x \in$

[31] Attempts known to us to judge the relative sizes of $\mathscr{S}$ and $\mathscr{N}$ which do not rely on such a mapping are described in the first, third and sixth of the cases listed below, and rely upon imposing a topology directly on the set of sequences $[k]^{\mathbb{N}}$.

[32] See for example (Eggleston, 1949; Olsen, 2004; Gu & Lutz, 2011; Bishop & Peres, 2017; Albeverio et al., 2017).

[33] This non-uniqueness of the representation will not affect the results below because $\{\tilde{x}_1 = \tilde{x}_2 \in [0,1] \colon x_1 \neq x_2\} = \mathbb{Q} \cap [0,1]$ and is of cardinality $\aleph_0$, whereas $|[k]^{\mathbb{N}}| = |[0,1]| = \aleph_1$.

$\mathcal{N}^{\star\star}\}$ (what Olsen (2004) calls "extremely non-normal numbers", but we use "extremely" for the larger set $\tilde{\mathcal{N}}^{\star}$). Then $\tilde{\mathcal{N}}^{\star\star}$ is comeagre (in the usual topology of real numbers) (Aveni & Leonetti, 2022; Olsen, 2004). An even stronger result holds. Let A denote a (not necessarily uniform) finite averaging operator and let $\mathcal{N}^{\star\star\star} := \{x \in [k]^{\mathbb{N}} : \mathrm{CP}(A(r^x)) = \Delta^k\}$ and $\tilde{\mathcal{N}}^{\star\star\star} := \{\tilde{x} \in [0,1] : x \in \mathcal{N}^{\star\star\star}\}$. Observe $\mathcal{N}^{\star\star\star} \subset \mathcal{N}^{\star\star}$ and $\tilde{\mathcal{N}}^{\star\star\star} \subset \tilde{\mathcal{N}}^{\star\star}$. Then $\tilde{\mathcal{N}}^{\star\star\star}$ is also comeagre (Stylianou, 2020)!

Most (Lebesgue measure) sequences are non-stochastic There exist a range of representations of real numbers called Q^*-representations (Q^* is a $k \times \infty$ matrix valued parameter of the representation); see (Albeverio et al., 2005, Section 4) for details. Let $\tilde{x}^{Q^*}$ denote the Q^*-representation of a sequence $x \in [k]^{\mathbb{N}}$, and $\tilde{\mathcal{S}}^{Q^*} := \{\tilde{x}^{Q^*} : x \in \mathcal{S}\}$ and $\tilde{\mathcal{N}}^{Q^*} := \{\tilde{x}^{Q^*} : x \in \mathcal{N}\}$. Then there exist Q^* such that $\mu_{\text{leb}}\left(\tilde{\mathcal{N}}^{Q^*}\right) = 1$. (Albeverio et al., 2005, p. 627). Thus if the size of $\mathcal{N}$ is judged via certain Q^* representations, Lebesgue almost all sequences are non-stochastic!

An obvious conclusion to draw from the above examples is that in answering the question of the preponderance of non-stochastic sequences, one can get essentially whatever answer one wants by choosing a range of different precise formulations of the question. At the very least, this should make us skeptical of any purely mathematical attempts to reason whether one might expect to encounter non-stochastic sequences in practice — the topic to which we now turn.

D.3 Typical Real Sequences

The laws of large numbers cannot be applied for describing the statistical stabilization of frequencies in sampling experiments. — Andrei Khrennikov (2009, p. 20)

What do the above points imply about the likelihood one will encounter stochastic or non-stochastic sequences when performing real measurements?

Nothing.

This is not to say that in actuality we will often encounter non-stochastic sequences. Rather our point is that no amount of purely theoretical reasoning will be able to tell us in advance how "likely" it is to do so. What is at issue is whether stochastic sequences are in fact "typical" in our world.

Perhaps the most surprising thing about the mathematical results summarized above is the extent to which different notions of typicality affect the conclusions. This raises the question of whether some notions of typicality are more justified when wishing to consider real sequences that have been measured in the world. In the study of physics (especially aspects of physics that are apparently intrinsically statistical) such questions have been raised, and below we briefly summarize what is known.

Traditionally, "probability" is considered as a primitive, and notions of typicality are derived from that in terms of their "probability" of occurring. And the above examples illustrate that attempts to argue for the Lebesgue measure having a privileged role as the "right" notion of typicality are barking up the wrong tree; confer (Pitowsky, 2012). But this will not do for our question. Typicality is a more fundamental notion (Galvan, 2006; 2007) — arguably the "mother of all" notions of probability (Goldstein, 2012). Typicality is at the core of questions of non-stochastic randomness in physics, thus (consistent with the perspective of the present paper) leading to non-additive measures of typicality (Galvan, 2022) (essentially defining a measure of typicality inspired by a coherent upper probability) which allows the extension to notions of mutual typicality necessary to reason about situations such as that referred to in footnote 30.

In fact, typicality plays an even stronger role than answering questions regarding the preponderance of non-stochastic sequences. As Dürr & Struyve (2021, p. 36) observe "the notion of typicality is necessary to understand what the statistical predictions of a physical theory really mean." They note that the usual appeal to the law of large numbers misses the point because while its conclusion is true (convergence of relative frequencies) *if* one sees typical sequences, but "What needs to be explained is why we only see typical sequences! That's actually the deep question underlying the meaning of probability theory from its very beginning ... " (Dürr & Struyve, 2021, p. 37). In classical mechanics, appeal to Liouville's theorem suggests an "invariant measure" as being a natural choice; in the quantum realm, there is an analogous choice (invariant to Bohmian flow) (Dürr & Struyve, 2021, p. 41). But these situations are rather special from the perspective of a statistician. The situation is well summarized by Dürr (2001, p. 130): "What is typicality? It is a notion for defining the smallness of sets of (mathematically inevitable) exceptions and thus permitting the formulation of law of large numbers type statements. Smallness is usually defined in terms of a measure. What determines the measure? In physics, the physical theory." Confer (Dürr & Teufel, 2009, Chapter 4) who observe that from a *scientific* perspective (where one wants to make claims about the world) establishing the pre-conditions for the law of large numbers to hold is "exceedingly difficult"[34].

Very well one might say, but the arguments in favor of typicality of non-stochastic sequences given above all rely on topological arguments, or unusual encodings of sequences to numbers. What is the justification for topological notions of typicality when considering sequences of measurements obtained from the world? Sklar (2000, p. 270) has actually argued that the topological perspective might offer a foundational perspective with *fewer* opportunities for claims of arbitrariness than measure theoretical approaches. See also (Sklar, 1995, p. 185) and the discussion in (Guttmann, 1999, Chapter 4) which reframes the problem away from typicality to viewing the whole question from an approximation perspective where the notion of smallness of sets is naturally one of meagreness. Our point is that even within the restricted realm of physics, there are compelling arguments at least not to take the measure-based notion of typicality for granted. Once that is accepted, non-stochastic sequences seem less unusual.

D.4 Violations of the Law of Large Numbers

A typical universe is in equilibrium; but "our universe is atypical or in non-equilibrium" (Dürr & Teufel, 2009, p. 81) and "what renders knowledge at all possible is nonequilibrium" (Dürr et al., 1992, p. 886) so we should not be surprised if it is not "typical". And indeed that is what we see as long as we look: "The so-called law of large numbers is also invalid for social systems with finite elements during transition" (Chen, 1991). Gorban (2011; 2017; 2018) has documented many examples of real phenomena failing to be statistically stable. Such failures are held to explain departures from "normal" distributions (Philip & Watson, 1987). But more importantly, they mean we should not expect even convergence of relative frequencies in non-equilibrium situations.

Such was the conclusion of Prigogine in his ground-breaking studies of non-equilibrium thermodynamics where he spoke of a "breakdown of the 'law of large numbers'" (Nicolis & Prigogine, 1977, p. 9 and 228); see also (Prigogine, 1978, p. 781), (Prigogine & Stengers, 1985, p. 180) and (Prigogine, 1980, p. 131). And more recently, studies of the use of machine learning systems "in the wild" have recognized that non-stochasticity is not so exotic after all (Katsikopoulos et al., 2021). Thus perhaps its time to downgrade this "law" of nature.

D.5 Repeal of the Law of Large Numbers

A typical universe shows statistical regularities as we perceive them in a long run of coin tosses. It looks as if objective chance is at work, while in truth it is not. There is no chance. That is the basis of mathematical probability theory. — Detlef Dürr and Stefan Teufel (2009, p. 64).

[34] The example they give is for the Galton board, or quincunx, a device often appealed to in order to teach the reality of the central limit theorem — an even stronger claim than the law of large numbers. The irony is that it is rarely checked empirically. And when it has been, it has been found to be untrue! (Bagnold, 1983, Figure 8).

Desrosières (1998) in his history of statistical reasoning has observed the awe with which stable frequencies were viewed when they were first encountered; the effect been interpreted as a hidden divine order[35]. And indeed in many practical situations, stable frequencies do arise. But that does not mean we should take such situations as the only ones that can occur. We may well legitimately call them "normal." But we can better understand the normal by studying the pathological (Canguilhem, 1978, p. 19–20). Ironically in his attempt to clarify the notion of "normal" Canguilhem (1978, p. 103) considered whether "normal" was simply "average" and concluded "the concepts of norm and average must be considered as two different concepts". As we have seen, averages can indeed be far from normal, and potentially quite often.

Perhaps we have been misled by the strange name given to the famous theorem we are considering: by calling it a "law" we are inheriting a lot of baggage as to what we mean by that, baggage that has been traced to notions of divine origin of lawfulness (Zilsel, 1942).[36] And we hanker after lawfulness:

> We . . . naturally hope that the world is orderly. We like it that way... All of us . . . find this idea sustaining. It controls confusion, it makes the world seem more intelligible. But suppose the world should happen in fact to be not very intelligible? Or suppose merely that we do not know it to be so? Might it not then be our duty to admit these distressing facts? (Midgley, 2013, p. 199)

Perhaps the theory of imprecise probabilities presented in this paper which we have grounded in the instability of relative frequencies may help us to admit this "distressing fact." It does suggest to us that the law of large numbers, while a fine and true theorem, as a "law" might be in need of repealing.

[35]"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error.' The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along." (Galton, 1889, p. 66). See also (Rose, 2016) for a recent discussion on statistical normality.

[36]The contrary views regarding the historical origin of the notion of a scientific law (Milton, 1981; Ruby, 1986; Weinert, 1995) do not contradict our point. Ironically, Zilsel was convinced that the law of large numbers *was* a natural law (indeed the most fundamental of natural laws!) (Zilsel, 1916), and in his PhD thesis attempted to argue the case for this philosophically, although after errors were pointed out, he renounced the argument and made no further reference to it (Lenhard & Krohn, 2022, p. 125).